UNIVERSITY OF ALBERTA

CLASSIFICATION, IDENTIFICATION AND PREDICTION FOR NON-LINEAR DYNAMICS

by

Cristina Adela Popescu

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Applied Mathematics

Department of Mathematical and Statistical Sciences

Edmonton, Alberta
Fall 2004

Your file Votre référence
ISBN: 0-612-96002-1
Our file Notre référence
ISBN: 0-612-96002-1

ACKNOWLEDGEMENT

I would like to express my deep gratitude to my supervisor professor Yau Shu Wong for his academic guidance and great encouragement. I also wish to thank Professors Ivan Mizera, Joseph So and Xiaobo Li for helpful advices and generous support throughout my Ph.D. study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The present thesis results form a series of papers [18], [19], [20], [21], [22], [23] appeared or submitted to scientific publications. The thesis consists of two parts, namely Chapters 2 and 3 form the first part, and Chapters 4 and 5 the second part.

In the first part, we focus on classification and prediction for nonlinear dynamical systems. Particularly, we are interested in applications to nonlinear aeroelasticity, which has been receiving considerable attention in the aerospace community in recent years. The second part of the thesis deals with state-space models involving Markov switching. A nested Monte Carlo Expectation Maximization algorithm for learning the parameters of the model is developed. The effectiveness of the proposed method is demonstrated by testing the algorithm for simulated and experimental data.

1.1 Nonlinear aeroelastic models

An understanding of the nonlinear aeroelastic response is a crucial problem for the aerospace community, since complex aeroelastic phenomena play an important role in the safe design of an aircraft. Currently, a major effort is being focussed on the prediction of limit cycle oscillation (LCO) and the flutter boundary. Classical linear theory is not appropriate for studying LCO, and it may give inaccurate results for predicting the flutter boundaries. For example, for flutter testing with several

external store configurations of an F-16A aircraft, Denegri [3] reports the appearance of LCO as well as the flutter and some sudden high-amplitude oscillations. However, he shows that linear flutter analysis fails to determine the oscillation amplitude or the oscillation onset velocity.

In a nonlinear formulation, the aeroelastic problem has been studied by many researchers via mathematical analysis and numerical simulations. In the recent survey paper [16], Lee, Price and Wong report different types of nonlinearities encountered in aeroelastic behavior of aircraft structures. Numerical simulations are used to study bifurcation and chaos for aeroelastic systems with structural nonlinearities. These nonlinearities are illustrated for a two-dimensional airfoil oscillating in pitch and plunge. The governing integro-differential equations of the airfoil motion are reformulated as a nonlinear system of ordinary differential equations (ODE). Numerical and analytical methods for solving the aeroelastic system are reported, including Humbolt's finite difference method, Runge-Kutta time-integrating scheme, and the describing function technique. The same problem is studied by Liu [17] using center manifold theory and the point transformation technique.

Several mathematical models are employed to study the LCO and flutter for various types of airfoils and nonlinearities. Tang and Dowell [26] apply an ONERA stall aerodynamic model for a non-rotating helicopter blade with a parabolic or cubic and freeplay torsional stiffness nonlinearity. A reduced-order model based upon Peters' finite state model for a two-dimensional aerodynamic flow is applied by Tang, Dowell and Virgin [27] to study a three-degree of freedom aeroelastic model with freeplay. For a low-aspect-ratio delta wing structure at low sub-sonic flows, Tang, Henry and Dowell [28] propose the vortex lattice aerodynamic model and the reduced-order aerodynamic technique. In the previously cited papers, the mathematical models are based on nonlinear ordinary differential equations which can be expressed in a state-space form. The systems of nonlinear differential equations are studied using numerical approaches or the describing function technique. The results are found to be in a good agreement with the experimental investigations carried out in the wind tunnel at Duke University.

For the nonlinear control of a prototypical wing section with torsional nonlinearity, Ko, Kurdila and Strganac [13] also consider a model based on a state-space form. The bifurcation structure and parametric stability of the resulting closed-loop dynamical systems are studied using the center manifold method [14]. The results are validated by experiments at Texas A&M University [15].

In order to solve the corresponding nonlinear ODE system analytically or numerically, the system parameters must be known. However, in practice, such as in ground vibration tests and actual flight tests, only the dynamical response corresponding to a given excitation is available. Hence, it is desirable to develop a technique such that one could predict the LCO and other complex aeroelastic phenomena from the given dynamical response.

Given a time series, $X = [x_1, x_2, \ldots, x_n]$, which contains a limited number of transient observations, we wish to predict the subsequent values $[x_{n+1}, x_{n+2}, \ldots]$. In a classical approach, linear time series models or the Kalman filter are commonly used to perform one step ahead predictions or short term predictions. The first part of this thesis proposes two new approaches for the long-term prediction of the aeroelastic response, namely nonlinear time series models and the unscented filter (UF) [11]. The main feature of the two nonlinear approaches proposed in this thesis is the capability of making not only short term predictions, but also accurate long term predictions. Moreover, chapter 3 also presents a new and efficient method based on the UF and the expectation maximization (EM) algorithm for estimating the parameters of an aeroelastic model.

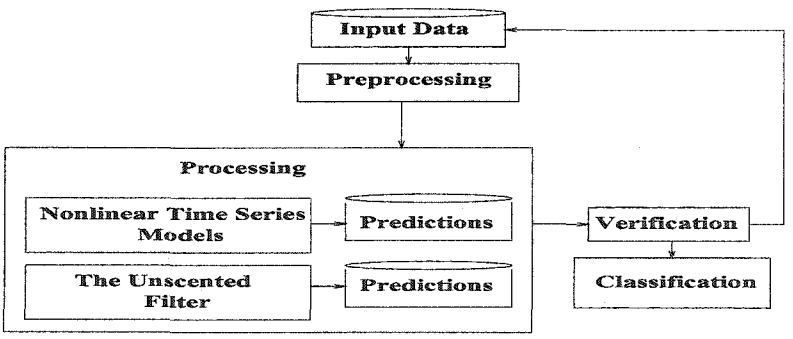

Figure 1.1: Expert data mining system

The nonlinear time series approach and the UF method can be employed to develop an expert data mining system (EDMS) [19]. The basic structure of the EDMS is illustrated in Fig. 1.1. A typical expert system consists of two core components, namely the knowledge base and the reasoning engine. The EDMS, can be especially designed to deal with aeroelastic data with structural nonlinearities. The knowledge base involves data analysis reported in the first part of the thesis, in which the nonlinear time series models and the UF are used to process the input data. Before a long term prediction is presented as output, the information obtained from the knowledge base must be reasoned and certain rules have to be satisfied. In our system, a simple rule is applied, namely the long-term predictions and their classification as LCO, stable or unstable oscillations are given if the two predictions from two different approaches in the processing step agree.

1.2 Models with hidden Markov switching

The second part of this thesis studies models with the nonlinearities given by the hidden Markov switching. In Chapter 4, we start with a result concerning the simplest type of such model, namely the hidden Markov models (HMM). The HMMs are especially known for their applications in automatic speech recognition [24]. Nevertheless, they have been successfully applied to many fields, such as handwriting recognition [2], pattern recognition in molecular biology [1] and fault detection [25].

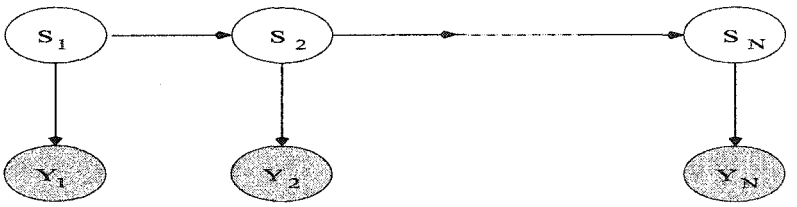

Figure 1.2: Hidden Markov Model

A HMM can be represented graphically by the Bayesian network [4] shown in Fig. 1.2. The figure displays a directed acyclic graph (DAG), in which each node corresponds to a random variable. The shaded nodes and the unshaded nodes

represent the observed variables and the hidden variables respectively. The conditional independence is specified by the edges: each node is conditionally independent from its non-descendants given its parents. Hence, a HMM is formed by the observed sequence $\{Y_n, n \geq 1\}$ and the hidden Markov chain $\{S_n, n \geq 1\}$ such that $P(Y_n|S_n, Y_{n-1}) = P(Y_n|S_n)$ and $P(S_n|Y_n, S_{n-1}) = P(S_n|S_{n-1})$. Here, we suppose that $\{S_n, n \geq 1\}$ is a finite homogeneous Markov chain with the transition matrix Φ and the initial probability distribution p. An important feature in the inference algorithms for HMMs is the prediction filter $p_n = P(S_n|Y_{n-1}, \ldots, Y_1)$, $n \geq 2$, and $p_1 = p$. Under appropriate hypotheses, we present a new approach based on the Ionescu Tulcea - Marinescu ergodic Theorem [8] for proving the geometric ergodicity of the Markov chain $\{p_n, n \geq 1\}$.

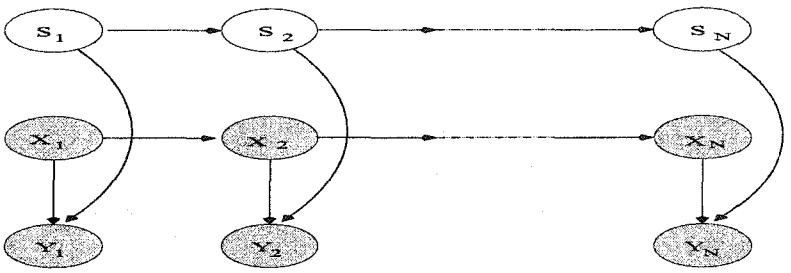

Figure 1.3: Model with hidden Markov switching

We then study the ergodic properties of the model represented by the Bayesian network displayed in Fig. 1.3. The model is more complex and it is formed with the observed sequences $\{Y_n, n \geq 1\}$ and $\{X_n, n \geq 1\}$, and the hidden, finite, homogeneous Markov chain $\{S_n, n \geq 1\}$. We suppose that both $\{Y_n, n \geq 1\}$ and $\{X_n, n \geq 1\}$ have continuous distributions, and $\{X_n, n \geq 1\}$ is a Markov chain. Now, the corresponding prediction filter is given by $w_n = P(S_n|Y_{n-1}, X_{n-1} \ldots, Y_1, X_1)$, $n \geq 2$ and $w_1 = p$. We associate a random system with complete connections (RSCC) [9] with this model. Based on the properties of the RSCC, we show that the Markov chain $\{(w_{n+1}, X_n), n \geq 1\}$ is geometrically ergodic. Then we study the ergodic properties of the model with misspecified parameters, and we prove the exponential forgetting of the initial conditions.

In Chapter 5, we present a nested Monte Carlo EM (NMCEM) algorithm [22] for

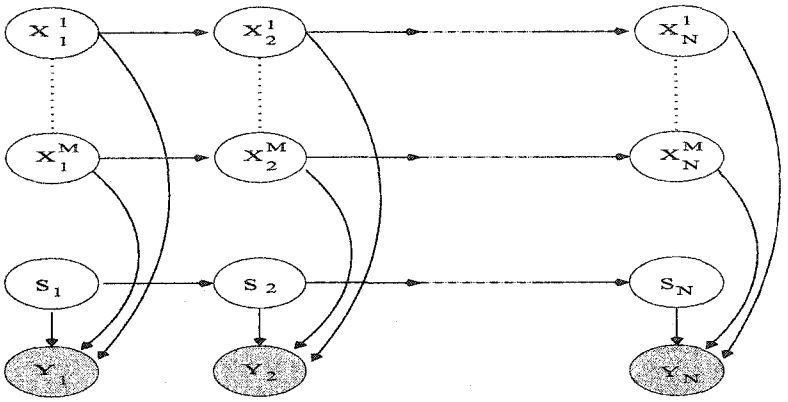

Figure 1.4: Switching State-Space Model

learning the parameters of the switching state-space model (SSM) introduced in [5]. Fig. 1.4. displays the Bayesian network corresponding to the SSM. Since it combine the state-space models and the HMMs, the SSM is regarded as a hybrid model. It can be viewed as an extension of the mixture of experts [10] or the switching regression models [6]. It is also closely related to models arising from economics ([7], [12]). Here, $\{Y_n, n \geq 1\}$ are the observations, $\{X_n^m, n \geq 1\}$, $m = 1, \ldots, M$, are the M sequences of the hidden state variables, and $\{S_n, n \geq 1\}$ is the hidden, finite, homogeneous Markov chain. The NMCEM algorithm is based on the multi-move Gibbs sampler, but the convergence rate is improved by introducing nesting and the Rao-Blackwellised forms [29]. The performance of the algorithm is illustrated for experimental medical data.

Bibliography

[1] P. Baldi and S. Brunak. *Bioinformatics, the Machine Learning Approach.* MIT Press, 1998.

[2] Y. Bengio, Y. LeCun, C. Nohl, and C. Burges. Lerec: A NM/HMM hybrid for on-line handwriting recognition. *Neural Computation*, 7(5):1289–1303, 1995.

[3] C. M. Denegri Jr. Limit cycle oscillation flight test results of a fighter with external stores. *J. of Aircraft*, 37(5):761–769, 2000.

[4] Z. Ghahramani. An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1):9–42, 2001.

[5] Z. Ghahramani and G.E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–996, 2000.

[6] S.M. Goldfeld and R.E. Quandt. A Markov model for switching regression. *J. Econometrics*, 1:3–16, 1973.

[7] J.D. Hamilton. Analysis of time series subject to changes in regime. *J. Econometrics*, 45:37–70, 1990.

[8] H. Hennion. Sur un théoréme spectral et son application aux noyaux lipchitziens. *Proc. Amer. Math. Soc.*, 118(2):627–634, 1993.

[9] M. Iosifescu and S. Grigorescu. *Dependence with Complete Connections and Its Applications.* Cambridge University Press, Cambridge, 1990.

[10] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixture of local experts. *Neural Computation*, 3:79–87, 1991.

[11] S.J. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte. A new approach for filtering nonlinear systems. In *American Control Conference, Seattle, Washington*, pages 1628–1632, 1995.

[12] C.J. Kim. Dynamic linear models with Markov-switching. *J. Econometrics*, 60:1–22, 1994.

[13] J. Ko, A. J. Kurdila, and T. W. Strganac. Nonlinear control of a prototypical wing section with torsional nonlinearity. *AIAA J. Guid.,Contr., Dynamics*, 20(6):1181–1189, 1997.

[14] J. Ko, T. W. Strganac, and A. J. Kurdila. Stability and control of a structurally nonlinear aeroelastic system. *AIAA J. Guid.,Contr., Dynamics*, 21(5):718–725, 1998.

[15] A. J. Kurdila, T. W. Strganac, J.L. Junkins, J. Ko, and M. R. Akella. Nonlinear control methods for high-energy limit-cycle oscillations. *AIAA J. Guid.,Contr., Dynamics*, 24(1):185–192, 2001.

[16] B. H. K. Lee, S. J. Price, and Y. S. Wong. Nonlinear aeroelastic analysis of airfoils: bifurcation and chaos. *Progress in Aerospace Sciences*, 35:205–334, 1999.

[17] L. Liu, Y. S. Wong, and B. H. K Lee. Application of the centre manifold theory in nonlinear aeroelasticity. *J. Sound and Vibration*, 234(4):641–659, 2000.

[18] C. A. Popescu and Y. S. Wong. Applications of the EM algorithm for the study of the aeroelastic systems with structural nonlinearities. In *Proceedings of the 3rd International Workshop on Scientific Computing*, Hong Kong, 2003. Nova Science Publisher Inc.

[19] C. A. Popescu and Y. S. Wong. A nonlinear statistical approach for aeroelastic response prediction. *AIAA J. Guid.,Contr., Dynamics*, 26(4):565–572, 2003.

[20] C. A. Popescu and Y. S. Wong. The Unscented and Extended Kalman filter for systems with polynomial restoring forces. In *Proceedings of the 44th AIAA/ ASME/ ASCE/ AHS/ ASC/ Structures, Structural Dynamics, and Materials Conference*, Norfolk, VA, USA, 2003. AIAA Paper 2003-1410.

[21] C. A. Popescu and Y. S. Wong. Ergodicity in models with hidden Markov switching. submitted to Math. Control Signals Systems, 2004.

[22] C. A. Popescu and Y. S. Wong. Monte Carlo approach for switching state-space models. In C. Orchard, B. Yang and M. Ali, editors, *Proceedings of the 17th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems*, volume 3029 of *Lecture Notes in Artificial Intelligence*, pages 945–955, Ottawa, 2004. Springer-Verlag.

[23] C. A. Popescu and Y. S. Wong. Nested Monte Carlo EM algorithm for switching state-space models. submitted to IEEE Tran. on Knowledge and Data Engineering, 2004.

[24] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceeding of the IEEE*, volume 77, pages 257–286, 1989.

[25] P. Smith. Hidden Markov models for fault detection in dynamical systems. *Pattern Recognition*, 27(1):149–164, 1994.

[26] D. M. Tang and E. H. Dowell. Comparison of theory and experiment for nonlinear flutter and stall response of a helicopter blade. *J. of Sound and Vibration*, 165(2):251–276, 1993.

[27] D. M. Tang, E. H. Dowell, and L. N. Virgin. Limit cycle behavior of an airfoil with a control surface. *J. Fluids and Structures*, 2:839–858, 1998.

[28] D. M. Tang, J. K. Henry, and E. H. Dowell. Limit cycle oscillations of delta wing models in low subsonic flow. *AIAA J.*, 37(11):1355–1362, 1999.

[29] D.A. Van Dyk. Nesting EM algorithms for computational efficiency. *Statist. Sinica*, 10:203–225, 2000.

Chapter 2

Nonlinear Time Series Models

2.1 Introduction

In recent years, there has been an increasing interest in the studies of nonlinear dynamics across broad ranges of disciplines. The main reason is that nonlinear dynamical systems could be used to model many complex phenomena, such as the jump discontinuity, amplitude dependent frequency profile, limit cycle oscillation, sub-harmonic motion, and even chaos. In particular, the ability to predict the response of a dynamical system is essential in controlling a physical system. However, if observation data are the only available information in a given application, then it is important to construct a time-series model which is capable of characterizing the nonlinear dynamics.

In this chapter, we are concerned with the dynamics of a nonlinear aeroelastic system. The corresponding data are obtained either from numerical simulations or actual wind-tunnel experiments. One of the most important aspects in the study of nonlinear aeroelasticity is the appearance of limit cycle oscillations, which are characterized as sustained periodic oscillations with neither increasing nor decreasing amplitude over time for a given flight condition. It should be noted that limit cycle

The material presented in this chapter was previously published in [18], [19], [27].

oscillations are an undesirable feature, since they can cause structural fatigue, and the vibrations can also induce pilot fatigue. Hence, an accurate prediction of limit cycle oscillations is of great interest to the aerospace industry. Another related topic is the prediction of oscillations in nonlinear flight dynamics. When information on possible unstable oscillations is provided to the pilot, certain controls may be activated in order to ensure a safe cruising performance.

The general form of an linear autoregressive (AR) model of order p is given by

$$X_n = \sum_{i=1}^{p} a_i X_{n-i} + e_n, \quad n > p.$$

Here, p is a positive integer, a_i, $i = 1, \ldots, p$ are the model coefficients, and e_n, $n \geq p+1$ are independent, identically distributed Gaussian variables with zero mean. A linear moving average (MA) model of order q is given in the following equation

$$X_n = \sum_{i=1}^{q} b_i e_{n-i} + e_n, \quad n > q,$$

where q is a positive integer and b_i, $i = 1, \ldots, p$ are the model coefficients. Combining these two models we get an autoregressive moving average (ARMA) model of order (p, q):

$$X_n = \sum_{i=1}^{p} a_i X_{n-i} + \sum_{i=1}^{q} b_i e_{n-i} + e_n, \quad n > q, \quad n > p.$$

Linear time series models [21], such as ARMA models, are often used for system identification of the aeroelastic system and flutter prediction. Torii and Matsuzaki ([24], [25]) regard the response of the wing excited by the turbulence of an airstream as a random vibration, and use an ARMA model for the system identification. The flutter prediction parameter is calculated using the estimated AR coefficients. The results are validated by numerical simulations under non-stationary conditions and by experiments performed during the wind-tunnel flutter testing.

The ARMA models can be generalized to auto-regressive moving average with

exogenous variable (ARMAX) models:

$$X_n = Au_n + \sum_{i=1}^{p} a_i X_{n-i} + \sum_{i=1}^{q} b_i e_{n-i} + e_n, \quad n > q, \quad n > p,$$

where u_n is a vector of inputs (the exogenous variable). Andrighettoni and Mantegazza [1] consider an ARMAX model for the identification of the model to be used in designing an adaptive flutter suppression system for a built-in wing model fitted with a leading- and trailing-edge control surface and two accelerometers.

However, the ARMA or ARMAX are linear models. Since the nonlinearities due to the aerodynamics or the structure of an aircraft can critically affect the aeroelastic behavior, it is therefore desirable to develop models which are capable of incorporating information about the nonlinearities of the aeroelastic phenomena. Two nonlinear time series models are being considered, namely the amplitude-dependent exponential auto-regressive models (EXPAR) [10], and the self-exciting autoregressive models (SETAR) [22]. The EXPAR models are suitable for polynomial structural nonlinearities, and the SETAR models include the threshold structure specific to freeplay or hysteresis.

The main difficulty associated with nonlinear time series models is the estimation of the model parameters, in which a complex nonlinear solver is usually needed. In the present EXPAR and SETAR models, we do not require any direct application of a nonlinear procedure because the problem is reformulated to the parameter estimation for AR models. Consequently, it can be solved by the well-known singular value decomposition method. Moreover, the selection of the best model can be done in a similar way to the selection of the best ARMA model using the Akaike information criterion (AIC) ([10]).

Another important aspect of the proposed approach is the capability to handle input data corrupted with noise. In [8], Dimitriadis and Cooper notice that the performances of ARMA and ARMAX models are very sensitive to the measurements errors. Tests were carried out at the noise levels of 5, 10, 15 and 20%, and the mean flutter prediction using ARMA-based methods was found to be acceptable

only for the case with 5% noise level. To overcome this difficulty, we have implemented wavelet de-noising methods as preprocessing. Consequently, the parameter estimating procedures are applied to the denoised data instead of the original noisy input data.

2.2 EXPAR models

The EXPAR models incorporate both the amplitude-dependent frequency and the limit cycle behavior. The basic form of an EXPAR model of order p is given by

$$X_n = (\Phi_1 + \pi_1 e^{-\gamma X_{n-1}^2})X_{n-1} + \ldots + (\Phi_p + \pi_p e^{-\gamma X_{n-1}^2})X_{n-p} + e_n, \tag{2.1}$$

where Φ_i, π_i, $i = 1, \ldots p$ and γ are constants and e_n, $n \geq p+1$ are independent, identically distributed Gaussian variables with zero mean. Such a model implies that X_n is a symmetric process, although X_n is not constrained to being Gaussian.

Let $Y_n = (X_n, \ldots, X_{n-p+1})^t$, $E_n = (e_n, 0, \ldots, 0)^t$, $f(Y_{n-1}) = AY_{n-1} + F(Y_{n-1})$,

$$A = \begin{pmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_{p-1} & \Phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}, F(Y_{n-1}) = e^{-\gamma X_{n-1}^2} \begin{pmatrix} \pi_1 X_{n-1} + \cdots + \pi_p X_{n-p} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Here ^{t} denotes the transpose of a matrix or a vector. Notice that we have

$$Y_n = f(Y_{n-1}) + E_n. \tag{2.2}$$

From (2.2), we can easily see that Y_n is a homogeneous Markov chain, and without the noise term E_n,

$$Y_n = f(Y_{n-1}) \tag{2.3}$$

is an autonomous deterministic difference equation. The equilibrium states (i.e. the

solutions of $f(Y) = Y$) are $\bar{Y}_1 = (0, \ldots, 0)^t$, and

$$\bar{Y}_2 = \left(\pm\sqrt{-\frac{1}{\gamma}\ln\left(\frac{1-\sum_{i=1}^{p}\Phi_i}{\sum_{i=1}^{p}\pi_i}\right)}, \ldots, \pm\sqrt{-\frac{1}{\gamma}\ln\left(\frac{1-\sum_{i=1}^{p}\Phi_i}{\sum_{i=1}^{p}\pi_i}\right)} \right)^t,$$

if $0 < (1 - \sum_{i=1}^{p}\Phi_i) / (\sum_{i=1}^{p}\pi_i) < 1$. Now we recall the following classical definition (see, e.g. [9], pp. 166)

Definition 2.1 The equilibrium state $\bar{Y}$ is

1. Stable, if for any given $\epsilon > 0$, there exists $\delta > 0$ such that for any $\|Y_1 - \bar{Y}\| < \delta$, we have $\|Y_n - \bar{Y}\| < \epsilon$ for the solution Y_n;

2. Attractive, if there exists $\delta > 0$ such that, for any $\|Y_1 - \bar{Y}\| < \delta$, we have

$$\lim_{n\to\infty} Y_n = \bar{Y};$$

3. Asymptotically stable, if it is stable and attractive;

4. Exponentially stable, if there exists $\delta > 0$, $K > 0$, and $0 \leq c < 1$, such that if $\|Y_1 - \bar{Y}\| < \delta$, then

$$\|Y_n - \bar{Y}\| \leq K\|Y_1 - \bar{Y}\|c^n.$$

The stability condition for the equilibrium states of (2.3) is given as follows.

Theorem 2.1 *1. The zero solution $\bar{Y}_1$ is exponentially asymptotically stable, if all the solutions of the equation*

$$\Lambda^p - (\Phi_1 + \pi_1)\Lambda^{p-1} - \cdots - \Phi_p - \pi_p = 0 \tag{2.4}$$

are inside the unit disc.

2. The non-null solution $\bar{Y}_2$ is exponentially asymptotically stable, if all the solutions of the equation

$$\Lambda^p - h_1\Lambda^{p-1} - \cdots - h_p = 0 \tag{2.5}$$

are inside the unit disc,where

$$h_1 = \left(\pi_1 + \Phi_1 \sum_{j=1}^{p} \pi_j - \pi_1 \sum_{j=1}^{p} \Phi_j\right) / \sum_{j=1}^{p} \pi_j + 2\left(1 - \sum_{j=1}^{p} \Phi_j\right) \ln\left[(1 - \sum_{j=1}^{p} \Phi_j)/ \sum_{j=1}^{p} \pi_j\right]$$

$$h_i = \left(\pi_i + \Phi_i \sum_{j=1}^{p} \pi_j - \pi_i \sum_{j=1}^{p} \Phi_j\right) / \sum_{j=1}^{p} \pi_j \qquad (i = 2, 3, \ldots, p).$$

PROOF. Elementary calculations give the Jacobian matrix $\partial f/\partial Y(Y_{n-1}) = A + H(Y_{n-1})$, where:

$$H(Y_{n-1}) = e^{-\gamma X_{n-1}^2} \begin{pmatrix} (\pi_1 - 2\gamma \sum_{i=1}^{p} \pi_i X_{n-1} X_{n-i}) & \ldots & \pi_{p-1} & \pi_p \\ 0 & \ldots & 0 & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & \ldots & 0 & 0 \end{pmatrix}.$$

It is easy to verify that the equation (2.4) gives the eigenvalues of $\partial f/\partial Y(\bar{Y}_1)$, and equation (2.5) gives the eigenvalues of $\partial f/\partial Y(\bar{Y}_2)$. Hence, the conclusion of the theorem is a consequence of Corollary 4.7.2 in [12]. □

Remark 2.1 The same result is reported as Condition A in [16] using a slightly different terminology.

A limit cycle of (2.3) is a periodic solution with period $q > 1$. Analogously with Definition 2.1 we have

Definition 2.2 A limit cycle $\bar{Y}_n$, $n \geq n_0 > p$ is asymptotically stable if the orbit $\Gamma = \{\bar{Y}_{n_0+1}, \ldots, \bar{Y}_{n_0+q}\}$, is asymptotically stable; that is

1. for any neighborhood $\mathcal{N}$ of Γ, there exists a neighborhood U of Γ such that, if $Y_1 \in U$ then the solution $Y_n \in \mathcal{N}$, for n large enough;

2. there exists a neighborhood $\mathcal{N}$ of Γ such that, if $Y_1 \in \mathcal{N}$, then for the solution

Y_n we have

$$\lim_{n\to\infty} d(Y_n, \Gamma) = 0.$$

We have the following theorem concerning the stability of a limit cycle of (2.3) (see Proposition 2.2 in [17]).

Theorem 2.2 *A limit cycle of period q, $\{\bar{Y}_{n_0+1}, \ldots, \bar{Y}_{n_0+q}\}$, $n_0 > p$, of the model (2.3) is asymptotically stable when all the eigenvalues of the matrix,*

$$M = \partial f/\partial Y(\bar{Y}_{n_0+q}) \cdot \partial f/\partial Y(\bar{Y}_{n_0+q-1}) \cdots \partial f/\partial Y(\bar{Y}_{n_0+1})$$

are inside the unit disk.

PROOF. To study what happens to the solutions in a neighborhood of $\Gamma = \{\bar{Y}_{n_0+1}, \ldots, \bar{Y}_{n_0+q}\}$, we introduce the variable $Z_n = Y_n - \bar{Y}_n$, $n > n_0 + 1$. Then, we have

$$Z_n = f(Y_{n-1}) - f(\bar{Y}_{n-1}) = \partial f/\partial Y(\bar{Y}_{n-1}) Z_{n-1} + O(\|Z_{n-1}\|^2).$$

Hence, we can associate the following difference equation with periodic coefficients

$$Z_n = \partial f/\partial Y(\bar{Y}_{n-1}) Z_{n-1}, n > n_0 + 1. \tag{2.6}$$

The zero solution of the equation (2.6) is asymptotically stable if the eigenvalues of the matrix M are inside the unit disk (Theorem 4.4.1, in [12]). This proves the theorem, since the limit cycle is asymptotically stable if $Z_n \to 0$ as $n \to \infty$. □

The conditions in Theorems 2.1 and 2.2 can replace the usual stability conditions for the ARMA models, and they can be used to generalize the approach in [25] for finding the flutter margin.

For complex dynamic predictions, the EXPAR model (2.1) can be extended [16] to the following form

$$X_n = (\Phi_1 + f_1(X_{n-1})e^{-\gamma X_{n-1}^2})X_{n-1} + \ldots + (\Phi_p + f_p(X_{n-1})e^{-\gamma X_{n-1}^2})X_{n-p} + e_n \tag{2.7}$$

where $f_i(X_{n-1}), \quad i = 1, \ldots, p$, are the polynomials:

$$f_i(X_{n-1}) = \pi_0^{(i)} + \pi_1^{(i)} X_{n-1} + \ldots + \pi_{r_i}^{(i)} X_{n-1}^{r_i}. \tag{2.8}$$

The model given in (2.7) admits more sophisticated nonlinear dynamics. For example, if the order r_i of the polynomials $f_i(X_{n-1})$ are odd, then X_n is not constrained to being a symmetric process.

We now briefly present an efficient procedure for estimating the coefficients of the EXPAR model (2.7). In general, the estimation of the order p, and the coefficients $\{\gamma, (\Phi_i, \pi_j^{(i)}; j = 0, \ldots, r_i, i = 1, \ldots, p)\}$ requires a nonlinear optimization procedure. However, this optimization problem can be reduced to fitting a linear regression [10]. For $n = m+1, \ldots, N$, we rewrite (2.7) as

$$X^{(N)} = A\beta + \mathbf{e}, \tag{2.9}$$

where

$$\begin{aligned}
&X^{(n)} = (X_n, X_{n-1}, \ldots, X_{n-(N-m-1)})^t, \quad \mathbf{e} = [e_N, e_{N-1}, \ldots, e_{m+1}]^t, \\
&Y_i^{(n)} = (X_n e^{-\gamma X_{N-1}^2} X_{N-1}^i, X_{n-1} e^{-\gamma X_{N-2}^2} X_{N-2}^i, \ldots, X_{n-(N-m-1)} e^{-\gamma X_m^2} X_m^i)^t. \\
&\beta = (\Phi_1, \pi_0^{(1)}, \ldots, \pi_{r_1}^{(1)}, \ldots, \Phi_p, \pi_0^{(p)}, \ldots, \pi_{r_p}^{(p)})^t, \\
&A = [X^{(N-1)}, Y_0^{(N-1)}, \ldots, Y_{r_1}^{(N-1)}, \ldots, X^{(N-p)}, Y_0^{(N-p)}, \ldots, Y_{r_p}^{(N-p)}].
\end{aligned}$$

First, the parameter γ is selected from a grid in a range such that $e^{-\gamma X_{n-1}^2}$ is different from both 0 and 1 for most values of X_{n-1}. For each value of γ, β is estimated using a singular value decomposition in (2.9). The choice of the maximum order m is subjective and it depends on the sample size. The order p of the fitted model is determined using the AIC criterion [10], i.e.,

$$AIC(p) = (N-m) \ln \hat{\sigma}_p^2 + 2(2p + \sum_{i=1}^{p} r_i + 1), \tag{2.10}$$

where $\hat{\sigma}_p^2$ is the least square estimate of the residual variance of the model. The last

term in (2.10) represents the double of the number of the estimated parameters in the model, including the fitted mean. For each γ, the fitted models are compared using the AIC criterion, and the model with the smallest AIC is chosen.

The complexity of the singular value decomposition method is proportional to

$$(N-m)(2p+\sum_{i=1}^{p} r_i)^2 + (2p+\sum_{i=1}^{p} r_i)^3.$$

Thus, the estimation method is computationally efficient for models of reasonably large order.

A further improvement in terms of parameters estimation can be achieved by replacing $\pi_0^{(i)}$ by $\pi_0^{(i)} \exp(-\gamma\mu_2)$, where $\mu_2 = E[X_{n-1}^2]$ is estimated by the sample second moment ([22], page 328).

To conclude this section, we mention the following modified EXPAR models:

$$X_n = (\Phi_1 + f_1(X_{n-1})e^{-\gamma X_{n-1}^2})X_{n-1} + \ldots + (\Phi_p + f_p(X_{n-p})e^{-\gamma X_{n-p}^2})X_{n-p} + e_n,$$

where f_i, $i = 1, \ldots, p$ are given in (2.8). Proceeding as in Example 4.3, page 129 in [22], we can easily prove that the Markov chain formed with $Y_n = (X_n, \ldots, X_{n-p+1})^t$ is ergodic, provided that all the solutions of the equation

$$\Lambda^p - \Phi_1\Lambda^{p-1} - \cdots - \Phi_p = 0$$

are inside the unit disc.

2.3 SETAR models

The essential idea underlying the SETAR models is a piece-wise linearisation of the nonlinear models over the state space with the introduction of thresholds.

Let $\{t_0, t_1, \ldots, t_l\}$ denote the thresholds, i.e. a linearly ordered subset of real numbers, such that $t_0 < t_1 < \ldots < t_l$, where $t_0 = -\infty$ and $t_l = +\infty$. A self-exciting threshold autoregressive model of order $(l; p, \ldots, p)$ or SETAR $(l; p, \ldots, p)$ where p

is repeated l times, is a univariate time series $\{X_n\}$ of the form

$$X_n = a_0^{(j)} + \sum_{i=1}^{p} a_i^{(j)} X_{n-i} + e_n, \quad t_{j-1} < X_{n-d} \leq t_j, \tag{2.11}$$

for $j = 1, 2, \ldots l$, where d is a fixed integer belonging to $\{ 1, 2, \ldots, p\}$, and $\{e_n\}$ is a Gaussian, independent, identically distributed white noise sequence. If for $j = 1, 2, \ldots, l$, we have $a_i^{(j)} = 0$ for $i = p_j + 1, p_j + 2, \ldots, p$, then $\{X_n\}$ is known as a SETAR$(l; p_1, p_2, \ldots, p_l)$ model. Hence, a SETAR $(1, p)$ model is equivalent to a linear autoregressive (AR) model of order p. To specify the terminology for the rest of this thesis, we give the following definition:

Definition 2.3 A Markov chain (Y_n) defined on the probability space (Ω, Σ, P) with values in $(E, \mathcal{E})$ is

1. ergodic, if there exists a probability p on $(E, \mathcal{E})$ such that, for all $y \in E$,

$$\lim_{n \to \infty} \|P^n(y, \cdot) - p(\cdot)\| = 0.$$

2. geometric ergodic, if there exists a probability p on $(E, \mathcal{E})$, a constant $0 \leq r < 1$ and a function M defined on E with positive real values, such that

$$\int_E |M(x)| p(dx) < \infty, \quad \|P^n(y, \cdot) - p(\cdot)\| \leq M(y) r^n,$$

for any positive integer n, and all $y \in E$.

Here, $P^{n+1}(y, A) = P(Y_n \in A | Y_1 = y), \quad A \in \mathcal{E}$, and for any signed measure μ on $(E, \mathcal{E})$, $\|\mu\|$ denotes the total variation:

$$\|\mu\| = \mu(E^+) - \mu(E^-),$$

where $E = E^+ \cup E^-$ is a Hahn decomposition of E with respect to μ.

The ergodicity implies the existence of a unique stationary measure. It has also many applications in statistical inference. For example, we can get asymptotic

results, such as versions of the central limit theorem or the law of large numbers.

Let $Y_n = (X_n, \ldots, X_{n-p+1})^t$. Then, we have $Y_n = f(Y_{n-1}) + E_n$, where $E_n = (e_n, 0, \ldots, 0)^t$, $f(Y_{n-1}) = (h(Y_{n-1}), X_{n-1}, \ldots, X_{n-p+1})^t$, and $h(Y_{n-1})$ is given in the right side of the equation (2.11). Obviously (Y_n) is a homogeneous Markov chain. Moreover, we have

Theorem 2.3 *The Markov chain (Y_n) is geometrically ergodic if*

$$\max_j \left(\sum_{i=1}^{p} |a_i^{(j)}| \right) < 1.$$

PROOF. See Lemma 3.1 in [2].

We now briefly describe the parameter estimation [23] for a SETAR$(2; p_1, p_2)$ model. First, let d and m be predefined, where m is the maximum regression order of the two linear AR models, and let n_0 be the maximum of d and m. The choice of m is subjective and usually depends on the sample size. To find the threshold t_1, we try some of the sample quantiles - e.g. $\{Q_{0.30}, Q_{0.40}, Q_{0.50}, Q_{0.60}, Q_{0.70}\}$ -, where by definition, for any $0 < q < 1$, exactly $100q\%$ of the data are less than Q_q. For each choice of t_1, the data set is re-arranged into two subsets and two sub-systems of linear autoregressive equations are set up. The first subset contains the observations less than or equal to t_1, and the second subset contains the observations greater than t_1. The coefficients are then estimated using a singular value decomposition for each of the corresponding matrices. For each value of t_1 and d, we apply Akaike's Information Criterion to determine p_1 and p_2, the orders of the two linear AR's. Specifically, if N_1 is the number of observations less than or equal to t_1 and $\hat{\sigma}_1(p_1)$ is their residual sum of squares, then the estimated $\hat{p_1}$ corresponds to the minimum $AIC(p_1)$, where

$$AIC(p_1) = N_1 \ln(\hat{\sigma}_1(p_1)/N_1) + 2(p_1 + 1).$$

The most difficult task for these models is to find the thresholds and the delay parameter d. Apart from trying some of the sample quantiles for finding the thresholds, we have also used some exploratory data analysis. The method presented in [22], Chapter 7.2.3, suggests to study the non-parametric lag regression estimates.

Let $m_j(x) = E(X_n|\ X_{n+j})$. A non-parametric kernel estimate $\hat{m}_j(x)$ for $m_j(x)$ is given by

$$\hat{m}_j(x) = \sum_{i=-j+1}^{N} X_i \delta_N(x - X_{i+j}) / \sum_{i=-j+1}^{N} \delta_N(x - X_{i+j}), \tag{2.12}$$

for $j = -p, \ldots, -1$, where p is a positive integer much smaller than the size N of the data set (see page 218 in [22] and Chapter 4.1.5 in [26]). Here, $\delta_N(\cdot)$ is a function defined by

$$\delta_N(z) = \begin{cases} (1 - |z|/h_N)/h_N & \text{if } |z| \leq h_N \\ 0 & \text{otherwise,} \end{cases}$$

where h_N is chosen such that $h_N \to 0$ as $N \to \infty$. A similar formula can be written for the non-parametric estimates $\hat{v}_j(x)$ of the variance $v_j(x) = VAR(X_n|X_{n+j})$. Analysing the plots of $\hat{m}_j(x)$ and $\hat{v}_j(x)$ for several values of j, we can determine the thresholds and the delay parameter. In addition to the simple kernels $\delta_N(\cdot)$, there are also other possible choices for the kernels (see [26], page 139), e.g the Gaussian or the Epanechnikov kernels. The bandwidth h_N can be chosen using the leave-on-out cross-validation ([26], page 141). Asymptotic properties of $\hat{m}_j(x)$ are given in [20].

To predict the aeroelastic response of a system with freeplay, we use a combination of a SETAR and an EXPAR model. To construct this model, we replace the linear autoregressive models in (2.11) with the EXPAR models given by (2.7). The parameters are estimated combining the algorithms for EXPAR and SETAR models.

2.4 Applications

In this section we present the results obtained using EXPAR models and combined EXPAR and SETAR models. We consider three classes of data

1. simulated data from mathematical functions;

2. numerically generated aeroelastic data corresponding to polynomial and freeplay structural nonlinearities;

3. experimental aeroelastic data recorded during wind tunnel tests.

To measure the accuracy of a long-term prediction, the available input data are divided into two subsets: the training set and the test set. The training set is used to estimate the parameters of the models to be used for predictions. The test set is used for checking the accuracy of the predictions. The performance assessment is based on the study of the residuals on both the training and the test sets. An advantage of this method is that it emphasizes the predictive aspect of the model selection.

2.4.1 Preprocessing

Generally speaking, all real data are contaminated by noise. However, the noise effect may vary. For instance, the noise content from a typical ground vibration test in aerospace industry is usually small and the majority is caused by the measurement noise. On the other hand, in a flight flutter test, the amount of noise corruption due to turbulence is often significant. Since the nonlinear time series models are very sensitivity to noise, it is important to reduce the noise effects. In this study, we apply a wavelet filtering[1]. ([14], Chapter 10), which uses a transform-based thresholding working in the following steps:

1. Transform the noisy data into an orthogonal domain.

2. Apply soft or hard thresholding to the resulting coefficients, thereby suppressing those coefficients smaller than a certain amplitude.

3. Transform back into the original domain.

The wavelet transform is based on a multiresolution analysis [7]. By multiresolution, a wavelet transform can be organized as a ladder of component stages, such that each involves simply the application of digital filtering to certain discrete time "signals". Consequently, this leads to a fast and efficient orthogonal transform of order $O(N)$.

De-noising using hard and soft thresholding and orthogonal maximally decimated wavelets transforms can cause certain visual artifacts, some of them due to the lack

[1]Our C++ implementation of the wavelets filters is based on the Matlab programs of WAVELAB, available online at http://www-stat.stanford.edu/ wavelab/.

of translation invariance of the wavelet basis. To overcome these difficulties, we apply translation invariant transformations [3]: the noisy signal is shifted, then de-noised with wavelet thresholding and finally unshifted. The implementation of this method over the range of all circulant shifts is of order $O(N \log_2 N)$.

Wavelet bases are not well-suited to representing signals containing sinusoidal oscillations of moderate duration. Coifman and Meyer [4] have proposed local cosine bases and Coifman, Meyer, and Wickerhauser [5] have introduced wavelet packets. There are many orthogonal bases, a method of selecting the best basis that minimizes a certain measure of the entropy is reported in [6] and [14], Chapter 8.

In addition to the de-noising step, the input data is transformed to the interval $[-1, 1]$, and we work with the mean deleted time series. Moreover, we apply the standard linearity, stationarity and gaussianity tests to the input data ([22], Chapter 5).

2.4.2 Simulated data

Data from a typical flight test usually include responses from more than one mode of vibrations, and they can be expressed as a combination of exponential and sine functions. In Fig. 2.1 we display three discrete signals simulated according to the formula

$$X_t = \mathrm{e}^{-0.4t} \sin\left(2\pi t + 0.1 \sin\left(6\pi t\right)\right) + \mathrm{e}^{\alpha t} \sin\left(10\pi t + 0.1 \sin\left(4\pi t\right)\right), \qquad (2.13)$$

where the parameter $\alpha = -1, 0, 1$ and the step $\delta t = 2\pi/128$.

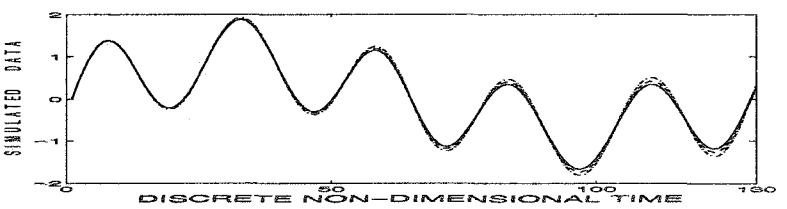

Figure 2.1: Simulated data:'-'($\alpha = -1$), '- -'($\alpha = 0$),'-.-'($\alpha = +1$)

It is clear that the differences between these data sets are small during the

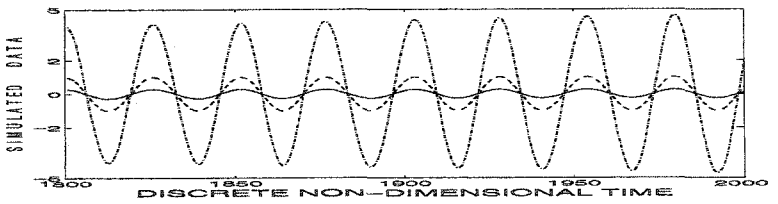

Figure 2.2: Last 200 simulated data:'-'($\alpha = -1$), '- -'($\alpha = 0$),'-.-'($\alpha = +1$)

initial states (i.e. the first 130 data points). However, their asymptotic behaviors are quite different as shown in Fig. 2.2, and they correspond to a damped oscillation ($\alpha = -1$), a limit cycle oscillation ($\alpha = 0$) and a divergent oscillation ($\alpha = +1$).

The EXPAR models have been successfully applied to study nonlinear random vibrations ([15]). Hence, it is natural to fit EXPAR models for this class of simulated data. Indeed, using only the first 130 points in Fig. 2.1 to estimate the parameters for the EXPAR model (2.7), we obtain a very accurate prediction of the nonlinear response. In Figs. 2.3-2.5, we compare the simulated and the predicted signals for the last 500 simulated data, from 1500 to 2000, for $\alpha = -1, 0$ and 1, respectively. Despite the fact that it is difficult to guess the asymptotic state looking only at the first 130 data points in Fig. 2.1, not only we are able to correctly classify the oscillations as damped, limit cycle and divergent, but we also actually achieve excellent long term predictions. The structural parameters used in equation (2.7) are reported in Table 2.1.

α	-1	0	1
p	28	18	28
γ	18.02	3.11	9.08
$\max_{1\le i\le p} r_i$	0	1	0

Table 2.1: EXPAR model parameters for simulated data

To test the proposed nonlinear time series models with more realistic data, we introduce an additive random Gaussian white noise in equation (2.13), such that the signal to noise ratio is 5, and we generate data according to the formula:

$$X_t = \mathrm{e}^{-0.4t}\sin(2\pi t + 0.1\sin(6\pi t)) + \mathrm{e}^{\alpha t}\sin(10\pi t + 0.1\sin(4\pi t)) + e_t, \qquad (2.14)$$

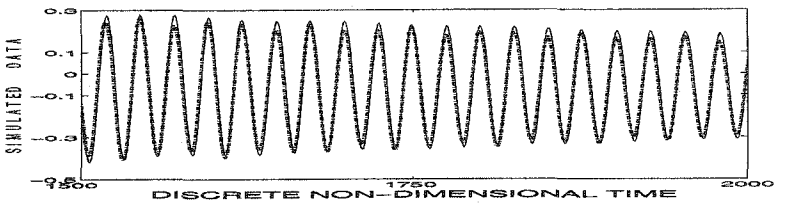

Figure 2.3: EXPAR prediction '- -' and simulated damped oscillation '-'($\alpha = -1$)

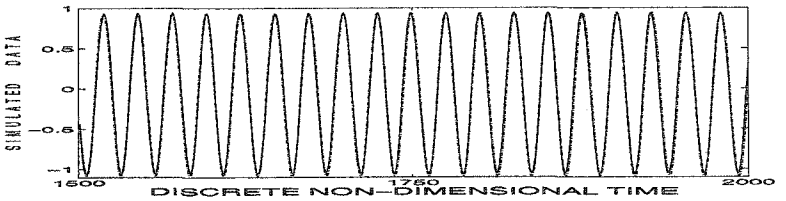

Figure 2.4: EXPAR prediction '- -' and simulated limit cycle oscillation '-'($\alpha = 0$)

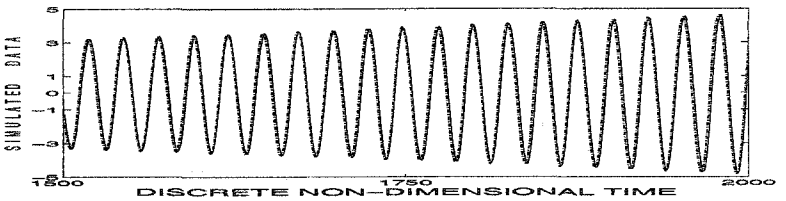

Figure 2.5: EXPAR prediction '- -' and simulated divergent oscillation '-'($\alpha = 1$)

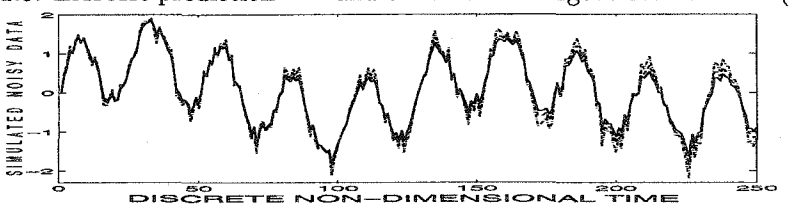

Figure 2.6: Simulated noisy data set:'-'($\alpha = -1$), '- -'($\alpha = 0$),'-.-'($\alpha = +1$)

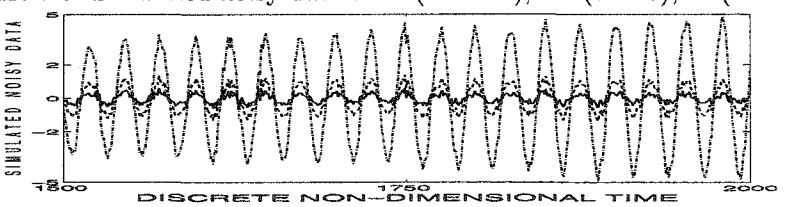

Figure 2.7: Last 200 simulated noisy data:'-'($\alpha = -1$), '- -'($\alpha = 0$),'-.-'($\alpha = +1$)

where again the parameter $\alpha = -1, 0, 1$ and the step $\delta t = 2\pi/128$. The differences between these noisy data sets are small during the first 250 data points, but the asymptotic states are very different (see Figs. 2.6 and 2.7).

The EXPAR model (2.7) fails to estimate the parameters when taking the noisy data directly as training set. However, to overcome this difficulty, a de-noising based on a local cosine base wavelet package is first applied to the noisy data. For the limit cycle oscillation ($\alpha = 0$), this leads to smooth initial data as illustrated in Fig. 2.8.

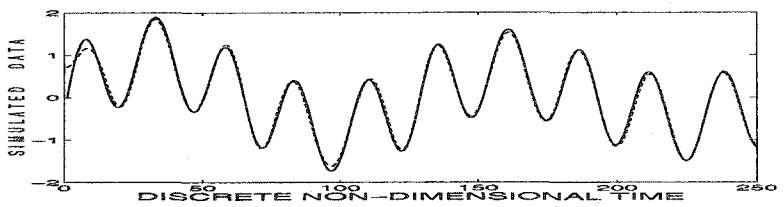

Figure 2.8: Clean ('-') and filtered ('- -') signals($\alpha = 0$, additive noise)

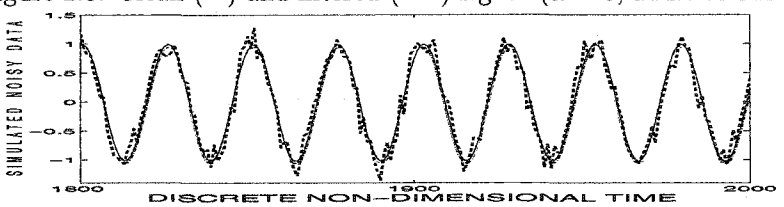

Figure 2.9: Noisy ('- -') and predicted ('-') signals ($\alpha = 0$, additive noise)

Using the filtered data, the EXPAR model can accurately predict the long-term nonlinear behavior. In Fig. 2.9, we compare the noisy and the predicted signal for the limit cycle oscillation. Similar results are obtained for the noisy signals corresponding to $\alpha = -1, 1$. In all three cases the training set contains the filtered observations from 126 to 250. The structural parameters are reported in Table 2.2. Compared to the clean signals, we need more complicated models for the noisy signals.

The same approach works well even if the noise is added within the sine function. For example, Fig. 2.10 displays simulated data according to the formula

$$x_t = \mathrm{e}^{-0.4t} \sin\left(2\pi t + 0.1 \sin\left(6\pi t\right)\right) + \mathrm{e}^{t} \sin\left(10\pi t + e_t\right), \qquad (2.15)$$

α	-1	0	1
p	16	16	14
γ	24.18	21.5	7.35
$\max_{1\leq i\leq p} r_i$	4	3	0

Table 2.2: EXPAR model parameter for noisy simulated data

where e_t is Gaussian white noise such that the signal to noise ratio is again 5.

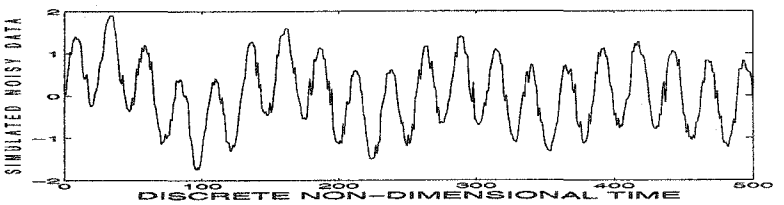

Figure 2.10: Simulated data with non-additive noise ($\alpha = 0$)

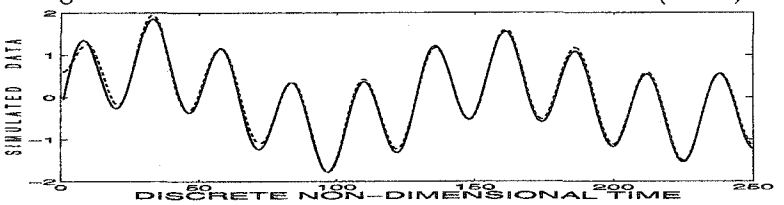

Figure 2.11: Clean ('-') and filtered ('- -') signals($\alpha = 0$, non-additive noise)

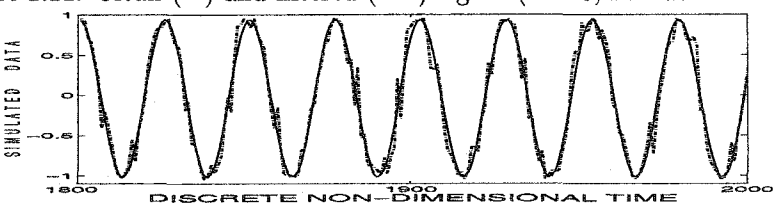

Figure 2.12: Noisy ('- -') and predicted ('-') signals($\alpha = 0$, non-additive noise)

In Fig. 2.11, we compare the clean signal and the filtered signal. In formula (2.15) the Gaussian noise is not additive and comparing Fig. 2.8 with Fig. 2.11, we notice a slightly better performance for removing the additive noise in formula (2.14). The performances of the implemented de-noising algorithms determine the goodness of the fit. The predictions are compared to the noisy signal, in Fig. 2.12. The training set contains again the observations from 126 to 250, $\gamma = 6.1$, $p = 15$ and we use a model (2.7) with polynomials of degree 3.

2.4.3 Simulated aeroelastic data

Now, we apply the non-linear time series models to an aeroelastic system with structural concentrated nonlinearities. The mathematical model associated with a two-degree-of-freedom airfoil oscillating in pitch and plunge is expressed by a nonlinear system [13]:

$$\begin{aligned} &\xi'' + x_\alpha \alpha'' + 2\zeta_\xi \frac{\tilde{\omega}}{U^*}\xi' + \left(\frac{\tilde{\omega}}{U^*}\right)^2 G(\xi) = -\frac{1}{\pi\mu}C_L(t) \\ &\frac{x_\alpha}{r_\alpha^2}\xi'' + \alpha'' + 2\frac{\zeta_\alpha}{U^*}\alpha' + \frac{1}{U^{*2}}M(\alpha) = \frac{2}{\pi\mu r_\alpha^2}C_M(t) \end{aligned} \tag{2.16}$$

The plunging deflection is denoted by ξ, and α is the pitch angle about the elastic axis. $G(\xi)$ and $M(\alpha)$ are the nonlinear plunge and pitch stiffness terms, respectively. $C_L(t)$, $C_M(t)$ are the lift and pitching moment coefficients, and they are expressed by integral terms for the subsonic flows. By introducing four new variables, the integro-differential system (2.16) can be reformulated [13] as

$$X_t' = AX_t + F(X_t), \tag{2.17}$$

where A is the matrix containing the system coefficients, and F is a nonlinear function, $X = [\alpha, \alpha', \xi, \xi', \omega_1, \omega_2, \omega_3, \omega_4]^t$, where ω_1, ω_2, ω_3, ω_4 are the four new variables being introduced to eliminate the integral terms due to $C_L(t)$ and $C_M(t)$.

The nonlinearities in the function F are resulting form the nonlinear plunge and pitch stiffness term. For a cubic spring model $M(\alpha) = \beta_\alpha \alpha + \beta_{\alpha^3}\alpha^3$, where β_α and β_{α^3} are the spring constants. For a freeplay model, $M(\alpha)$ is given by

$$M(\alpha) = \begin{cases} M_0 + \alpha - \alpha_f & \alpha < \alpha_f, \\ M_0 + M_f(\alpha - \alpha_f) & \alpha_f \le \alpha \le \alpha_f + \delta, \\ M_0 + \alpha - \alpha_f + \delta(M_f - 1) & \alpha > \alpha_f + \delta, \end{cases} \tag{2.18}$$

where M_0, δ, α_f, and M_f are the freeplay constants. Similar expressions can be obtained for $G(\xi)$ by replacing α by ξ in the above formulas.

In order to demonstrate the effectiveness of the nonlinear time series models, we present results for two case studies. In each test case, the data set is generated by solving the eight dimensional nonlinear ODE system (2.17) using a fourth-order Runge-Kutta time integration scheme. The system parameters are chosen so that the aeroelastic responses correspond to LCO. A typical input data consists of the 150-400 transient observations. The majority of these input data is used as training set, and the remaining data constitutes the test set. In the following figures, the x-axis displays the non-dimensional time and the y- axis the pitch angle measured in radians, or the non-dimensional plunging deflection.

Cubic spring model

In Figs. 2.13-2.14, we display the pitch and the plunge motions for the aeroelastic system with cubic springs applied to both $G(\xi)$ and $M(\alpha)$. In order to investigate a more realistic test case, we add an extra white noise with variance 0.053 and 0.082, respectively. Figs. 2.13-2.14 show also the noisy data with the signal-to-noise ratio 5 (the dot-dashed signals).

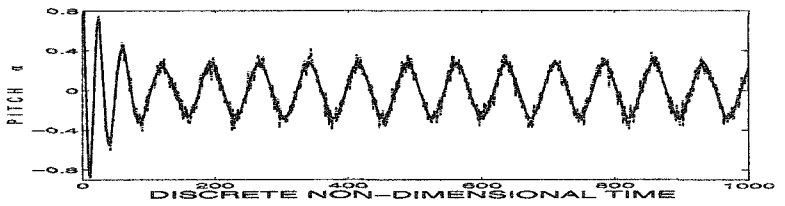

Figure 2.13: Cubic spring model, pitch motion: '-' clean, '-.-' noisy

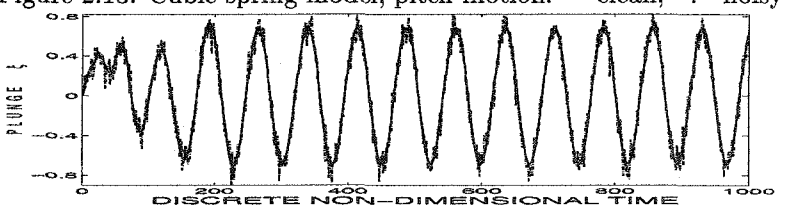

Figure 2.14: Cubic spring model, plunge motion: '-' clean, '-.-' noisy

The EXPAR-models can be applied directly to the clean signals shown in Figs. 2.13 - 2.14. In order to obtain accurate predictions for the noisy signals, we first perform a de-noising procedure in the preprocessing step. In Fig 2.15, we compare

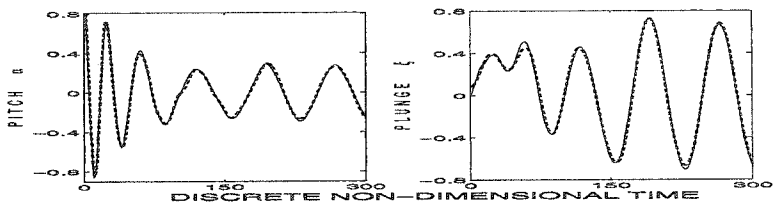

Figure 2.15: Cubic spring model: '-' clean and '- -'filtered signals

the clean signals (solid line) with the de-noised signals (dashed line) using the cosine bases filters. Without giving any information about the structure or the parameters of the associated aeroelastic system, excellent results are obtained applying these filters.

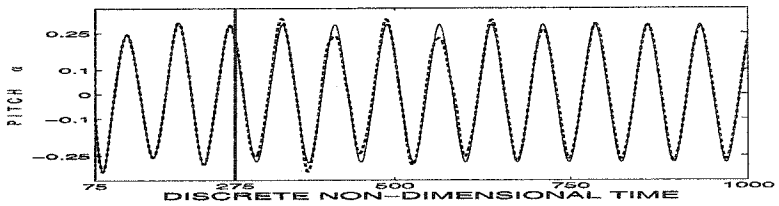

Figure 2.16: Cubic spring model, pitch motion: '-' clean, '- -' predicted

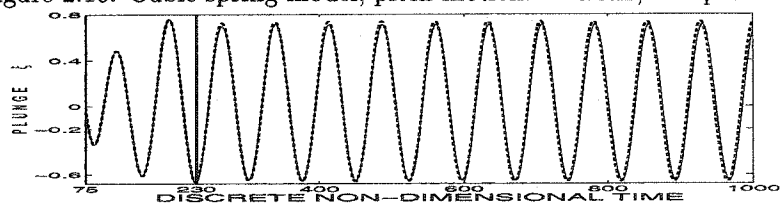

Figure 2.17: Cubic spring model, plunge motion: '-' clean, '- -' predicted

Using the filtered signals, we obtain the predictions displayed in Figs. 2.16-2.17. For the pitch angle α, the training set is taken from $n = 75$ to $n = 274$ and the prediction starts at $n = 275$ (see Fig. 2.16). For the plunging motion, the training set is taken from $n = 75$ to $n = 229$ and the prediction starts at $n = 230$ (see Fig. 2.17). For both case studies, we fit the EXPAR models (2.7) with polynomials of degree 4 and $\gamma = 16.9$, $p = 16$, for the pitch, and $\gamma = 32.62$, $p = 5$, for the plunge, respectively. We notice that even for these noisy signals, the amplitude and the frequency of the LCO are correctly predicted. Excellent predictions are observed when the EXPAR models are applied using the clean data given in Figs. 2.13 - 2.14.

Freeplay model

In Figs. 2.18 - 2.19 (solid line), we display the initial time history for the pitch and plunge motions corresponding to an aeroelastic system with a freeplay in $M(\alpha)$ and a linear spring in $G(\xi)$. The asymptotic state is again a LCO. The aeroelastic response is more complex compared to the first example. The first 375 observations represent the transient of the plunge and pitch motion. In Figs. 2.18 - 2.19 (dot-dashed line), we also display the data corrupted with white noise with variances 0.078 and 0.25, respectively (the signal-to-noise ratio is 5).

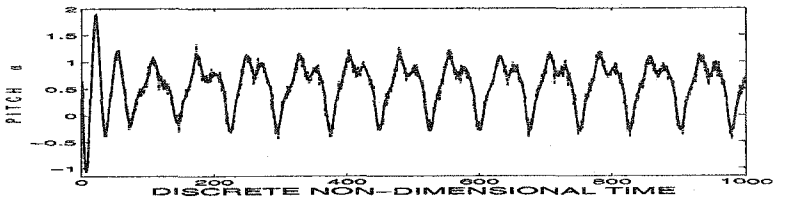

Figure 2.18: Freeplay model, pitch motion: '-' clean, '-.-' noisy

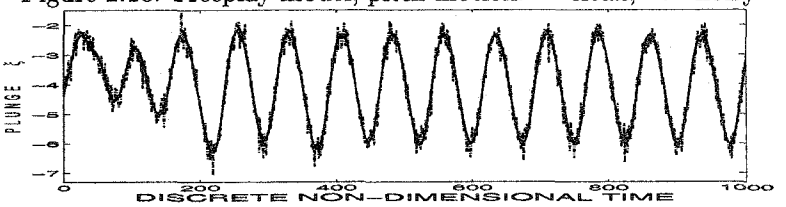

Figure 2.19: Freeplay model, plunge motion: '-' clean, '-.-' noisy

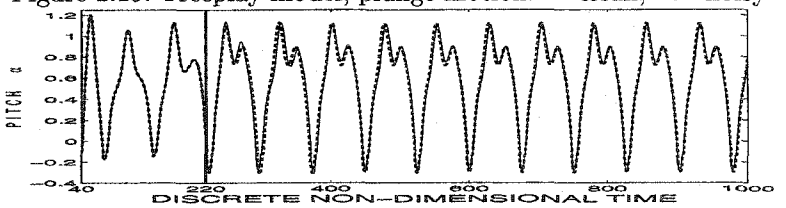

Figure 2.20: Freeplay model, pitch motion: '-' clean, '- -' predicted

Firstly, we fit the EXPAR models for the clean signals shown in Figs. 2.18 - 2.19. The EXPAR models are selected with polynomials of degree 3, $\gamma = 16.3$, $p = 16$ and the training set from $n = 40$ to $n = 219$ for the pitch, and $\gamma = 11.1$, $p = 8$ and the training set from $n = 70$ to $n = 219$ for the plunge. In Fig. 2.20, we display the results for the simulated pitch motion (solid line) and the predicted motion (dashed

line). The predictions are excellent and they begin at $n = 220$. The results for the simulated plunge motion are similar.

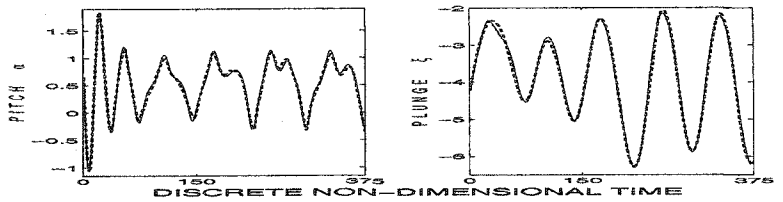

Figure 2.21: Freeplay model: '-' clean and '- -'filtered signals

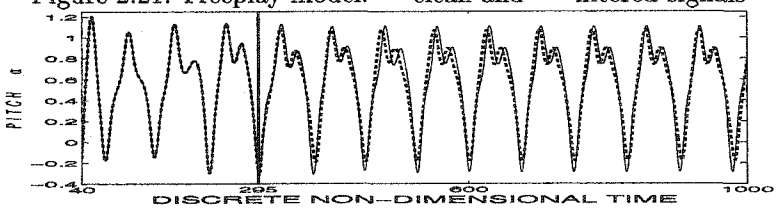

Figure 2.22: Freeplay model,noisy pitch motion: '-' clean, '- -' predicted

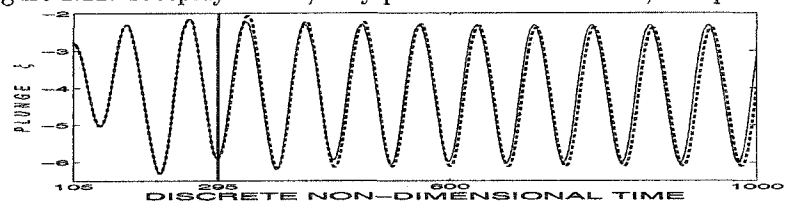

Figure 2.23: Freeplay model, noisy plunge motion: '-' clean, '- -' predicted

Since the predictions for the clean signals are very accurate, we now consider the noisy data given in Figs. 2.18 - 2.19. For the pitch data, we first apply a de-noising procedure using a translation invariant hard thresholding with the Daubechies orthogonal wavelets. For the noisy plunge data, we apply a local cosine-bases de-noising. In Fig. 2.21, we compare the clean signals (solid line) with the filtered signals (dashed line). Using the de-noised signal for the pitch data, we fit an EXPAR model with polynomials of degree 3, $\gamma = 11.89$, $p = 20$ and training sets from $n = 40$ to $n = 294$. For the corresponding de-noised plunge data, we fit an EXPAR model with polynomials of degree 3, $\gamma = 42.1$, $p = 20$ and training sets taken from $n = 105$ to $n = 294$. The predictions, starting at $n = 295$, are displayed in Figs. 2.22 - 2.23. Compared to the results obtained for a clean signal (Fig. 2.20), the predictions are less accurate. Thus, the goodness of the fit is sensitive to the signal-

to-noise ratio. However, the predicted frequencies and amplitudes of LCOs are in good agreement with the simulated data.

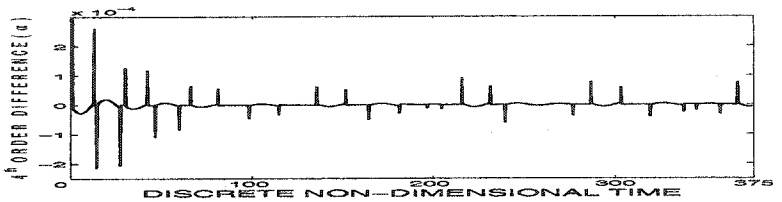

Figure 2.24: Freeplay model,4th order difference

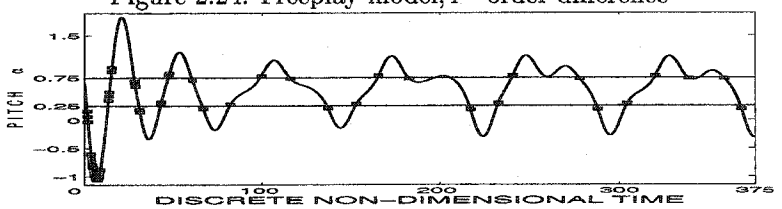

Figure 2.25: Thresholds for the freeplay model

The EXPAR models do not require any information about the threshold structure of the freeplay nonlinearity. However, from the study of the fourth order differences of the data for the pitch motion, we notice periodically changes in the dynamics of the model (see Fig. 2.24). This is caused by the non-differentiability of the function $M(\alpha)$ near the switching points. The values corresponding to the peaks shown in Fig. 2.24 are plotted in Fig. 2.25. Indeed, these values correspond to the exact locations of the switching points in the freeplay model, where $\alpha = .25$ and $\alpha = .75$ in the present case study. Thus, for a clean signal, from studying the differences, we can determine whether we have a freeplay nonlinearity and we can also estimate the thresholds.

Since the Gaussian noise is non-differentiable, the study of the differences is not very helpful for a noisy signal. To find the thresholds, we perform an exploratory data analysis as presented in Section 2.3. Hence, we estimate the conditional means $m_j(x) = E(X_n|X_{n+j})$ and the conditional variances $v_j(x) = VAR(X_n|X_{n+j})$, $j = -80, -30, -20, -18, -8 \ldots -1$, using an non-parametric approach (see equation 2.12).

Fig. 2.26 shows the results for the non-parametric lag-regressions estimates

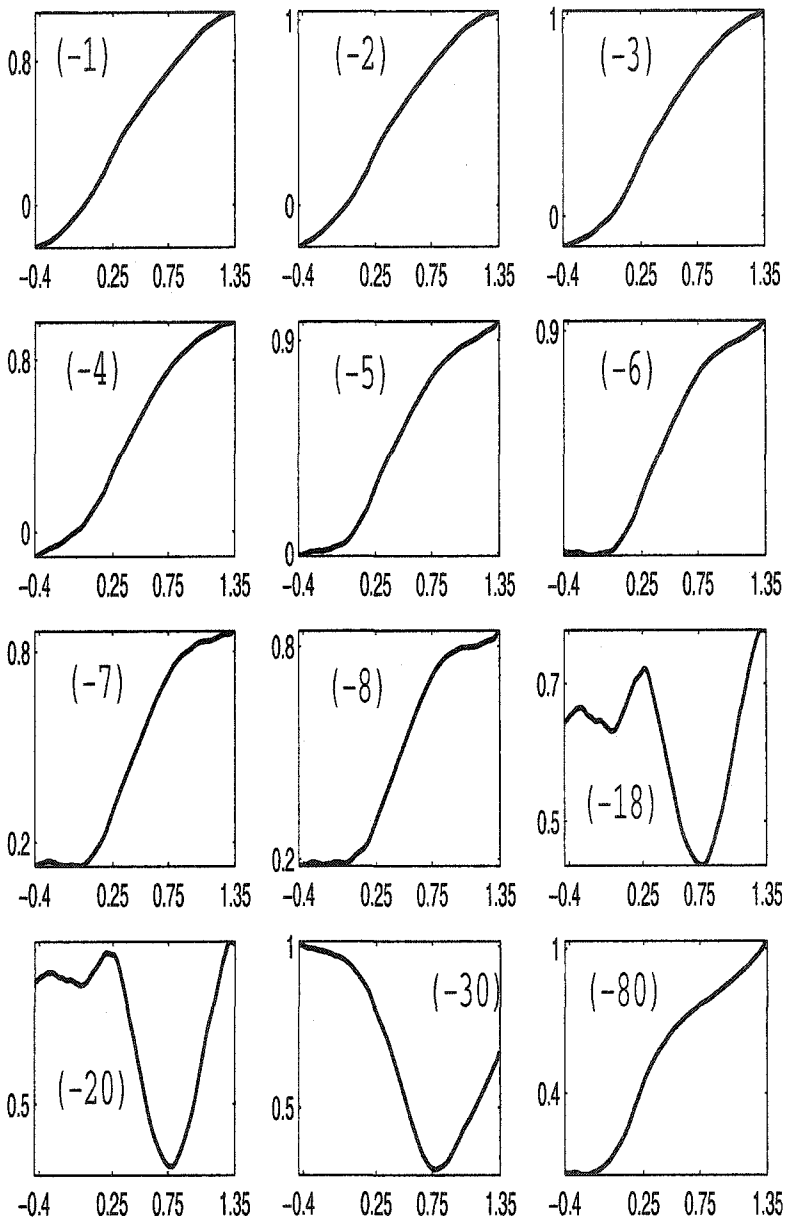

Figure 2.26: The non-parametric lag-regression estimates of the conditional mean $\hat{m}_j(x) = E(x_n|x_{n+j})$ as a function of x. The values of 'j' are shown in parentheses.

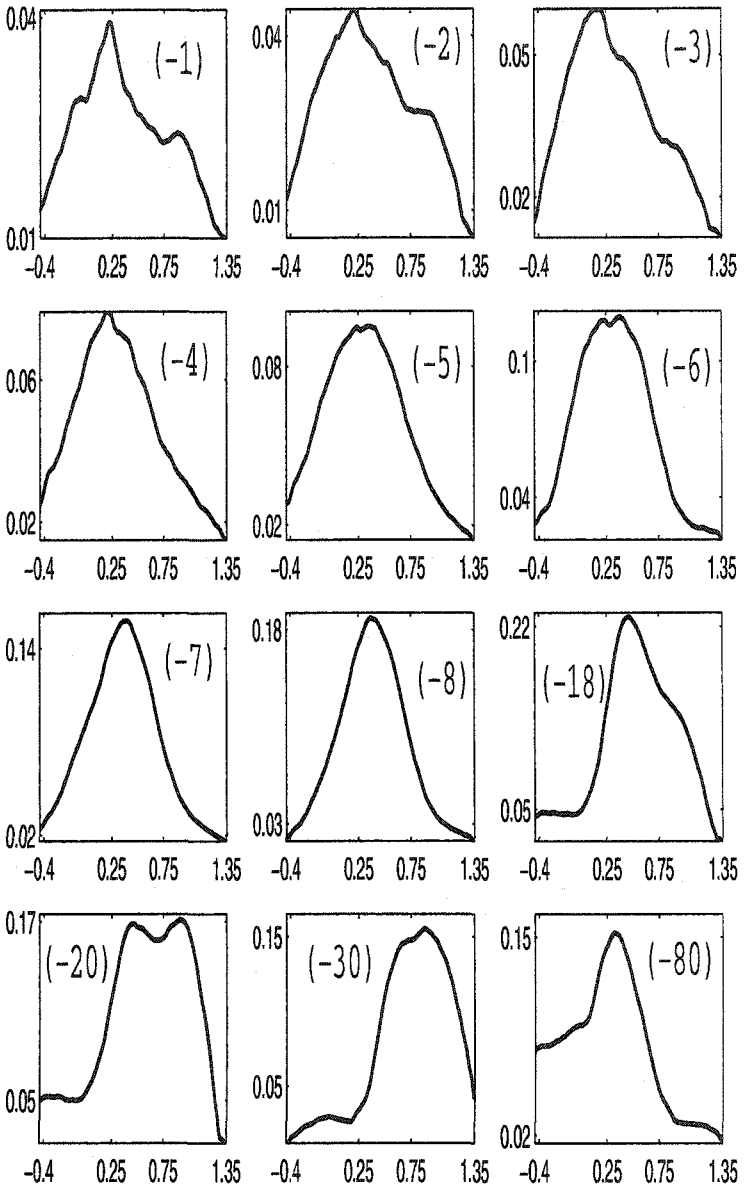

Figure 2.27: The non-parametric lag-regression estimates of the conditional variance $\hat{v}_j(x) = VAR(x_n|x_{n+j})$ as a function of x. The values of 'j' are shown in parentheses.

$\hat{m}_j(x)$. The values of x and $\hat{m}_j(x)$ are represented on the x and the y axes, respectively. We notice a gradual shift from an almost linear function to a curve with two main inflexion points back to an almost linear function. The inflexion points are approximately at 0.25 and 0.75. Thus, they corresponds to the values of the thresholds. They are visible first for the values of j around -5, -4. Hence, the delay parameter for the SETAR model can be $d = 4$ or $d = 5$ (we have tried both values and obtained similar results).

The non-parametric estimates of the variances $\hat{v}_j(x)$ give the same informations about the threshold structure. In Fig. 2.27, the values of x and $\hat{V}_j(x)$ are represented on the x and the y axes, respectively. The thresholds are most transparent for $j = -1$, when we have a two-hump curve with the turning points at 0.25 and 0.75. The rest of the pictures shows two-hump curves or one-hump curves with the inflexion points around 0.25 and 0.75.

To improve our nonlinear prediction, we implement the threshold structure into the nonlinear time series model by combining the SETAR and the EXPAR models. We use an EXPAR model (2.1) for the first region, and extended EXPAR models (2.7) with polynomials of degrees 3 and 2, respectively, for the other two regions. The results are very similar to those displayed in Figs. 2.22 - 2.23, but the model becomes more complicated.

2.4.4 Experimental aeroelastic data

With the success obtained for the simulated aeroelastic data, we investigate the performances of the EXPAR models for experimental wind tunnel data[2] recorded at the Texas A&M University. We consider two case studies, one corresponds to a LCO (data from the file DN04J.dat) and the other corresponds to a steady state (data from the file DN04A.dat). The mathematical model associated with the experimental data contains a nonlinear spring stiffness term, approximated by a fourth order polynomial [11]. Since the noise effect is not too severe, no de-noising procedure is necessary when fitting EXPAR models.

[2]The data are available online at http://aerounix.tamu.edu/aeroel.

The results for the LCO case are displayed in Figs. 2.28 and 2.29 for the pitch and the plunge motions, respectively. For the pitch motion, we fit an EXPAR model (2.7) with polynomials of degree 3, $p = 18$ and $\gamma = 46.1$. The training set contains the observations from $n = 100$ to $n = 349$, and the prediction starts at $n = 350$. For the plunge motion, the prediction starts at $n = 350$, but the training set contains only the observations between $n = 200$ and $n = 349$. In the EXPAR model (2.7) polynomials of degree 3 are employed with $p = 12$ and $\gamma = 43.1$.

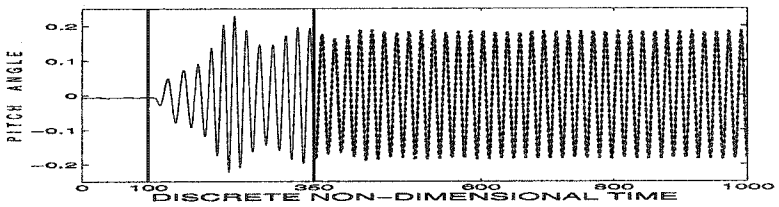

Figure 2.28: Experimental LCO: '-' pitch motion, '- -' EXPAR prediction

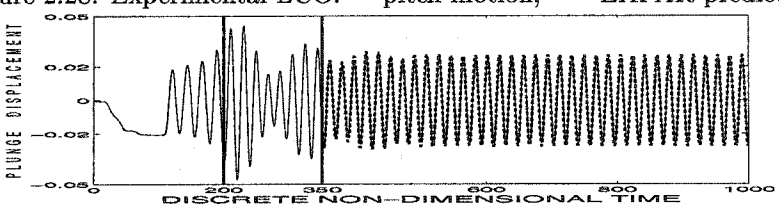

Figure 2.29: Experimental LCO: '-' plunge motion, '- -' EXPAR prediction

From the first 350 transient observations, the prediction leading to an LCO is not obvious. However, the results presented in Figs. 2.28 and 2.29 demonstrate that the fitted EXPAR models are capable to provide excellent long term predictions for the pitch angle and the plunge displacement.

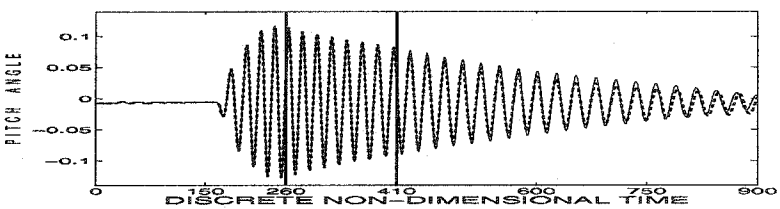

Figure 2.30: Experimental damped oscillation: '-' pitch motion, '- -' EXPAR prediction

Now, we present the results obtained for the pitch motion when the aeroelastic

system exhibits a steady damped oscillation. For the plunge motion, the quality of the prediction is similar and they are not reported here. The predictions (dashed line) starting at $n = 410$ and the measured signal (solid lines) are compared in Fig. 2.30. In this model, polynomials of degree 1 are used with $\gamma = 0.1$ and $p = 2$. The training set contains the observations corresponding to $n = 260$ to $n = 409$. The EXPAR model is capable of providing accurate predictions of the damped oscillations.

Bibliography

[1] M. Andrighettoni and P. Mantegazza. Multi-input/multi-output adaptive active flutter suppression for a wing model. *J. Aircraft*, 35(3):462–469, 1998.

[2] K.S Chan and H. Tong. On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Adv. Appl. Probab.*, 17:666–678, 1985.

[3] R.R. Coifman and D.L. Donoho. Translation invariant de-noising. Technical report, Stanford University and Yale University, 1995. http://www-stat.stanford.edu/ donoho.

[4] R.R. Coifman and Y. Meyer. Remarques sur l'analyse de Fourier a fênetre. *C. R. Acad. Sci. Paris*, (312):259–261, 1991.

[5] R.R. Coifman, Y. Meyer, and M.V. Wickerhauser. Wavelet analysis and signal processing. In M.B. Ruskai et al., editor, *Wavelets and Their Applications*, pages 153–178. Jones and Bartlett, Boston, 1992.

[6] R.R Coifman and M.V. Wickerhauser. Entropy based algorithms for best-basis selection. *IEEE Trans. Info. Theory*, 38:713–718, 1992.

[7] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, Philadelphia, 1992.

[8] G Dimitriadis and J. E. Cooper. Flutter prediction from flight flutter test data. *J. Aircraft*, 38(2):355–367, 2001.

[9] S.N. Elaydi. *An Introduction to Difference Equations.* Springer Verlag, New York, 1996.

[10] V. Haggan and T. Ozaki. Modeling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika*, 68(1):189–196, 1981.

[11] J. Ko, T. W. Strganac, and A. J. Kurdila. Stability and control of a structurally nonlinear aeroelastic system. *AIAA J. Guid.,Contr., Dynamics*, 21(5):718–725, 1998.

[12] V. Lakshmikantham and D. Trigiante. *Theory of Difference Equations: Numerical Methods and Applications.* Marcel Dekker, Inc., New York, Basel, 2001.

[13] B. H. K. Lee, S. J. Price, and Y. S. Wong. Nonlinear aeroelastic analysis of airfoils: bifurcation and chaos. *Progress in Aerospace Sciences*, 35:205–334, 1999.

[14] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, 1999.

[15] T. Ozaki. Nonlinear time series models for nonlinear random vibrations. *J. Appl. Prob.*, 17:84–93, 1980.

[16] T. Ozaki. The statistical analysis of perturbed limit cycle processes using nonlinear time series models. *J. Time Ser. Anal.*, 3(1):29–41, 1982.

[17] T. Ozaki. Nonlinear time series models and dynamical systems. In E.J. Hannan et al., editor, *Handbook of Statistics*, pages 25–83. North-Holland, 1985.

[18] C. A. Popescu and Y. S. Wong. A nonlinear statistical approach for aeroelastic response prediction. In *Proceedings of the 43rd AIAA/ ASME/ ASCE/ AHS/ ASC/ Structures, Structural Dynamics, and Materials Conference*, Denver, CO, USA, 2002. AIAA Paper 2002-1281.

[19] C. A. Popescu and Y. S. Wong. A nonlinear statistical approach for aeroelastic response prediction. *AIAA J. Guid.,Contr., Dynamics*, 26(4):565–572, 2003.

[20] P.M. Robinson. Non-parametric estimation for time series models. *J. Time Ser. Anal.*, 4:185–208, 1983.

[21] R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications.* Springer Verlag, New York, 2000.

[22] H. Tong. *Nonlinear Time Series.* The Clarendon Press Oxford University Press, New York, 1990.

[23] H. Tong and K.S. Lim. Threshold autoregression, limit cycle and cyclical data. *J. Royal Stat. Soc.*, B42(3):245–292, 1980.

[24] H. Torii and Y. Matsuzaki. Flutter boundary prediction based on nonstationary data measurement. *J. Aircraft*, 34(3):427–432, 1997.

[25] H. Torii and Y. Matsuzaki. Flutter margin evaluation for discrete-time systems. *J. Aircraft*, 38(1):42–47, 2001.

[26] R.S. Tsay. *Analysis of Financial Time Series.* J. Wiley & Sons, New York, 2002.

[27] Y. S. Wong, C. A. Popescu, and O. Voitcu. Nonlinear dynamic prediction and feature extraction. In *Proceedings of the 48th Annual Conference of Canadian Aeronautics and Space Institute*, pages 173–205, Toronto, 2001.

Chapter 3

Nonlinear State-Space Models

3.1 Introduction

In this chapter, we propose an alternative approach to the nonlinear time series models. Many mathematical models corresponding to aeroelastic systems are based on nonlinear ordinary differential equations (ODE) that can be expressed in a state-space form. Here, our approach is based on the state-space form of the associated aeroelastic models. For aeroelastic systems with polynomial or freeplay nonlinearities, we compare the performances of the unscented filter (UF) [10] against the extended Kalman filter (EKF) ([7], Chapter 9). Not only are these filters used for noise removal, but they also can be employed for estimation and prediction.

Generally speaking, the implementation of nonlinear time series models is an illustration of Takens' [24] remarkable theorem which proves that for almost any deterministic nonlinear system with a d dimensional space, the state can be effectively reconstructed by observing $2d + 1$ time lags of its outputs. Hence, instead of building a state-space model, we can construct an autoregressive model directly on the observations that nonlinearly relates the previous outputs and the current output. The main advantage is the possibility to work with only one dimensional

The material presented in this chapter was previously published in [18], [19], [20]

signals. However, unlike the noise-free case of the Takens' theorem, when the dynamics are noisy, the optimal prediction of the observation would depend on the entire history of past observations. Any truncation of this time history could throw away potentially valuable information about the unobserved state. The state-space formulation of nonlinear dynamical systems allows us to overcome the limitations of the nonlinear autoregressive models. The price needed to be paid is that it requires more informations regarding the nature of the nonlinearities.

Kalman estimators have a long history in the study of aeroelastic phenomena [14]. The EKF was often applied as a real-time parameter identifier. For example, Roy and Walker [22] use the EKF to improve the estimation of the damping for identification of the flutter stability; Block and Strganac [5] use Kalman estimators with a linear quadratic regulator to control the plunge and pitch motions of a wing.

The UF was introduced by Julier and Uhlmann [10]. This filter does not require the calculation of the Jacobians, and it is computationally at most as expensive as the EKF. Thus, the UF can be applied for continuous non-differentiable nonlinearities such as the freeplay model.

3.2 The Filters

Consider a general nonlinear discrete system

$$\begin{aligned} x_{k+1} &= f[x_k, u_{k+1}] + v_{k+1}, \\ y_{k+1} &= h[x_{k+1}, u_{k+1}] + w_{k+1}. \end{aligned} \tag{3.1}$$

Here $f[\cdot,\cdot]$ is the process model, x_k is the state of the system at the k-time step, u_k is the input vector, y_k is the observation vector, $h[\cdot,\cdot]$ is the observation model, v_k is the noise process and w_k is the additive measurement noise. We assume that the noise vectors, v_k and w_k, are Gaussian and from uncorrelated white sequences: $E[v_k] = E[w_k] = 0$, for all k, and $E[v_i v_j^t] = \delta_{ij} Q_i$, $E[w_i w_j^t] = \delta_{ij} R_i$, $E[v_i w_j^t] = 0$, for all i, j, where δ_{ij} is the Kronecker symbol. We use the notations $\bar{x}_i = E[x_i | y_1, \ldots, y_i]$ and $\hat{x}_{i+1} = E[x_{i+1} | y_1, y_2, \ldots, y_i]$ for the filtered and the predicted values, respectively.

The corresponding conditional covariances are $\bar{P}_i = E\left[(x_i - \bar{x}_i)(x_i - \bar{x}_i)^t | y_1, \ldots, y_i\right]$ and $\hat{P}_{i+1} = E\left[(x_{i+1} - \hat{x}_{i+1})(x_{i+1} - \hat{x}_{i+1})^t | y_1, \ldots, y_i\right]$. The classical Kalman update equations ([3], Chapter 4) at time $k+1$ are

$$\begin{aligned}\bar{x}_{k+1} &= \hat{x}_{k+1} + G_{k+1}\nu_{k+1}, \\ \bar{P}_{k+1} &= \hat{P}_{k+1} - G_{k+1}\hat{P}_{\nu_{k+1}\nu_{k+1}}G_{k+1}^t.\end{aligned}$$

Here, $\nu_{k+1} = y_{k+1} - \hat{y}_{k+1}$ is the innovation, $\hat{P}_{\nu_{k+1}\nu_{k+1}}$ is its conditional covariance and $G_{k+1} = \hat{P}_{x_{k+1}y_{k+1}}\hat{P}^{-1}_{\nu_{k+1}\nu_{k+1}}$ is the Kalman gain. With this updating scheme, the remaining problem is to determine the optimal predictions $\hat{x}_{k+1}$ and $\hat{P}_{k+1}$. In the linear case, the Kalman filter calculates these quantities exactly. For nonlinear models, both the EKF and the UF require approximations of these quantities.

3.2.1 The Extended Kalman Filter (EKF)

For the EKF, we suppose that the estimated mean $\bar{x}_k$ is approximately equal to the true state x_k, and then

$$\hat{x}_{k+1} = f(\bar{x}_k, u_{k+1}). \tag{3.2}$$

Similarly, the predicted observation is given by

$$\hat{y}_{k+1} = h(\hat{x}_{k+1}, u_{k+1}). \tag{3.3}$$

The covariances are then determined by linearizing the equations (3.1) using Taylor series expansions and neglecting the second and higher order terms:

$$\hat{P}_{k+1} = \frac{\partial f}{\partial x}(\bar{x}_k)\bar{P}_k\frac{\partial f}{\partial x}^t(\bar{x}_k) + Q_{k+1}, \tag{3.4}$$

$$\hat{P}_{\nu_{k+1}\nu_{k+1}} = \frac{\partial h}{\partial x}(\hat{x}_{k+1})\hat{P}_{k+1}\frac{\partial h}{\partial x}^t(\hat{x}_{k+1}) + R_{k+1}. \tag{3.5}$$

The cross-covariance is given by

$$\hat{P}_{x_{k+1}y_{k+1}} = \hat{P}_{k+1}\frac{\partial h}{\partial x}^t(\hat{x}_{k+1}). \tag{3.6}$$

Since the estimates in (3.2)-(3.6) are based on a first order approximation of the nonlinear terms, the EKF is generally sub-optimal for nonlinear systems. However, second order versions of the EKF can be developed, but they are usually computationally expensive.

Based on the available observations $y_1 \dots y_N$, smoothed values $x_i^N = E[x_i|y_1 \dots y_N]$ and $P_i^N = E[(x_i - x_i^N)(x_i - x_i^N)^t|y_1 \dots y_N]$ can be calculated. Using the linearized system

$$\begin{aligned} x_{k+1} &= f(\bar{x}_k, u_{k+1}) + \frac{\partial f}{\partial x}(\bar{x}_k)(x_k - \bar{x}_k) + v_{k+1}, \\ y_k &= h(\bar{x}_k, u_k) + \frac{\partial h}{\partial x}(\bar{x}_k)(x_k - \bar{x}_k) + w_k, \end{aligned} \tag{3.7}$$

we have the following backward recursions

$$x_{k-1}^N = \bar{x}_{k-1} + J_{k-1}(x_k^N - \hat{x}_k), \tag{3.8}$$

$$P_{k-1}^N = \bar{P}_{k-1} + J_{k-1}(P_k^N - \hat{P}_k)J_{k-1}^t, \tag{3.9}$$

where

$$J_{k-1} = \bar{P}_{k-1}\frac{\partial f}{\partial x}^t(\bar{x}_{k-1})\hat{P}_k^{-1}. \tag{3.10}$$

3.2.2 The Unscented Filter (UF)

Both the EKF and the UF approximate the state distribution with a Gaussian one. However, instead of using the EKF linearization approach, the UF employs a deterministic sampling. The sample points completely capture the true mean and the true covariance. In contrast to the first-order accuracy of the EKF, the UF is capable to accurately capture the true posterior mean and the covariance up to the third order for a nonlinear system. Moreover, the UF is computationally attractive, it does not require the calculations of the Jacobians, and it can be applied to aeroelastic systems with structural nonlinearities given by freeplay and hysteresis models.

In order to state the UF equations, let us define the $2n$-dimensional augmented state vector $X_k^a = (x_k^t, v_{k+1}^t)^t$, where n is the dimension of the state space, and

denote $F[(x_k^t, v_{k+1}^t)^t, u_{k+1}] = f[x_k, u_{k+1}] + v_{k+1}$ and $H[(x_k^t, w_k^t)^t, u_k] = h[x_k, u_k] + w_k$. For the UF filtering algorithm, the following steps must be carried out [8]:

1. Compute the set of the translated sigma points from the augmented covariance matrix $\tilde{P}_k^a$

$$\sigma_i^a(k|k) \leftarrow 2(n+n) \text{ rows or columns from } \pm\sqrt{(2n+\gamma)\tilde{P}_k^a},$$
$$\chi_0(k|k) = \bar{X}_k^a,$$
$$\chi_i(k|k) = \sigma_i^a(k|k) + \bar{X}_k^a.$$

2. Evaluate $\chi_i(k+1|k) = F[\chi_i(k|k), u_{k+1}]$, for $i = 0, \ldots, 4n$.

3. Compute the predicted mean as

$$\hat{x}_{k+1} = \frac{1}{2n+\gamma}\{\gamma\chi_0(k+1|k) + \frac{1}{2}\sum_{i=1}^{4n}\chi_i(k+1|k)\}.$$

4. Compute the predicted covariance as

$$\hat{P}_{k+1} = \frac{1}{2n+\gamma}\{\gamma[\chi_0(k+1|k) - \hat{x}_{k+1}][\chi_0(k+1|k) - \hat{x}_{k+1}]^t$$
$$+ \frac{1}{2}\sum_{i=1}^{4n}[\chi_i(k+1|k) - \hat{x}_{k+1}][\chi_i(k+1|k) - \hat{x}_{k+1}]^t\}$$

5. Predict the expected observation $\hat{y}_{k+1}$ and the innovation covariance $\hat{P}_{\nu_{k+1}\nu_{k+1}}$ using similar formulas and $Y_i(k+1|k) = H[\chi_i(k+1|k), u_{k+1}]$.

6. Predict the cross-correlation matrix $\hat{P}_{x_{k+1}y_{k+1}}$ using $Y_i(k+1|k)$ and $\chi_i(k+1|k)$.

Using the multi-dimensional Taylor series expansions, it can be shown that the new filter estimate the mean and the covariance exactly up to the third order terms (see Theorem 2 in [8]). Choosing $\gamma = 3 - 2n$ we minimizes the difference between the moments of the standard Gaussian and the sigma points up to the fourth order ([8]). Thus, without calculating the Jacobians or the Hessians, we can get the same order of accuracy as the Gaussian second order filter [2]. Furthermore, Julier and

Uhlmann have extended the algorithm to manipulate even higher order properties of the distribution [9].

Now we present a method to obtain the smoothed values x_k^N and P_k^N, $k = 1, \ldots, N$. Notice that, for the linearized system (3.7), we have

$$\bar{P}_{x_k x_{k+1}} = E[(x_k - \bar{x}_k)(x_{k+1} - \hat{x}_{k+1})^t | x_1, \ldots, x_k] = \bar{P}_k \frac{\partial f^t}{\partial x}(\bar{x}_k).$$

Thus, using (3.10), we get

$$J_k = \bar{P}_{x_k x_{k+1}} \hat{P}_{k+1}^{-1}, . \tag{3.11}$$

Hence, the unscented smoothing algorithm can be described as follows

1. Similarly to the UF, compute the set of the translated sigma points from the augmented covariance matrix $\bar{P}_k^a$

$$\begin{aligned}
&\sigma_i^a(k|k) \leftarrow 4n \text{ rows or columns from } \pm\sqrt{(2n+\gamma)\bar{P}_k^a}, \\
&\chi_0(k|k) = \bar{X}_k^a, \\
&\chi_i(k|k) = \sigma_i^a(k|k) + \bar{X}_k^a.
\end{aligned}$$

2. Evaluate $\chi_i(k+1|k) = F[\chi_i(k|k), u_{k+1}]$, for $i = 0, \ldots, 4n$.

3. Compute $\bar{x}_{k+1}$, $\bar{P}_{k+1}$, $\hat{x}_{k+1}$, and $\hat{P}_{k+1}$ using the UF.

4. Compute

$$\begin{aligned}
\bar{P}_{x_k x_{k+1}} = \frac{1}{2n+\gamma}\{&\gamma[\chi_0(k|k) - \bar{x}_k][\chi_0(k+1|k) - \hat{x}_{k+1}]^t \\
&+ \frac{1}{2}\sum_{i=1}^{4n}[\chi_i(k|k) - \bar{x}_k][\chi_i(k+1|k) - \hat{x}_{k+1}]^t\}
\end{aligned}$$

5. Compute x_k^N, P_k^N, J_k starting with $k = N$ and using the backwards recursions given in (3.8), (3.9) and (3.11).

The recursions (3.8) and (3.9) are based on the linearization. Thus, unlike the UF, this smoothing method is dependent on the accuracy of the approximation with the

linearized system (3.7). However, compared to the EKF smoothing, it still has the advantage of not requiring the calculation of the Jacobians.

3.2.3 Comparisons

To examine the effectiveness of the EKF and the UF, we apply both methods to data generated from numerical simulations. We consider a two degree of freedom airfoil oscillating in pitch and plunge, according to the model expressed by the system of ODEs given in equation 2.16. We simulate a chaotic aeroelastic response [12] of an oscillating airfoil with cubic restoring forces in the pitch and plunge:

$$M(\alpha) = \beta_\alpha \alpha + \beta_{\alpha^3} \alpha^3,$$
$$G(\xi) = \beta_\xi \xi + \beta_{\xi^3} \xi^3,$$

with the spring constants $\beta_\alpha = \beta_\xi = 1$, $\beta_{\alpha^3} = 50$ and $\beta_{\xi^3} = 0.01$.

A fourth order Runge-Kutta scheme was employed to solve the equivalent ODEs system given in equation 2.17. The two filters were implemented in discrete time starting from the same initial guess. We suppose that the coefficients of the system are known, but only a few transient values of α and ξ are observed. Hence, the observations space contains only the noisy observations for the plunging deflection ξ and the pitch angle α. The discrete form corresponding to the system (2.17) is given by

$$x_{k+1} = f(a_0, x_k) + v_{k+1},$$
$$y_{k+1} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{bmatrix} x_{k+1} + w_{k+1},$$

where a_0 is a constant vector containing the values of the parameters. Here, x_k is an eight dimensional vector corresponding to X in (2.17), and y_k is a two dimensional vector corresponding to the noisy observations for the plunging deflection and the pitch angle. The noises v_{k+1}, w_{k+1} are as in the general description (3.1), and f is a polynomial function. The dimension of the state space is 8 and the dimension of

the observation space is 2.

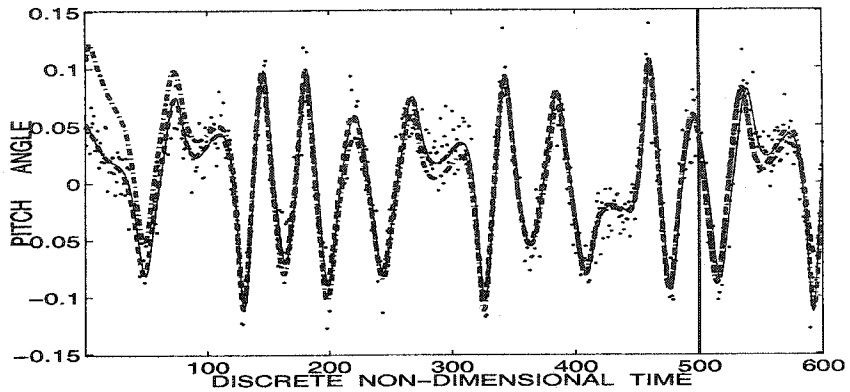

Figure 3.1: Pitch motion: the noisy (..), clean (-), smoothed and predicted signals using the EKF (.-.) and the UF (- -)

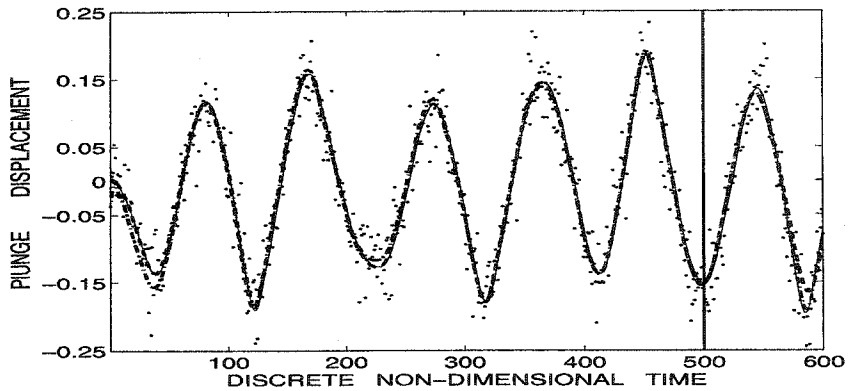

Figure 3.2: Plunge motion: the noisy (..), clean (-), smoothed and predicted signals using the EKF (.-.) and the UF (- -)

The results presented in what follows are obtained by averaging one hundred Monte Carlo trials. The performances of the two filters are compared on the basis of the mean square error, and the UF and EKF estimated covariance matrices.

Figs. 3.1 - 3.2 show the pitch and plunge smoothed and predicted values generated by the EKF and the UF, respectively. The original noisy time series has signal to noise ration SNR=3. The predictions start at $n = 500$. Both graphs indicate that the UF produces more accurate smoothed values than the EKF. However, the accuracy of the predictions for these chaotic signals is comparable for both filters.

In Figs. 3.3 - 3.4, we compare the mean square estimation errors (MSEs) es-

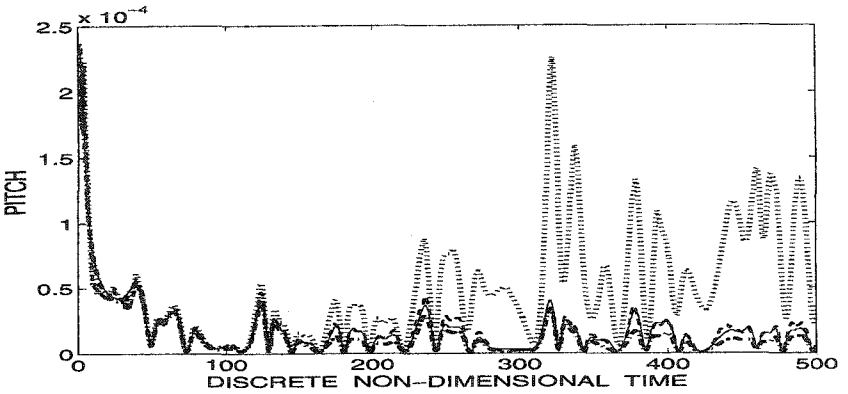

Figure 3.3: Pitch motion: the EVARs for the UF (- -) and the EKF (..) compared with the MSEs for the UF (-) and the EKF (.-.)

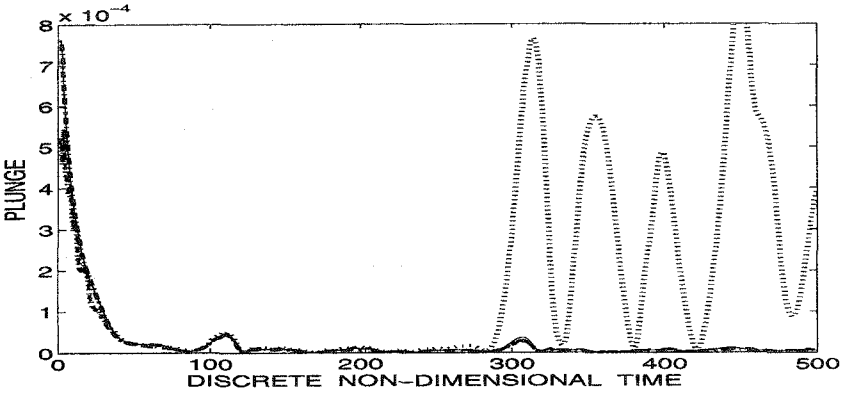

Figure 3.4: Plunge motion: the EVARs for the UF (- -) and the EKF (..) compared with the MSEs for the UF (-) and the EKF (.-.)

timated using the UF and the EKF (i.e. the diagonal elements of $\bar{P}_k$) with the variances (EVARs) estimated using the Monte Carlo simulations. We notice that the EVARs corresponding to the UF is considerable smaller than the one corresponding to the EKF for both the pitch angle and the plunge deflection. Moreover, the values of the MSE and the EVARs corresponding to the UF are similar, but the EVARs are larger than the MSEs for the EKF. The UF smoothing mean square errors and the variances are compared in Figs. 3.5 - 3.6, and we observe again that the values are very close.

In conclusion, the UF estimates its mean square errors accurately and thus we can be confident in the filter estimates. The results for the EKF are less accurate,

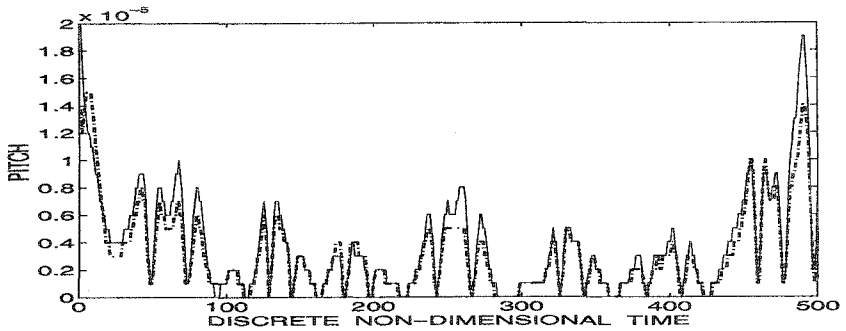

Figure 3.5: Pitch motion: the EVARs (-) and the MSEs (.-.) for the UF smoothing

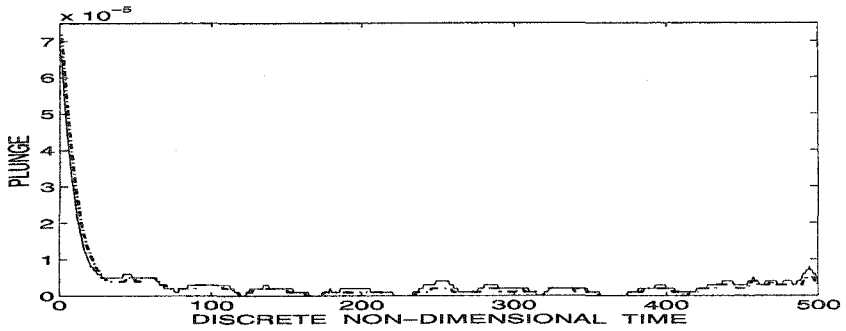

Figure 3.6: Plunge motion: the EVARs (-) and the MSEs (.-.) for the UF smoothing

and they tend to underestimate the mean square errors.

3.3 The UF and the EKF as parameter estimators

In this section, we consider aeroelastic models that can be expressed in the discrete state-space form (3.1), with h a linear function and f a polynomial or a continuous piece-wise linear function. Supposing that the parameters of the system are unknown, and only a few transient observations $y_1, \ldots, y_N$ are recorded during a wind tunnel test, we are interested in estimating the parameters of the system. Then, using the estimated values, we determine long term predictions for the state variable x_k. Here, we present two methods using the UF or the EKF as parameter estimators.

First, the state space is augmented to include the system parameters:

$$\begin{bmatrix} x_{k+1} \\ a_{k+1} \end{bmatrix} = \begin{bmatrix} f[x_k, a_k] \\ a_k \end{bmatrix} + \begin{bmatrix} v_{k+1} \\ v_{k+1}^{(2)} \end{bmatrix},$$
$$y_{k+1} = h[x_{k+1}, a_{k+1}] + w_{k+1}.$$

Here, a_k is the vector formed with the parameters of the system at time k, and v_{k+1}, $v_{k+1}^{(2)}$, and w_{k+1} are uncorrelated Gaussian white noises. The EKF or the UF can be directly applied to the new system, and together with the filtered values $\bar{x}_k$, we can estimate the parameters $\bar{a}_k$.

After applying the filter to the given noisy observations, we use the predictor to study the asymptotic behavior. We fix the parameters to be the last values estimated using the filter $a_0 = \bar{a}_N$. For the predictor, we have the following state-space formulation:

$$\begin{aligned} x_{k+1} &= f(x_k, a_0) + v_{k+1} \\ y_{k+1} &= h[x_{k+1}, a_0] + w_{k+1}. \end{aligned} \tag{3.12}$$

Under the detectability and the stabilizability hypotheses for the linear Kalman filter, the effects of the initial guesses are forgotten for a sufficiently large number of data, and the computations are stable ([1]). Unfortunately, it is difficult to extend these results to the nonlinear case. Even as a linear dynamics parameter estimator, the EKF can have convergence problems [13]. When estimating nonlinear dynamics, the convergence of the EKF or the UF depends on the initial guess. A general result for the convergence of the EKF is given in Theorem 4.1 in [21]. Provided that the initial estimation error is small enough, this theorem establishes sufficient conditions for the estimation error to be bounded. Thus, although it can produce very accurate predictions, the estimation method based on EKF or UF requires an extensive filter tuning.

Next, we consider a dual method by running simultaneously two filters, one for the state variable x_k and the other for the parameters a_k. The state-space

representation for the state variable is (3.12), and for the parameters is

$$\begin{aligned} a_{k+1} &= a_k + v_{k+1}^{(2)}, \\ y_{k+1} &= h(a_{k+1}, x_{k+1}) + w_{k+1}. \end{aligned} \tag{3.13}$$

At each time step k, the current estimate $\bar{x}_k$ is used in the EKF or UF which corresponds to the system (3.13). Then, the EKF or UF current estimation $\bar{a}_k$ replaces the vector a_0 in (3.12).

The dual EKF has been used for estimating linear dynamics [17]. For the nonlinear case, various dual combinations of the EKF and the UF have been compared and better performances of the UF are reported [26]. However, no theoretical proof of the unconditional convergence is known. From our experience, the results are also strongly dependent on the initial guess. In what follows, we present results obtained using the previous two methods applied to numerically simulated and experimental data.

3.3.1 Simulated aeroelastic data

We consider the same simulated data as described in Section 2.4.3. Comparing the performance with the results obtained using nonlinear time series models, we obtain a better assessment of the parameter estimation methods based on the EKF or the UF.

Cubic spring model

Now, we apply the UF directly for the noisy data shown in Figs. 2.13, 2.14. The training set contains data from $k = 1$ to $k = 250$. In order to check the performance of the UF, we run a Monte Carlo simulation. After tuning the filter, we apply the filter to 50 different sets of data simulated using the clean data given in Figs. 2.13 and 2.14, and additive white noise with variance 0.053 and 0.082, respectively (i.e. similarly to the noisy data displayed in Figs. 2.13, 2.14). For the pitch and plunge motions, the clean signal and the average of the filtered signal are plotted in Fig.

3.7. The results of the filtering are comparable with the performance of the wavelet de-noising reported in Fig. 2.15.

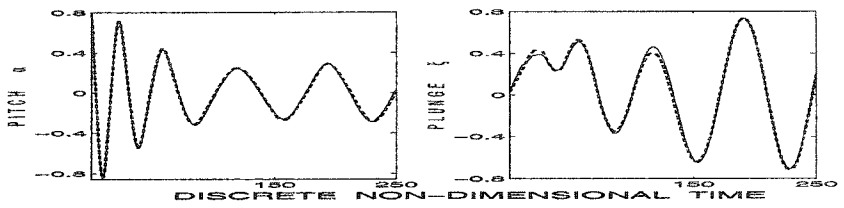

Figure 3.7: Cubic spring model: clean (-) and filtered (- -) signal with the UF

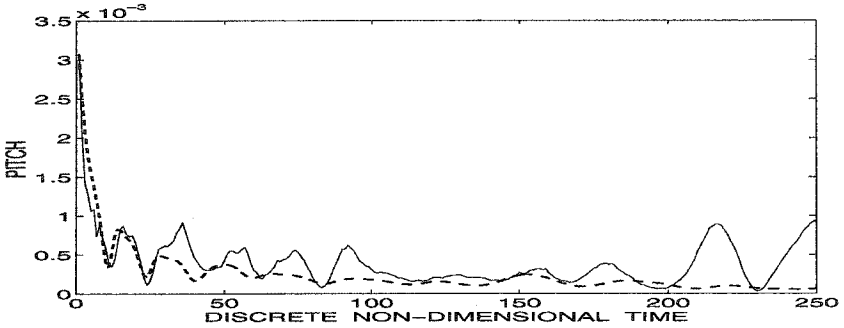

Figure 3.8: Cubic spring model: EVARs (-) and the UF MSEs (- -)

In Fig. 3.8, we present the estimations of the mean square errors (MSEs) calculated using the UF, together with an estimations of the same variances (EVARs) using the Monte Carlo simulations. We have plotted only the results for the pitch motion since the results for the plunge motion are similar. We notice that the performances of the filter are excellent, the UF estimations of the variances are very accurate. Thus, the filter can be regarded to perform good predictions.

The results of the predictor are reported in Figs. 3.9-3.12. We observe that the long term predictions are accurate, not only for the plunging and pitching motions, but also for their derivatives (hidden variables). For the pitch and the plunge motion, the predictions are similar to the results obtained using EXPAR models and displayed in Figs. 2.16-2.17.

The UF can be used also to predict the divergent signals. To illustrate this, we generate noisy divergent signals corresponding to the pitch and plunge motions. The signal corresponding to the pitch motion is displayed in Fig. 3.13 (dots). Since the

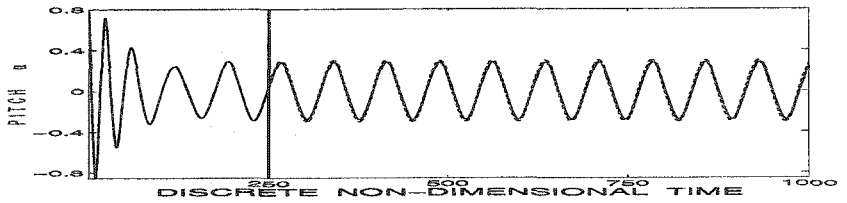

Figure 3.9: Cubic spring model pitch motion: clean '-' and predicted '- -' signal using UF

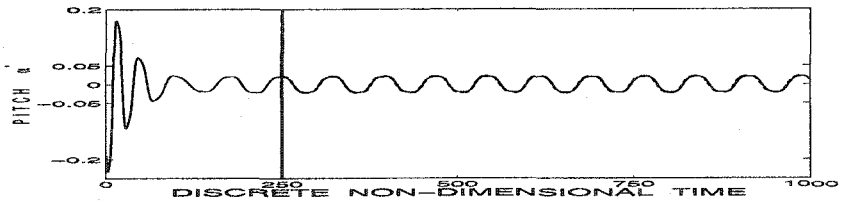

Figure 3.10: Cubic spring model α': clean '-' and predicted '- -' signal using UF

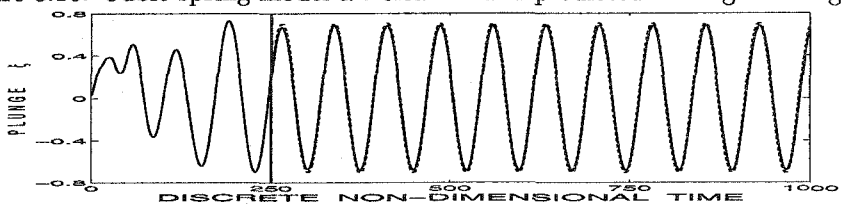

Figure 3.11: Cubic spring model plunge motion: clean '-' and predicted '- -' signal using UF

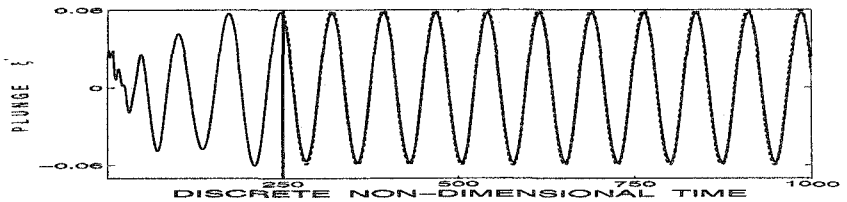

Figure 3.12: Cubic spring model ξ': clean '-' and predicted '- -' signal using UF

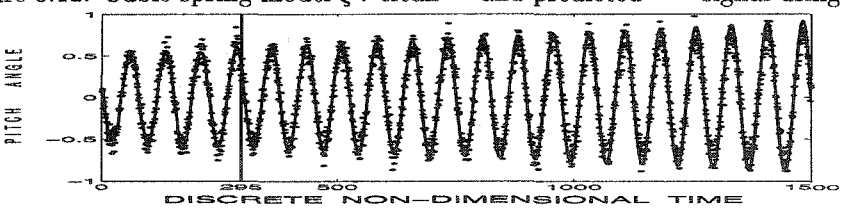

Figure 3.13: Unstable oscillations of the pitch angle (..) and the UF prediction '-'

results for the plunge are similar, we present only the predictions obtained for the pitch angle. The training set contains the first 294 observations (dots). Looking only at this transient data, it seems that the asymptotic state of the system might be a LCO. However, the UF accurately predicts the long term behavior of this divergent system. The filtered signal (solid line, $k = 1$ to $k = 294$) and the predictions (solid line, $k = 295$ to $k = 1500$) for the pitch motion are displayed in Fig. 3.13.

Freeplay model

As in the cubic spring case, we apply the UF directly to the noisy data displayed in Figs. 2.18 and 2.19, and we perform again a Monte Carlo simulation. The training set contains data from $k = 1$ to $k = 295$. For the pitch angle and the plunge displacement, the clean signal and the average of the filtered signal are plotted in Fig. 3.14. Fig. 3.15 displays the average of the variance calculated using the UF (MSEs) together with the estimations of the same variance using the Monte Carlo simulations (EVARs).

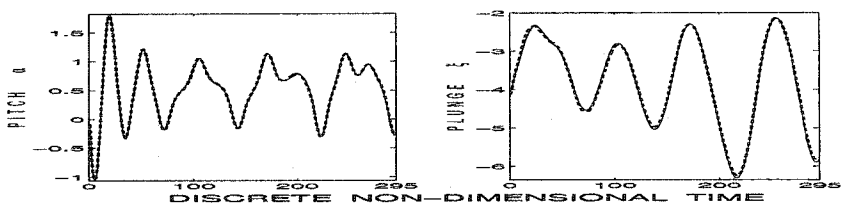

Figure 3.14: Freeplay model: clean (-) and filtered (- -) signal with the UF

Here, the UF is dealing with a continuous piece-wise linear function. Despite the fact that this function is not differentiable, the performances of the filter are excellent. Moreover, if we compare the results displayed in Figs. 2.21 and 3.14, we notice that the UF is better than the wavelet filters.

The results of the predictor are presented in Figs. 3.16-3.19. We see that the long term predictions are very accurate, not only for the plunging deflection and the pitch angle, but also for their derivatives (hidden variables). The results are better than those obtained using the EXPAR models and the noisy signals. They are comparable to the results for the clean pitch motion, shown in Fig. 2.20. Thus,

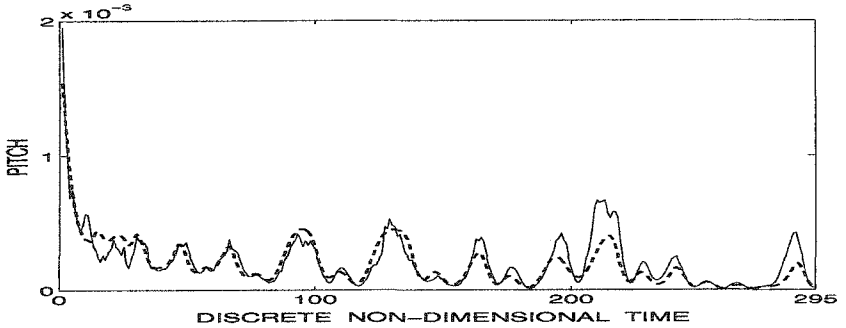

Figure 3.15: Freeplay model: EVARs (-) and the UF MSEs (- -)

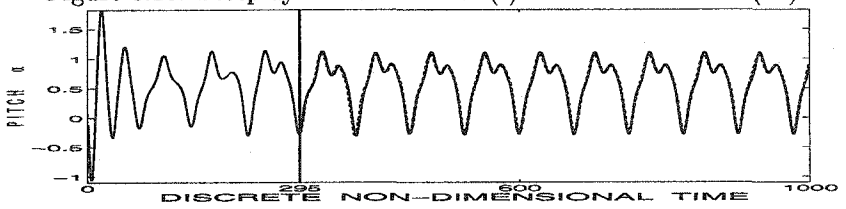

Figure 3.16: Freeplay model, α: clean '-' and predicted '- -' signal using UF

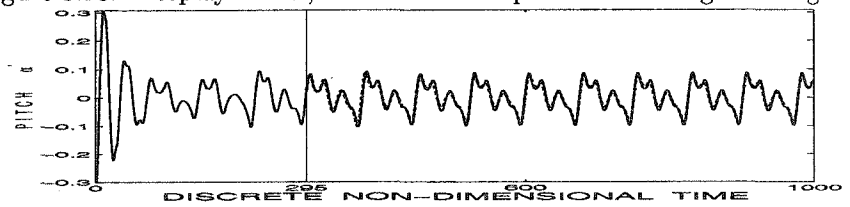

Figure 3.17: Freeplay model, α': clean '-' and predicted '- -' signal using UF

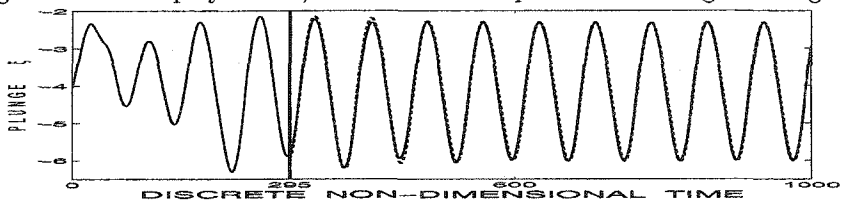

Figure 3.18: Freeplay model, ξ: clean '-' and predicted '- -' signal using UF

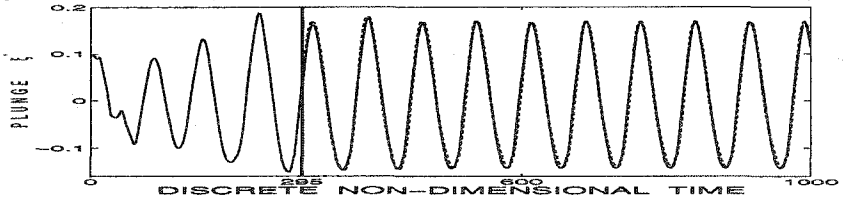

Figure 3.19: Freeplay model, ξ': clean '-' and predicted '- -' signal using UF

the UF is less sensitive to the signal-to-noise ratio, but we need extra information concerning the structure of the nonlinearities in order to implement this method.

3.3.2 Experimental data

Now, we compare the results obtained using the UF or EKF, with those obtained using EXPAR models and reported in Section 2.4.4. In order to apply the UF or the EKF as parameter estimators and predictors, we consider the state-space form [11] of the following mathematical model which corresponds to the experimental data:

$$\begin{bmatrix} m & mx_\alpha b \\ mx_\alpha b & I_\alpha \end{bmatrix} \begin{bmatrix} \xi'' \\ \alpha'' \end{bmatrix} + \begin{bmatrix} c_\xi & 0 \\ 0 & c_\alpha \end{bmatrix} \begin{bmatrix} \xi' \\ \alpha' \end{bmatrix} + \begin{bmatrix} k_\xi & 0 \\ 0 & k_\alpha(\alpha) \end{bmatrix} \begin{bmatrix} \xi \\ \alpha \end{bmatrix} = \begin{bmatrix} -L \\ M \end{bmatrix} \tag{3.14}$$

The term $k_\alpha(\alpha)$ denotes the nonlinear spring stiffness associated with the pitching motion, and it can be approximated by a polynomial [11].

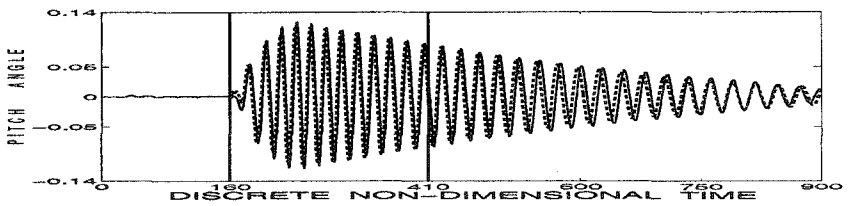

Figure 3.20: Experimental damped oscillation: '-' pitch motion, '- -' the UF prediction

First, we present the results obtained for the pitch motion when the aeroelastic system exhibits a steady damped oscillation (data from the file DN04A.dat). For the plunge motion, the quality of the prediction is similar and it is not reported here. In the UF approach, the training set contains the observations from $k = 160$ to $k = 409$. The measured signal (solid lines), the filtered signal (dashed lines, $k = 160$ to $k = 409$) and the predicted signal (dashed lines, $k = 410$ to $k = 900$) are displayed in Fig. 3.20. The results obtained using an EXPAR model are slightly more accurate (see Fig. 2.30), but the UF also gives a correct prediction for the asymptotic state of the aeroelastic system.

We also compare the results obtained with the EKF against those obtained

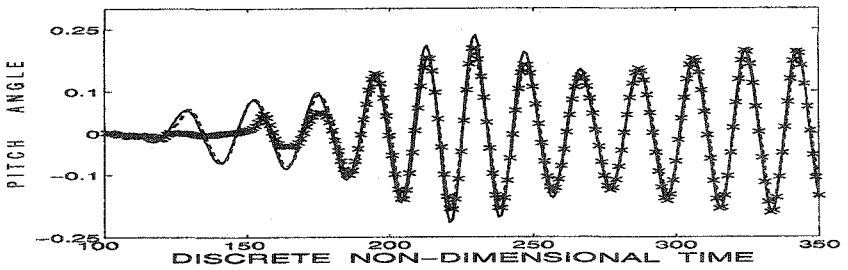

Figure 3.21: Pitch motion: experimental data '-', filtered data using the UF 'x-' and the EKF '- -'

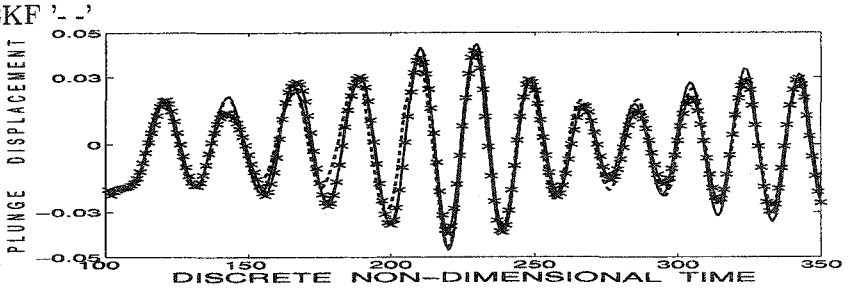

Figure 3.22: Plunge motion: experimental data '-', filtered data using the UF 'x-' and the EKF '- -'

with the UF for the LCOs displayed in Figs 2.28 and 2.29. First, the state space is augmented with the parameters, and the EKF or the UF are used as learning algorithms on the training set (see Figs. 3.21, 3.22). Then, the parameters are fixed, and the filters are used for predictions (see Figs. 3.23, 3.24).

The training set contains only the transient observations corresponding to $t = 100 - 350$. Both EKF and UF accurately predict the amplitude and the frequency of the LCO, but the long term predictions produced by the EKF are slightly shifted (see Figs. 3.23, 3.24). However, for the training set, the results obtained with the EKF are slightly better than those corresponding to the UF. Since the goodness of the fit depends on the initial guess, and thus on the tunning of the filters, it is difficult to conclude which filter has the best performance. Nevertheless, for the aeroelastic data considered here, both EKF and UF give reliable long-term predictions.

For the plunge motion, the long term predictions obtained using the dual EKF and the dual UF are presented in Fig. 3.25. The results obtained for the pitch motion are very similar. The training set contains the observations from $t = 100$

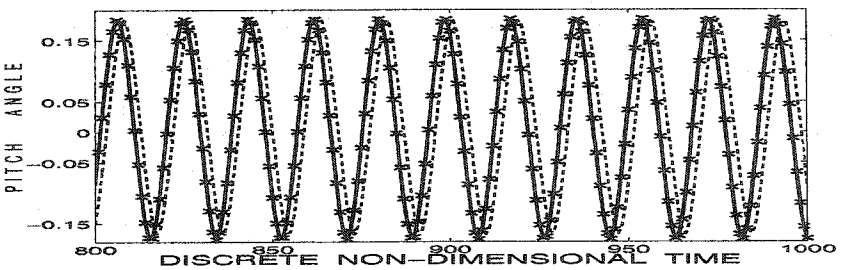

Figure 3.23: Pitch motion: experimental data '-', predicted data using the UF 'x-' and the EKF '- -'

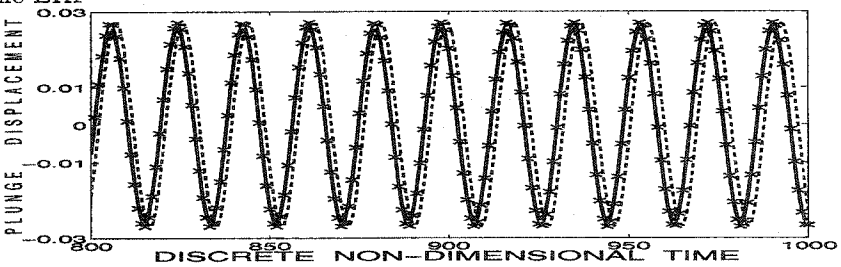

Figure 3.24: Plunge motion: experimental data '-', predicted data using the UF 'x-' and the EKF '- -'

to $t = 350$. We run two filters in parallel, one for the parameters and the other for the state variable. We notice that excellent results are obtained with both methods. Looking at Figs. 3.25 and 3.24, we conclude that, in general, the accuracy of the predictions for the dual method is comparable to the one obtained with the previous method. However, the phase shift error in the predictions using the dual EKF is reduced. This dual approach requires less tuning that the first method presented in this chapter.

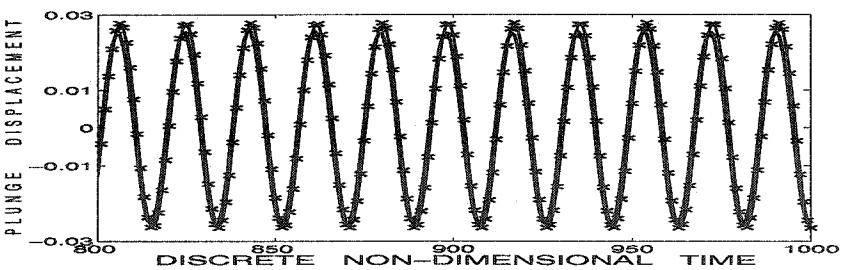

Figure 3.25: Plunge motion: experimental data '-', predicted data using the dual UF 'x-' and EKF '- -'

3.4 The EM algorithm

In this section, we consider a two degree of freedom airfoil oscillating in pitch and plunge with either polynomial restoring forces or freeplay structural nonlinearities. We consider mathematical models that can be expressed by the system of ordinary differential equations 2.16 or 3.14. Now, we propose a method for parameter estimation based on the expectation maximization (EM) algorithm [16].

The EM algorithm is especially useful when it is easier to calculate the likelihood of the model using not only the observed data Y_{obs}, but also the hidden data Y_{hid}. In our case, only the pitch angle and the plunging deflection can be measured, but the aeroelastic model involves also their derivatives. Hence the EM is based upon a data augmentation scheme, such that the observed data are a mapping of the augmented data $Y_{obs} = \mathfrak{m}(Y_{aug})$, where $Y_{aug} = \{Y_{obs}, Y_{hid}\}$. The algorithm starts with an initial guess θ_0 for the unknown parameters and iteratively compute the estimation θ^*. Each iteration consists of two steps: the expectation (E) and the maximization (M) step.

Using the current estimation θ_n of the parameters, the E-step computes the conditional expectation of the augmented data log-likelihood $Q(\theta|\theta_n) = E[\log p(\theta|Y_{aug})|\ Y_{obs}, \theta_n]$. Sometimes an approximation is needed during the E-step. Then, to justify the convergence of the algorithm, it is important to notice that, the negative of the free-energy is maximized [16] with respect to the distribution component:

$$Q_{n+1} = \arg\max_{Q} \mathfrak{F}(Q, \theta_n),$$

where

$$\mathfrak{F}(Q, \theta) = \int Q(Y_{hid}) \log p(\theta|Y_{aug}) \mathrm{d}Y_{hid} - \int Q(Y_{hid}) \log Q(Y_{hid}) \mathrm{d}Y_{hid}.$$

The M-step performs a maximization with respect to the parameters θ:

$$\theta_{n+1} = \arg\max_{\theta} Q(\theta|\theta_n).$$

In term of $\mathfrak{F}$, the M-step can be expressed as:

$$\theta_{n+1} = \arg\max_{\theta} \mathfrak{F}(Q_{n+1}, \theta).$$

Hence, an approximation can be used in either the E-step or the M-step as long as $\mathfrak{F}$ is increasing.

The main advantage of using the EM-algorithm is the guaranteed convergence ([28]), [16]). However, depending on the initial guess, the algorithm can only converge to a local maximum, and it is slower than the methods discussed in Section 3.3. Various strategies for choosing the initial guess are proposed in [4], and methods for accelerating the convergence are presented in [15].

The EM algorithm is a classical method for estimating the parameters of linear systems [23]. For a general nonlinear dynamics, the EM algorithm has been applied using the EKF smoothing [6], and it has been used in conjunction with neural networks [27]. The smoothing was done using the UF and an approximation based on a neural network for the backwards dynamics. In the next two subsections, we present the implementation of the EM algorithm for an aeroelastic system with freeplay and polynomial restoring forces.

3.4.1 The freeplay model

Let us consider a two degree of freedom airfoil oscillating in pitch and plunge with freeplay nonlinearities. The mathematical model is expressed by the system 2.16 and the nonlinearities are given by the formula 2.18.

Introducing four new variables, and taking into account the freeplay nonlinearity, the integro-differential system (2.16) can be reformulated [12] as follows:

$$X_t' = AX_t + F_1 \text{ if } X_t(1) < \alpha_f,$$
$$X_t' = BX_t + F_2 \text{ if } \alpha_f \leq X_t(1) \leq \alpha_f + \delta,$$
$$X_t' = AX_t + F_3 \text{ if } X_t(1) > \alpha_f + \delta,$$

where α_f and δ are constants, and $X(1) = \alpha$ is the first component of the eight dimensional vector $X = [\alpha, \xi, \alpha', \xi', \omega_1, \omega_2, \omega_3, \omega_4]^t$. Here, ω_1, ω_2, ω_3, ω_4 are the four new variables introduced to eliminate the integral terms $C_L(t)$ and $C_M(t)$, A, B are 8×8 matrices, F_i, $i = 1, \ldots, 3$ are eight dimensional vectors. Hence, we have three linear systems that can be solved analytically. With a sufficiently small sampling step τ, the solution can be expressed as

$$\begin{aligned} X_{t+\tau} &= A_1(\tau)X_t + b_1(\tau) \text{ if } X_t(1) < \alpha_f, \\ X_{t+\tau} &= A_2(\tau)X_t + b_2(\tau) \text{ if } \alpha_f \leq X_t(1) \leq \alpha_f + \delta, \\ X_{t+\tau} &= A_1(\tau)X_t + b_3(\tau) \text{ if } X_t(1) > \alpha_f. \end{aligned} \tag{3.15}$$

Since in practice only α and ξ can be measured, we associate with the system 3.15 the following linear discrete switching state-space system:

$$x_{k+1} = A_{S_{k+1}}x_k + b_{S_{k+1}} + v_{k+1}, \tag{3.16}$$

$$y_k = Cx_k + w_k, \tag{3.17}$$

where S_k is a discrete random variable given by

$$S_{k+1} = \begin{cases} 1 & \text{if } x_k(1) < \alpha_f, \\ 2 & \text{if } \alpha_f \leq x_k(1) \leq \alpha_f + \delta, \\ 3 & \text{if } x_k(1) > \alpha_f. \end{cases}$$

Here, A_i, $i = 1, 2, 3$ are 8×8 matrices, $A_3 = A_1$, b_i, $i = 1, 2, 3$ are eight-dimensional vectors, $y_k = [\alpha, \xi]^t$ is the two-dimensional observation vector, x_k is the eight-dimensional state vector, $v_k \sim N(0, Q_{S_k})$ and $w_k \sim N(0, R)$ are independent Gaussian white noise vectors, Q_i, $i = 1, 2, 3$ are 8×8 matrices, R is a 2×2 matrix, and C is the 2×8 matrix

$$C = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \end{bmatrix}.$$

Suppose that we know α_f and δ (e.g. we determine the thresholds α_f and $\alpha_f + \delta$ using

the non-parametric method presented in Section 2.3 and illustrated in Figs. 2.26, 2.27), thus the values of the switching variable S_k are known. We also assume that, conditional on $S_1 = i$, we have $x_1 \sim N(\mu_i, \Sigma_i)$, for $i = 1, 2, 3$. Hence, the unknown parameters of the previous model are $\Theta = \{A_i, b_i, Q_i, R, \mu_i, \Sigma_i, i = 1, 2, 3\}$. Once the parameters are estimated, we can predict the future values of x_i, $i = N+1, N+2, \ldots$, and we can then determine the asymptotic state of the aeroelastic system.

We estimate the parameters Θ using the EM algorithm. First, we augment the data with the hidden variables x_i, $i = 1, \ldots, N$ and we calculate the complete log-likelihood:

$$\log(L) = \log P(x_1, \ldots, x_N, y_1, \ldots, y_N, S_1, \ldots, S_N) = \log P(y_N | y_{N-1}, \ldots, y_1, x_N, \ldots, x_1, S_N, \ldots, S_1) + \ldots + \log P(y_1 | x_N, \ldots, x_1, S_N, \ldots, S_1) + \log P(x_N | x_{N-1}, \ldots, x_1, S_N, \ldots, S_1) + \ldots + \log P(x_1 | S_N, \ldots, S_1) + \log P(S_N, S_{N-1}, \ldots, S_1).$$

Using the equations (3.16) and (3.17), we obtain:

$$\log(L) = -5N \ln(2\pi) - \frac{1}{2}\sum_{i=1}^{N} \ln(|R|) - \frac{1}{2}\sum_{i=1}^{N} (y_i - Cx_i)^t R^{-1} (y_i - Cx_i) - \frac{1}{2}\sum_{i=2}^{N} \ln(|Q_{S_i}|) - \frac{1}{2}\sum_{i=2}^{N} (x_i - A_{S_i} x_{i-1} - b_{S_i})^t Q_{S_i}^{-1} (x_i - A_{S_i} x_{i-1} - b_{S_i}) - \frac{1}{2}\ln(|\Sigma_{S_1}|) - \frac{1}{2}(x_1 - \mu_{S_1})^t \Sigma_{S_1}^{-1} (x_1 - \mu_{S_1}) + \log P(S_N, \ldots, S_1).$$

The EM algorithm iteratively maximizes $\hat{E} = E[\log(L) | y_1, \ldots, y_N, S_1 = i_1, \ldots, S_n = i_n]$. Using the previous formulas and $tr(AB) = tr(BA)$, where tr is the notation for the trace of a matrix, we get

$$\hat{E} = -5N \ln(2\pi) - \frac{N}{2}\log(|R|) - \frac{1}{2}\sum_{n=1}^{N} tr[R^{-1}(y_n y_n^t - y_n x_{n|N}^t C^t - C x_{n|N} y_n^t + C P_n C^t)] - \frac{1}{2}\sum_{n=2}^{N} \log(|Q_{i_n}|) - \frac{1}{2}\sum_{n=2}^{N} tr[Q_{i_n}^{-1}(P_n - A_{i_n} P_{n,n-1}^t - P_{n,n-1} A_{i_n}^t + A_{i_n} P_{n-1} A_{i_n}^t - x_{n|N} b_{i_n}^t - b_{i_n} x_{n|N}^t + A_{i_n} x_{n-1|N} b_{i_n}^t + b_{i_n} x_{n-1|N}^t A_{i_n}^t + b_{i_n} b_{i_n}^t)]$$

$$-\frac{1}{2}\log(|\Sigma_{i_1}|)-\frac{1}{2}tr[\Sigma_{i_1}^{-1}(P_1-\mu_{i_1}x_{1|N}^t-x_{1|N}\mu_{i_1}^t+\mu_{i_1}\mu_{i_1}^t)]$$

Here, $P_n = E[x_n x_n^t|y_1,\ldots,y_N]$, $P_{n,n-1} = E[x_n x_{n-1}^t|y_1,\ldots,y_N]$, and $x_{n|N} = E[x_n|$ $y_1,\ldots,y_N]$ are the smoothed values. They can be calculated using the Kalman filtering and smoothing [23]. For the M-step, analytical update equations for the parameters Θ can be found taking derivatives with respect to the parameters Θ in the previous formula for $\hat{E}$ [23].

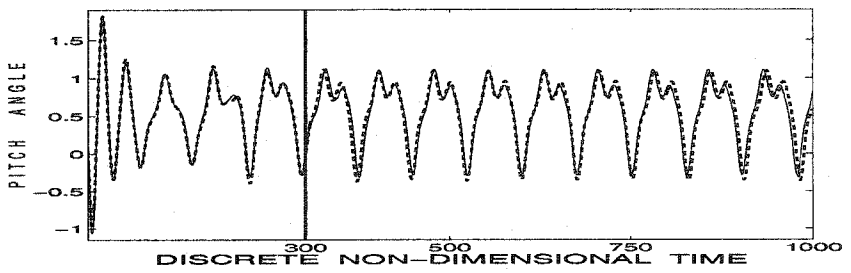

Figure 3.26: Pitch: clean data '-', filtered or predicted data using the UF '- -'

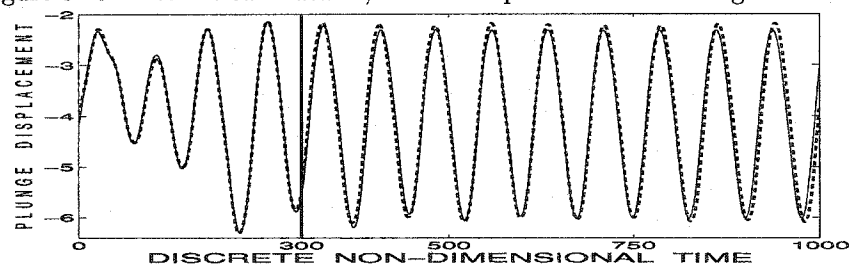

Figure 3.27: Plunge: clean data '-', filtered or predicted data using the UF '- -'

We apply the EM algorithm for the simulated noisy data displayed in Figs. 2.18, 2.19. The training set contains the first 300 observations, and is used for parameter estimation. In Figs 3.26 and 3.27, we compare the clean signal with the filtered signal (the first 300 observations) and the predicted signal (the last 700 observations). The amplitude and the frequency of the limit cycle oscillations are accurately predicted. The results reported in Figs. 3.16 and 3.18 are slightly better than those presented here, but the tuning required for the methods presented in Section 3.3 is very extensive. Moreover, the results obtained using the EM algorithm are more accurate than those obtained with the EXPAR or SETAR models displayed

in Figs. 2.22, 2.23.

3.4.2 The polynomial model

Now, we consider the model 3.14 corresponding to a wing section with two degrees of freedom and a control surface. The system of ordinary differential equations 3.14 can be rewritten in a state-space form [11]. Using the Euler integration scheme, the associated discrete system can be expressed as:

$$x_{k+1} = Ax_k + \begin{bmatrix} 0_{2\times 4} \\ B \end{bmatrix} [x_k^2(1), x_k^3(1), x_k^4(1), x_k^5(1)]^t + \begin{bmatrix} 0_{2\times 1} \\ b \end{bmatrix} + v_{k+1},$$

$$y_{k+1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} x_{k+1} + w_{k+1}.$$

Here $x_k = [\alpha, \xi, \alpha', \xi']^t$ is the state of the system at time step k, and $y_{k+1} = [\alpha, \xi]^t$ is the observation vector. We assume that the white noise vectors, v_k and w_k, are Gaussian and from uncorrelated white sequences. The 4×4 matrix A, the 2×4 matrix B, the two dimensional vector b , and the two covariance matrices corresponding to the noise vectors v_k and w_k represent the unknown parameters of the aeroelastic system.

The previous system has a nonlinear state equation. The EKF or the UF can be used for filtering, smoothing and prediction. Moreover, for the E-step, the likelihood and the conditional expectation of the likelihood can be approximated based on the linearization of the previous system as presented in (3.7). The formulas are similar to those reported in the previous section, but we have to use the EKF or UF smoothing (see Section 3.2.2) to compute the conditional expectations $x_{n|N}$, variances P_n, and covariances $P_{n,n-1}$.

We illustrate the parameter estimation based on the EM algorithm for the experimental data recorded at the Texas A&M University. First, we consider the data displayed in Figs. 2.28 and 2.29. In order to compare the learning algorithm with the neural network approach in [25], the training sets are formed with the observations corresponding to $k = 200 - 300$. Looking only at the training sets, one may

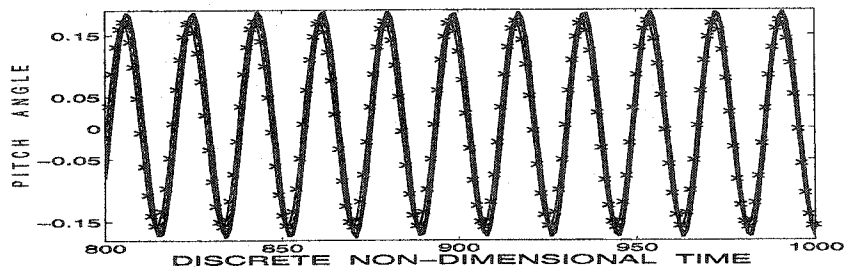

Figure 3.28: Pitch: experimental data '-', predicted data using the EM algorithm and the UF 'x-' or the EKF '- -'

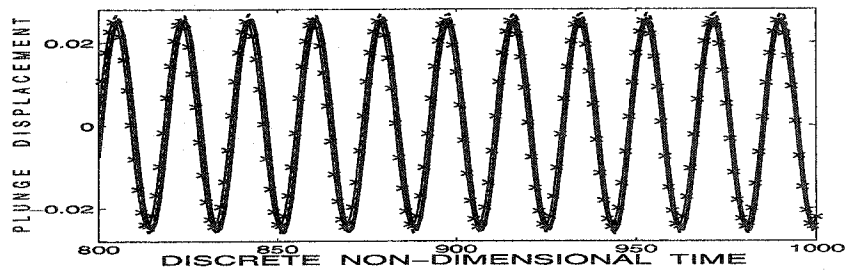

Figure 3.29: Plunge: experimental data '-', predicted data using the EM algorithm and the UF 'x-' or the EKF '- -'

guess that the signals may eventually decay to zero. However, both implementations of the EM algorithm accurately predict the LCO with the correct amplitude and frequency. The results obtained using the EM algorithm with the UF or the EKF smoothing are displayed in Figs. 3.28, 3.29. Practically no tuning was needed to obtain these results. Thus, this method is an attractive alternative to a neural network approach. Similar results are obtained using the transient observations corresponding to $k = 100 - 350$ for training, as in Figs. 3.21 - 3.24. From the graphs displayed in Figs. 3.28, 3.29, we notice the similarity of the results obtained using the UF and EKF.

Finally, we present the results obtained using the EM algorithm and the UF for the experimental data displayed in Fig. 3.30. These data have been also recorded at the Texas A&M University (the file DN04F.dat). The training set corresponds to $k = 250$ to $k = 519$ as indicated by the two vertical lines shown in Fig. 3.30. The predictions (dashed lines) start at $k = 520$ and they are in good agreement with the experimental data (solid lines). The results obtained for the pitch motion are

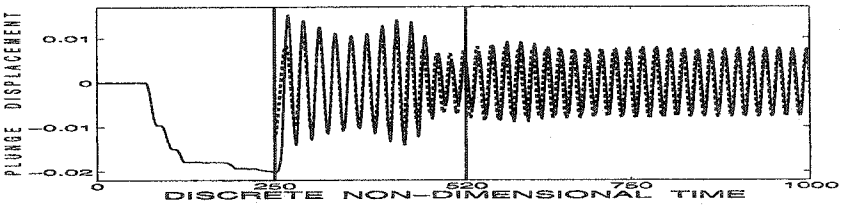

Figure 3.30: Plunge: experimental data '-', filtered or predicted data using the EM algorithm and the UF '- -'

similar and are not reported here.

In Fig. 3.31, we present the estimations of the likelihood corresponding to each iteration of the EM algorithm. One attractive feature of the EM algorithm is the monotonic increasing of the estimations of the likelihood, even for this nonlinear case. In addition, this method is guaranteed to converge at least to a local maximum. On the other hand, the results are still dependent on the initial guess and the algorithm converges rapidly at the beginning, and then deteriorates with a slow convergence.

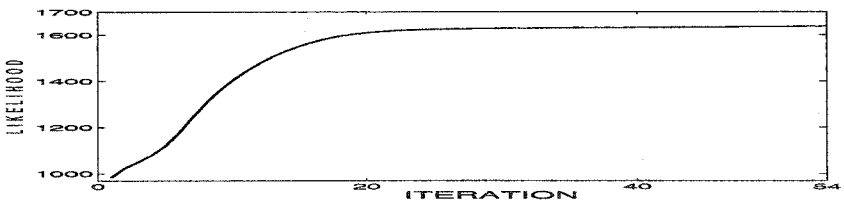

Figure 3.31: Estimation of the likelihood using the UF '-'

Bibliography

[1] B.D.O. Anderson and J.B. Moore. *Optimal filtering.* Pretince Hall, Inc., Englewood Cliffs, New Jersey 07632, 1979.

[2] M. Athans, R.P. Wishner, and A. Bertolini. Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements. *IEEE Trans. Aut. Control*, 13(6):504–518, 1968.

[3] A. Balakrishnan. *Kalman Filtering Theory.* Optimization Software, Inc., New York, 1984.

[4] C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Comput. Statist. Data Anal.*, 41:561–575, 2003.

[5] J. J. Block and T. W. Strganac. Applied active control for a nonlinear aeroelastic structure. *AIAA J. Guid.,Contr., Dynamics*, 21(6):838–845, 1998.

[6] Z. Ghahramani and S. Roweis. Learning nonlinear dynamical systems using an EM algorithm. In M.S. Kearns, S.A. Solla, and Cohn D.A., editors, *Advances in Neural Information Processing Systems*, pages 599–605. MIT Press, 1999.

[7] A.H. Jazwinski. *Stochastic processes and filtering theory.* Academic Press, New York and London, 1970.

[8] S. Julier, J.K. Uhlmann, and H.F. Durrant-White. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Trans. Aut. Control*, 45(3):477–482, 2000.

[9] S. J. Julier and J. K. Uhlmann. The scaled unscented transformation. In *Proceedings of the IEEE American Control Conference*, pages 4555–4559, Anchorage AK, USA, 8–10 May 2002.

[10] S.J. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte. A new approach for filtering nonlinear systems. In *American Control Conference, Seattle, Washington*, pages 1628–1632, 1995.

[11] J. Ko, T. W. Strganac, and A. J. Kurdila. Stability and control of a structurally nonlinear aeroelastic system. *AIAA J. Guid.,Contr., Dynamics*, 21(5):718–725, 1998.

[12] B. H. K. Lee, S. J. Price, and Y. S. Wong. Nonlinear aeroelastic analysis of airfoils: bifurcation and chaos. *Progress in Aerospace Sciences*, 35:205–334, 1999.

[13] L. Ljung. Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems. *IEEE Trans Aut. Control*, AC-24(1):36–50, 1979.

[14] M. G. Lyons, R. Vepa, S. C. McIntosh, and D. B. DeBra. Control law synthesis and sensor design for active flutter suppression. In *Proceedings of the AIAA Guidance and Control Conference*, pages 1–29, New York, 1973. AIAA Paper 73-832.

[15] X.L. Meng and D. Van Dyk. The EM algorithm - an old folk-song to a fast new tune. *J.R. Statist. Soc.B*, 59(3):511–567, 1997.

[16] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, an other variants. In M.I. Jordan, editor, *Learning in Graphical Models*. 1998.

[17] L.W. Nelson and E. Stear. The simultaneous on-line estimation of parameters ans states in linear systems. *IEEE Trans. Aut. Control*, pages 94–98, 1976.

[18] C. A. Popescu and Y. S. Wong. Applications of the EM algorithm for the study of the aeroelastic systems with structural nonlinearities. In *Proceedings of the*

3rd International Workshop on Scientific Computing, Hong Kong, 2003. Nova Science Publisher Inc.

[19] C. A. Popescu and Y. S. Wong. A nonlinear statistical approach for aeroelastic response prediction. *AIAA J. Guid.,Contr., Dynamics*, 26(4):565–572, 2003.

[20] C. A. Popescu and Y. S. Wong. The Unscented and Extended Kalman filter for systems with polynomial restoring forces. In *Proceedings of the 44th AIAA/ ASME/ ASCE/ AHS/ ASC/ Structures, Structural Dynamics, and Materials Conference*, Norfolk, VA, USA, 2003. AIAA Paper 2003-1410.

[21] K. Reif, S. Gunther, E. Yaz, and R. Unbehauen. Stochastic stability of the discrete-time extended Kalman filter. *IEEE Trans. Aut. Control*, 44(4):714–728, 1999.

[22] R. Roy and R. Walker. Real-time flutter identification. Technical report, NASA Science and Technical Information Branch, October 1985. NASA CR-3933.

[23] R.H. Shumway and D.S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.*, 3(4):253–264, 1982.

[24] F. Takens. Detecting strange attractors in turbulence. *Lectures notes in mathematics*, 898:366–381, 1981.

[25] O. Voitcu and Y. S. Wong. Neural network approach for nonlinear aeroelastic analysis. *AIAA J. Guid.,Contr., Dynamics*, 26(1):99–105, 2003.

[26] E. A. Wan and A. T. Nelson. Dual Kalman filtering methods for nonlinear prediction, estimation, and smoothing. In M. I. Jordan et al., editor, *Advances in Neural Information Processing Systems*. MIT Press, 1997.

[27] E. A. Wan, R. Van der Merwe, and A. T. Nelson. Dual estimation and the unscented transformation. In M. I. Jordan et al., editor, *Advances in Neural Information Processing Systems*. MIT Press, 2000.

[28] C.F.J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Chapter 4

Ergodicity in models with hidden Markov switching

4.1 Introduction

In the previous chapters, we have discussed methods for parameter estimation and prediction for nonlinear aeroelastic models. In this chapter, instead of a threshold switching, we consider models with Markovian switching. We present two different models with hidden discrete Markov switchings: one with mutually independent observations, and the other with Markovian observations given the sequence of the states of the hidden Markov switching. For both models, we study the ergodic properties of the prediction filters corresponding to the hidden Markov models.

A HMM is formed by a hidden Markov chain $\{S_n, n \geq 1\}$ and a stochastic process $\{Y_n, n \geq 1\}$, with distributions depending on $\{S_n, n \geq 1\}$. Usually, the hidden sequence $\{S_n, n \geq 1\}$ is a finite homogeneous Markov chain, and the observations $\{Y_n, n \geq 1\}$ are mutually independent given the sequence $\{S_n, n \geq 1\}$.

Example 4.1 An example of a HMM is a model ([12]) with observations $\{Y_n, n \geq$

The material presented in this chapter was submitted for publication in [14]

1} of the form

$$Y_n = h(S_n) + V_n,$$

where $\{V_n, n \geq 1\}$ is a Gaussian white noise sequence independent of $\{S_n, n \geq 1\}$, with symmetric and positive definite covariance matrix D, (i.e. $V_n \sim N(0, D)$), and h is a mapping from the finite set S to $\mathbb{R}^d$.

Two classical inference methods for HMMs are the forward-backward and the Baum-Welch algorithms ([13], pp. 329-333). During the forward part of the forward-backward algorithm, the prediction filter, $\{\mathbf{P}_\bullet(S_n|Y_{n-1}, \ldots, Y_1)\}$ is computed, where $(\Omega, \mathcal{F}, \mathbf{P}_\bullet)$ is the probability space. The backwards part uses the filtered values for smoothing. The Baum-Welch algorithm is a special case of the expectation maximization (EM) algorithm, and is used for parameter estimations. To infer the posterior probabilities of the hidden states in the E step, the forward-backward algorithm is applied. Hence, for estimation problems in HMMs, the ergodic properties of the prediction filter are of a special interest.

Different assumptions can be made about the state space and the distributions of $\{S_n, n \geq 1\}$ and $\{Y_n, n \geq 1\}$. Under compactness or local compactness hypotheses, the existence of an invariant probability distribution for the prediction filters is proved in [10] and [16]. Kaijser [9] studied the ergodicity of the filter under the assumption of subrectangularity, and his approach is based on the dependence with complete connections.

The first model considered here is a HMM similar to the one presented in [12]. With this model, we associate a random system with complete connections (RSCC), and as a consequence of the properties of the RSCC, we obtain the ergodicity of the Markov chain formed by the prediction filter. Our methods are related to the approach in [9], but we work under different hypotheses, and we illustrate how the Ionescu Tulcea - Marinescu ergodic Theorem [6] can be applied for HMMs.

The focus of this chapter is in the switching state-space models (SSM), which are a generalization of the HMMs and the state-space models. Since the hybrid representation combines discrete and continuous dynamics, they are capable to model many complex phenomena ([5], [15]). A SSM is formed with the observation process

$\{Y_n, n \geq 1\}$ and two hidden Markov chains: the finite Markov chain $\{S_n, n \geq 1\}$ and the state process $\{X_n, n \geq 1\}$.

Example 4.2 As a generalization of example 4.1, we consider the following hybrid model with observations $\{Y_n, n \geq 1\}$ and two sequences of hidden states $\{S_n, n \geq 1\}$ and $\{X_n, n \geq 0\}$ with $X_0 \sim N(\mu, \Sigma)$:

$$X_n = AX_{n-1} + W_n, \quad W_n \sim N(0, H), n \geq 1. \tag{4.1}$$

$$Y_n = C(S_n)X_n + V_n. \tag{4.2}$$

Here $\{S_n, n \geq 1\}$ is a homogeneous Markov chain with values on a finite set S, A is the $m \times m$ transition matrix, $C(i)$, $i \in S$, are the $d \times m$ output matrices for the state-space model, and H is the symmetric and positive definite covariance matrix of the normal distribution corresponding to the independent random vectors W_n, $n \geq 1$. The noise sequences $\{W_n, n \geq 1\}$ and $\{V_n, n \geq 1\}$ are independent. Moreover, the sequences $\{S_n, n \geq 1\}$ and $\{X_n, n \geq 0\}$ are independent. Conditional on $\{S_n, n \geq 1\}$, V_n, $n \geq 1$, are Gaussian and mutually independent such that, if $S_n = i$, $i \in S$, then $V_n \sim N(0, R(i))$, where $R(i)$ is the $d \times d$ symmetric and positive definite covariance matrix.

Since the SSM are mixture models with an exponentially increasing number of components, deterministic inference algorithms become intractable rapidly. For example, if we attempt to implement the EM algorithm for the model (4.1) - (4.2) in the E step, we need to use approximations for the smoothing. A possible solution is to use a stochastic algorithm, namely the Gibbs sampler [3]. This approach is based on an iterative algorithm. At each step l, we draw a sample $\{X_n(l)\}$ for the state variable. Then, using the prediction filter $\{\mathbf{P}_\bullet(S_n | Y_{n-1}, X_{n-1}(l) \ldots, Y_1, X_1(l))\}$, we draw a sample for the hidden sequence $\{S_n, n \geq 1\}$.

The second model considered in this chapter is a SSM. For this case, we study the ergodic properties of the prediction filter used in every step of the Gibbs sampler. Our approach extends the methods applied for HMMs in [12].

4.2 Random Systems with Complete Connections

In this section we summarize some results concerning the random systems with complete connections [8].

Definition 4.1 A random system with complete connection (RSCC) is a quadruple $\{(W, \mathcal{W}), (X, \mathcal{X}), u, P\}$, formed with

1. the measurable spaces $(W, \mathcal{W})$ and $(X, \mathcal{X})$;
2. the $(\mathcal{W} \times \mathcal{X}, \mathcal{W})$ measurable map $u : W \times X \to W$;
3. the transition probability function P from $(W, \mathcal{W})$ to $(X, \mathcal{X})$.

Let define recursively the maps $u^{(n)} : W \times X^n \to W$:

$$u^{(n+1)}(w, x^{(n+1)}) = \begin{cases} u(w, x_1), & \text{if } n = 0, \\ u(u^{(n)}(w, x^{(n)}), x_{n+1}), & \text{if } n > 0, \end{cases} \tag{4.3}$$

where $x^{(n)} = (x_1, \ldots, x_n) \in X^n$, for any positive integer n. We also define

$$P_r(w, A) = \begin{cases} P(w, A), & \text{if } r = 1, \\ \int_X P(w, dx_1) \ldots \int_X P(wx^{(r-1)}, dx_r) 1_A(x^{(r)}), & \text{if } r > 1, \end{cases} \tag{4.4}$$

$$P_r^n(w, A) = P_{n+r-1}(w, X^{n-1} \times A) \tag{4.5}$$

for any positive integers r, n and any $w \in W$, $A \in \mathcal{X}^r$. Here, and whenever no confusion is possible, we denote by wx^n the element $u^{(n)}(w, x^{(n)}) \in W$.

Definition 4.2 The homogeneous RSCC $\{(W, \mathcal{W}), (X, \mathcal{X}), u, P\}$ is uniformly ergodic if for any positive integer r there exist a probability P_r^∞ on $\mathcal{X}^r$ such that

$$\lim_{n \to \infty} P_r^n(w, A) = P_r^\infty(A),$$

uniformly with respect to $w \in W$, $A \in \mathcal{X}^r$, and r.

Definition 4.3 The RSCC $\{(W, \mathcal{W}), (X, \mathcal{X}), u, P\}$ is a RSCC with contraction if (W, d) is a separable metric space, $r_1 < \infty$, $R_1 < \infty$, and there exists a positive

integer k such that $r_k < 1$. Here, for any positive integer j we have

$$r_j = \sup_{w' \neq w''} \int_{X^j} P_j(w', dx^{(j)}) \frac{d(w'x^{(j)}, w''x^{(j)})}{d(w', w'')},$$

$$R_j = \sup_{A \in \mathcal{X}^j} \sup_{w' \neq w''} \frac{|P_j(w', A) - P_j(w'', A)|}{d(w', w'')}.$$

The following result reveals the probabilistic meaning of the quantities defined in 4.3 -4.5.

Theorem 4.1 *Let $\{(W, \mathcal{W}), (X, \mathcal{X}), u, P\}$ be a RSCC. For any $w_0 \in W$ there exist a probability space (Ω, Σ, P_{w0}) and two sequences of random variables $\{\xi_n, n \geq 1\}$ and $\{\tau_n, n \geq 0\}$ defined on Ω and with values on X and W respectively, such that $\tau_0 = w_0$, and for any positive integers m, n, r and any set $A \in \mathcal{X}^r$ we have $\tau_n = w_0\xi^{(n)}$,*

$$P_{w0}([\xi_n, \ldots, \xi_{n+r-1}] \in A) = P_r^n(w_0, A)$$
$$P_{w0}([\xi_{n+m}, \ldots, \xi_{n+m+r-1}] \in A | \xi^{(n)}) = P_{w0}([\xi_{n+m}, \ldots, \xi_{n+m+r-1}] \in A | \xi^{(n)}, \tau^{(n)})$$
$$= P_r^m(\tau_n, A), \quad P_{w0} \text{ a.s.}.$$

Moreover, the sequence $\{\tau_n, n \geq 0\}$ is a homogeneous Markov chain whose initial distribution is concentrated in w_0 and whose transition operator is given by

$$Uf(w) = \int_X P(w, dx) f(wx),$$

for any bounded, measurable, real valued function f defined on W.

PROOF. See Theorem 1.1.2 in [8]. □

Remark 4.1 The Markov chain $\{\tau_n, n \geq 0\}$ is called the Markov chain associated with the RSCC. For any bounded, measurable, real valued function f defined on W

and any positive integer n, the iterates of the operator U are given by

$$U^n f(w) = \int_{X^n} P_n(w, dx^{(n)}) f(wx^{(n)}), \quad w \in W.$$

Remark 4.2 If $\{(W, \mathcal{W}), (X, \mathcal{X}), u, P\}$ is an RSCC with contraction and (W, d) is a compact space, then the associated Markov chain is compact (see Proposition 3.2.3 in [8]).

Let $Y = \cup_{r>0} X^r$. For any $A \in \mathcal{X}$ we denote by $Y_n(A)$ the set of the elements of Y which contain among their components at least n which belong to A. Let $B_{L,A}(W, \mathcal{W})$ be the collection of bounded, measurable, real valued function f defined on W for which there exist a sequence $\{l_n, n > 0\}$ of positive numbers such that, for any positive integer n, any $x^{(r)} \in Y_n(A)$, w', $w'' \in W$ we have

$$|f(w'x^{(r)}) - f(w''x^{(r)})| \leq l_n.$$

Proposition 4.1 *Let $\{(W, \mathcal{W}), (X, \mathcal{X}), u, P\}$ be a RSCC for which there exist $A_0 \in \mathcal{X}$ and a positive integer ν such that*

1. *There exists $\gamma > 0$ such that for all $w \in W$ we have $P_1^\nu(w, A_0) \leq \gamma$;*

2. *If, for any positive integer n, we define*

$$a_n = \sup_{w', w'' \in W, x^{(r)} \in Y_n(A_0), A \in \mathcal{X}} |P(w'x^{(r)}, A) - P(w''x^{(r)}, A)|,$$

then we have $\sum_{n>0} a_n < \infty$.

For any positive integer h and all $w', w'' \in W$, $x^{(r)} \in Y_n(A_0)$ and $f \in B_{L,A_0}(W, \mathcal{W})$ we have

$$|U^h f(w'x^{(r)}) - U^h f(w''x^{(r)})| \leq \frac{2\nu}{\gamma} (\text{osc} f) \sum_{j \geq n} a_j + l_n,$$

where osc $f = \sup_{v,w \in W} |f(v) - f(w)|$.

PROOF. See Proposition 2.4.1 in [8]. □

4.3 Hidden Markov Models

We begin with some definitions and notations. For any set A we denote the indicator function by 1_A. Let $S = \{1, \ldots, M\}$ be a finite set, and let $\mathcal{B}(\mathbb{R}^n)$ be the collection of the Borel sets on $\mathbb{R}^n$, for any integer $n > 0$. We define the following L^1 norms

$$\|u\|_1 = \sum_{i=1}^{n} |u_i|, \quad \|Z\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |Z_{i,j}|,$$

where $u = (u_i) \in \mathbb{R}^n$, and $Z = (Z_{i,j})$ is a $n \times n$ matrix. Let $\|\cdot\|$ be the Euclidian (L^2) norm on $\mathbb{R}^n$ and $\|\cdot\|_2$ be the corresponding L^2 matrix norm:

$$\|Z\|_2 = \max_{\|x\|=1} \|Zx\| = \rho(Z^t Z)^{1/2},$$

where $\rho(\cdot)$ denotes the spectral radius of the matrix (which is equal to the maximum eigenvalue of the matrix, for any symmetric and non-negative definite matrix). By definition, the matrix Z is primitive with index of primitivity r, if Z^r has only positive entries and r is the smallest integer with this property.

We define

$$W = \{w \in \mathbb{R}^M : w_i \ge 0, i \in S, \sum_{i=1}^{M} w_i = 1\}, \tag{4.6}$$

and consider on W the topology induced by the Euclidian space $\mathbb{R}^M$. Let $\mathcal{W}$ denote the collection of the Borel sets on W. If $L(W)$ denotes the set of real-valued, bounded and Lipschitz continuous functions defined on W, then $L(W)$ is a Banach space for the norm $\|\cdot\|_{BL}$ defined by $\|g\|_{BL} = \|g\| + s(g)$, where

$$\|g\| = \sup_{p \in W} |g(p)|, \quad s(g) = \sup_{p \ne p', p, p' \in W} \frac{|g(p) - g(p')|}{\|p - p'\|_1}.$$

For any bounded linear operator V from $L(W)$ to $L(W)$, let

$$\|V\|_{BL} = \sup_{\|g\|_{BL}=1} \|Vg\|_{BL}.$$

For the first model, we consider the same HMM as in [12], which is formed

by the unobserved random sequence $\{S_n, n \geq 1\}$, and the observations sequence $\{Y_n, n \geq 1\}$, defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P}_\bullet)$ with values in S and in $\mathbb{R}^d$, respectively. We suppose that the sequence $\{S_n, n \geq 1\}$ is a homogeneous Markov chain with the initial probability distribution $\mathbf{p}_\bullet = (\mathbf{p}_\bullet{}^i)$, and the transition probability matrix $Q = (q^{i,j})$, $i, j \in S$. Hence, for any $i, j \in S$ and any integer $n \geq 1$, we have $\mathbf{P}_\bullet(S_1 = i) = \mathbf{p}_\bullet{}^i$, and $\mathbf{P}_\bullet(S_{n+1} = j | S_n = i) = q^{i,j}$. We also suppose that the observations $\{Y_n, n \geq 1\}$ are mutually independent given the sequence $\{S_n, n \geq 1\}$, that is, for any $i_1, \ldots i_n \in S$ and any sets $A_n, \ldots A_1 \in \mathcal{B}\left(\mathbb{R}^d\right)$, we have

$$\mathbf{P}_\bullet(Y_n \in A_n, \ldots, Y_1 \in A_1 | S_n = i_n, \ldots, S_1 = i_1) = \prod_{k=1}^{n} \mathbf{P}_\bullet(Y_k \in A_k | S_k = i_k).$$

Moreover, the conditional probability distribution of the observation Y_n given that $\{S_n = i\}$ is absolutely continuous with respect to a non-negative and σ-finite measure λ, and it has a positive density $b_\bullet = (b^1, \ldots, b^M)^t$. Thus, for any $A \in \mathcal{B}\left(\mathbb{R}^d\right)$ and any integers $n \geq 1$, $i \in S$, we have

$$\mathbf{P}_\bullet(Y_n \in A | S_n = i) = \int_A b^i(y) \lambda(dy).$$

Remark 4.3 The HMM presented in example 4.1 satisfies all the previous assumptions. In addition, this model corresponds to the Bayesian network displayed in Fig. 1.2.

4.3.1 The forward - backward and the Baum - Welch algorithms

Let suppose the matrix Q and the vector $\mathbf{p}_\bullet$ are unknown. To estimate the parameters $\theta = \{Q, \mathbf{p}_\bullet\}$, we use a special case of the EM algorithm, which is the Baum-Welch algorithm [4]. Augmenting with the hidden variables, we have

$$\log \mathbf{P}_\bullet(Y_1, S_1, \ldots, Y_N, S_N) = \sum_{i \in S} 1_{\{i\}}(S_1) \log \mathbf{P}_\bullet(S_1) + \sum_{n=1}^{N} \sum_{i \in S} 1_{\{i\}}(S_n)$$

$$\times \log \mathbf{P}_\bullet(Y_n|S_n = i) + \sum_{n=2}^{N} \sum_{i\in S} \sum_{j\in S} 1_{\{i\}}(S_n) 1_{\{j\}}(S_{n-1}) \log \mathbf{P}_\bullet(S_n = i|S_{n-1} = j).$$

Since the values of $\{S_n, n = 1, \ldots N\}$ are unknown, we condition the previous equation with respect to the observations $\{Y_n, n = 1, \ldots N\}$.

$$E\left[\log \mathbf{P}_\bullet(Y_1, S_1, \ldots, Y_N, S_N)\middle| Y_1, \ldots, Y_N\right] = \sum_{i\in S} E[S_1 = i|Y_1, \ldots, Y_N] \log \mathbf{p}_\bullet^i + \sum_{n=1}^{N}$$

$$\sum_{i\in S} E[S_n = i|Y_1, \ldots, Y_N] \log b^i(Y_n) + \sum_{i,j\in S} \log q^{j,i} \sum_{n=2}^{N} E[S_n = i, S_{n-1} = j|Y_1, \ldots, Y_N]$$

Taking the derivatives with respect to the parameters θ, we obtain the following M-step:

$$q^{i,j} = \frac{\sum_{n=2}^{N} E[S_n = j, S_{n-1} = i|Y_1, \ldots, Y_N]}{\sum_{n=2}^{N} E[S_{n-1} = i|Y_1, \ldots, Y_N]}, \quad \mathbf{p}_\bullet^i = E[S_1 = i|Y_1, \ldots, Y_N].$$

The necessary expectations are then computed in the E-step using the forward-backward algorithm [4].

The forward part recursively compute $\alpha_n = (\alpha_n^i, i \in S)$, where

$$\begin{aligned}\alpha_n^i &= \mathbf{P}_\bullet(S_n = i, Y_1, \ldots, Y_n) \\ &= \mathbf{P}_\bullet(Y_n|S_n = i) \sum_{j\in S} \alpha_{n-1}^j \mathbf{P}_\bullet(S_n = i|S_{n-1} = j) = b^i(Y_n) \sum_{j\in S} \alpha_{n-1}^j q^{j,i},\end{aligned}$$

for $n = 2, \ldots N$. The initial values are $\alpha_1^i = \mathbf{p}_\bullet^i b^i(Y_1)$, $i \in S$. In the backward step, we start with $\beta_N^i = 1$, $i \in S$, and we compute $\beta_n = (\beta_n^i, i \in S)$, $n = N-1, \ldots 1$:

$$\begin{aligned}\beta_n^i &= \mathbf{P}_\bullet(Y_{n+1}, \ldots, Y_N|S_n = i) \\ &= \sum_{j\in S} \beta_{n+1}^j \mathbf{P}_\bullet(Y_{n+1}|S_{n+1} = j) \mathbf{P}_\bullet(S_{n+1} = j|S_n = i) = \sum_{j\in S} \beta_{n+1}^j b^j(Y_{n+1}) q^{i,j}.\end{aligned}$$

Using $\{\alpha_n, \beta_n, n = 1, \ldots, N\}$, we compute the expectations needed in the M-step:

$$E[S_n = i|Y_1, \ldots, Y_N] = \frac{\alpha_n^i \beta_n^i}{\sum_{j\in S} \alpha_n^j \beta_n^j},$$

$$E[S_{n-1}=j, S_n=i|Y_1,\ldots,Y_N] = \frac{\alpha_{n-1}^j q^{j,i} b^i(Y_n)\beta_n^i}{\sum_{k\in S}\sum_{l\in S}\alpha_{n-1}^k q^{k,l} b^l(Y_n)\beta_n^l}.$$

For numerical stability, especially for large N, it is better to work with some normalized versions of $\{\alpha_n, \beta_n, n = 1,\ldots,N\}$. For example, corresponding to $\{\alpha_n^i, i \in S\}$, we have the filter $\{\mathbf{P}_\bullet(S_n = i|Y_1,\ldots,Y_n), i \in S\}$, $n = 1,\ldots,N$:

$$\mathbf{P}_\bullet(S_n = i|Y_1,\ldots,Y_n) = \frac{b^i(Y_n)p_n^i}{\sum_{j\in S} b^j(Y_n)p_n^j}.$$

Here, $p_n = (p_n^i)$, $i \in S$ is the prediction filter, $p_n^i = \mathbf{P}_\bullet(S_n = i|Y_{n-1},\ldots,Y_1)$. We have $p_1 = \mathbf{p}_\bullet$ and for $n \geq 1$

$$p_{n+1} = \frac{Q^t B_\bullet(Y_n)p_n}{b_\bullet^t(Y_n)p_n},$$

where $B_\bullet(y) = \mathrm{diag}(b^i(y))$, for any $y \in \mathbb{R}^d$ [12]. To emphasize the dependency with respect to the initial condition and the observations, we use the same notation as in [12], and we have

$$p_{n+1} = f[Y_n,\ldots,Y_1,p_1] = \frac{M_{n,1}p_1}{e^t M_{n,1}p_1} = M_{n,1}\cdot p_1, \tag{4.7}$$

where $e = (1,\ldots,1)^t$ is a M-dimensional vector, $\cdot$ denotes the projective product [11] and for any integers $n \geq l$, $M_{n,l} = Q^t B_\bullet(Y_n)\cdots Q^t B_\bullet(Y_l)$.

4.3.2 Geometric ergodicity of the prediction filter

Notice that $\{p_n, n \geq 1\}$ is a Markov chain with the transition probability

$$\Pi(p_{n+1} \in E|p_n = p) = \sum_{j\in S} p^j \int_{\mathbb{R}^d} b^j(y)1_E(f[y,p])\lambda(dy),$$

for any $n \geq 1$, $E \in \mathcal{W}$ and $p \in W$.

For any real-valued, bounded and $(\mathcal{W}, \mathcal{B}(R))$ measurable function g defined on

W, and any $p \in W$, we have

$$Ug(p) = \mathbf{E}_\bullet[g(p_{n+1})|p_n = p] = \sum_{j\in S} p^j \int_{\mathbb{R}^d} b^j(y) g(f[y,p])\lambda(dy), \tag{4.8}$$

$$\begin{aligned} U^n g(p) = \mathbf{E}_\bullet[g(p_{n+1})|p_1 = p] = &\sum_{i_1,\ldots,i_n \in S} p^{i_1} q^{i_1,i_2} \cdots q^{i_{n-1},i_n} \\ &\times \int_{\mathbb{R}^{nd}} b^{i_1}(y_1)\cdots b^{i_n}(y_n) g(f[y_n,\ldots,y_1,p])\lambda^{(n)}(dy^{(n)}), \end{aligned} \tag{4.9}$$

where $(\mathbb{R}^{nd}, \mathcal{B}(\mathbb{R}^{nd}), \lambda^{(n)})$ is the product measurable space, $\lambda^{(n)} = \lambda \otimes \ldots \otimes \lambda$, and $y^{(n)} = (y_1^t, \ldots, y_n^t)^t$, for any $n > 1$.

It is easy to verify that the quadruple $\{(W, \mathcal{W}), (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)), u, P\}$ is a RSCC, where $u : W \times \mathbb{R}^d \to W$,

$$\begin{aligned} u(p,y) &= f[y,p] = \frac{Q^t B_\bullet(y) p}{b_\bullet^t(y) p}, \\ P(p,E) &= \sum_{j\in S} p^j \int_E b^j(y)\lambda(dy), \end{aligned} \tag{4.10}$$

for all $p = (p^1, \ldots, p^M)^t \in W$, $y \in \mathbb{R}^d$, and $E \in \mathcal{B}(\mathbb{R}^d)$. For any $n \geq 1$, $E \in \mathcal{B}(\mathbb{R}^{nd})$ and $p \in W$, we define recursively the maps $u^{(n)} : W \times (\mathbb{R}^d)^n \to W$ and the transition probability functions P_n:

$$u^{(n)}(p,(y_1,\ldots,y_n)) = py^{(n)} = f[y_n,\ldots,y_1,p], \quad y_i \in \mathbb{R}^d, p \in W, \tag{4.11}$$

$$P_n(p,E) = \sum_{i_1,\ldots,i_n \in S} p^{i_1} q^{i_1,i_2}\cdots q^{i_{n-1},i_n} \int_E b^{i_1}(y_1)\cdots b^{i_n}(y_n)\lambda^{(n)}(dy^{(n)}). \tag{4.12}$$

Comparing (4.10) with (4.8) and (4.12) with (4.9), we notice that

$$\begin{aligned} Ug(p) &= \int_{\mathbb{R}^d} g(f[y,p])P(p,dy), \\ U^n g(p) &= \int_{\mathbb{R}^{nd}} g(f[y_n,\ldots,y_1,p])P_n(p,dy^{(n)}). \end{aligned}$$

Furthermore, using (4.10)-(4.12), it is easy to show that for any probability distribution $\mathbf{p}_\bullet \in W$, the random sequences $\{p_{n+1}, n \geq 0\}$ with $p_1 = \mathbf{p}_\bullet$ and $\{Y_n, n \geq 1\}$ are associated with the previously defined RSCC on the probability space $(\Omega, \mathcal{F}, \mathbf{P}_\bullet(\cdot|p_1$

$= \mathbf{p}_\bullet$) (see Theorem 4.1).

Throughout this subsection, we suppose the matrix Q is primitive with index of primitivity r. As a consequence, the Markov chain $\{S_n, n \geq 1\}$ is geometrically ergodic with a unique invariant probability distribution $\pi_\bullet = (\pi^i)$ on S. Thus, from Theorem 4.3 in [7] there exists positive constants a_1, $\hat{a}_1$ and $c_1 < 1$, $\hat{c}_1 < 1$ such that $Q^n = (q_n^{i,j})$, the n-th power of the matrix Q, satisfies

$$|q_n^{i,j} - \pi^j| \leq a_1 c_1^n, \tag{4.13}$$

$$\max_{i,i' \in S} \sum_{j \in S} |q_n^{i,j} - q_n^{i',j}| \leq \hat{a}_1 \hat{c}_1^n, \tag{4.14}$$

for any $n \geq 1$ and any $i, j \in S$. As in [12], let $\min^+$ denote the minimum over positive elements, and let

$$\delta(y) = \frac{\max_{i \in S} b^i(y)}{\min_{i \in S} b^i(y)}, \quad \Delta_{-1} = \min_{i \in S} \int_{\mathbb{R}^d} \delta^{-1}(y) b^i(y) \lambda(dy),$$

$$\epsilon = \min_{i,j \in S}{}^+ q^{i,j}, \quad \Delta = \max_{i \in S} \int_{\mathbb{R}^d} \delta(y) b^i(y) \lambda(dy), \quad R = \epsilon^r \Delta_{-1}^{r-1}.$$

Now we establish two important properties of the previously defined RSCC.

Proposition 4.2 *The RSCC* $\left\{(W, \mathcal{W}), \left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right), u, P\right\}$ *is uniformly ergodic.*

PROOF. For any integers $n, k \geq 1$, every $p \in W$ and $E \in \mathcal{B}\left(\mathbb{R}^{kd}\right)$, from (4.12), we get

$$P_k^n(p, E) = \sum_{i_1, \ldots, i_{k+1} \in S} p^{i_1} q_{n-1}^{i_1, i_2} q^{i_2, i_3} \cdots q^{i_k, i_{k+1}} \int_E b^{i_2}(y_1) \cdots b^{i_{k+1}}(y_k) \lambda^{(k)}(dy^{(k)}).$$

Let

$$P_k^\infty(E) = \sum_{i_1, \ldots, i_k \in S} \pi^{i_1} q^{i_1, i_2} \cdots q^{i_{k-1}, i_k} \int_E b^{i_1}(y_1) \cdots b^{i_k}(y_k) \lambda^{(k)}(dy^{(k)}).$$

Using (4.13), it is easy to prove that $\lim_{n \to \infty} P_k^n(p, E) = P_k^\infty(E)$, uniformly with respect to p, E and k. Thus the RSCC is uniformly ergodic. □

Proposition 4.3 *If* $\Delta < \infty$, *then* $\left\{(W, \mathcal{W}), \left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right), u, P\right\}$ *is an RSCC with*

contraction.

PROOF. Replacing in (4.10) and using the definition of the $\|\cdot\|_1$, we obtain

$$R_1 = \sup_{E\in\mathcal{B}(\mathbb{R}^d)} \sup_{p\neq p',p,p'\in W} \frac{|P(p,E)-P(p',E)|}{\|p-p'\|_1} \leq 1 < \infty. \tag{4.15}$$

From Lemma 2.2 in [11], we get

$$\|f[y,p]-f[y,p']\|_1 \leq \delta(y)\|p-p'\|_1 \quad p,p' \in W, y \in \mathbb{R}^d.$$

Hence,

$$r_1 = \sup_{p\neq p',p,p'\in W} \int_{\mathbb{R}^d} \frac{\|f[y,p]-f[y,p']\|_1}{\|p-p'\|_1} P(p,dy) \leq \Delta < \infty. \tag{4.16}$$

For $n > 2r$, from the first inequality in Theorem 2.1 and the inequality 5 in [12], we get

$$\begin{aligned} r_n &= \sup_{p\neq p',p,p'\in W} \int_{\mathbb{R}^{dn}} \frac{\|f[y_n,\ldots,y_1,p]-f[y_n,\ldots,y_1,p']\|_1}{\|p-p'\|_1} P_n(p,dy^{(n)}) \\ &\leq \epsilon^{-r}\Delta^r(1-R)^{n/r-2}. \end{aligned}$$

Thus, for n sufficiently large, $r_n < 1$, and together with (4.15) and (4.16), this implies that we have an RSCC with contraction. □

Remark 4.4 If the observation conditional densities b^i are Gaussian for any $i \in S$ (as in example 4.1), then $\Delta < \infty$ (Example 4.2 in [12]).

Now, we return to the Markov chain $\{p_n\}$ associated with the RSCC. Using the previous proposition, and the fact that $(W, \|\cdot\|_1)$ is a compact space, we obtain that $\{p_n\}$ is a compact Markov chain (see also Remark 4.2). Furthermore, let define

$$U_n = \frac{1}{n}\sum_{k=1}^{n} U^k, n \geq 1.$$

Theorem 4.2 *The Markov chain $\{p_n, n \geq 1\}$ is geometrically ergodic. If $\Delta < \infty$ and Q^∞ is the unique invariant probability distribution for the chain $\{p_n, n \geq 1\}$,*

then we also have $\lim_{n\to\infty} \|U_n - U^\infty\|_{BL} = 0$. *In addition, there exists positive constants* C *and* $\theta < 1$, *such that for any function* $g \in L(W)$

$$\|U^n g - U^\infty g\|_{BL} \leq C\theta^n \|g\|_{BL}.$$

Here, for any bounded and measurable real-valued function g, $U^\infty g$ *is defined by*

$$U^\infty g = \int_W g(p) Q^\infty(dp).$$

Lemma 4.1 *Let* $c = \max\{c_1, c_2\}$, *with* $c_2 = (1-R)^{1/r}$ *and* c_1 *as defined in (4.13). There exists a positive constant* K *such that for any positive integer* $n > r+1$, $g \in L(W)$ *and* p_1, $p_2 \in W$

$$|U^n g(p_1) - U^n g(p_2)| \leq nKc^n \|g\|_{BL}.$$

PROOF OF LEMMA 4.1. We follow the same ideas as in the proof of Theorem 3.5 in [12]. Using the second inequality in Theorem 2.1 and the inequality (5) in [12], we obtain a result similar to Proposition 3.7 in [12]:

$$\begin{aligned} \max_{i_l,\ldots,i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} & |g(f[y_n,\ldots,y_l,p_1]) - g(f[y_n,\ldots,y_l,p_2])| \\ & \times b^{i_l}(y_l)\ldots b^{i_n}(y_n)\lambda(dy_l)\ldots\lambda(dy_n) \leq 2s(g)c_2^{n-l+1-r}, \end{aligned} \tag{4.17}$$

for any positive integers n, l such that $n \geq l+r-1$, and any function $g \in L(W)$. Now we express

$$\begin{aligned} U^n g(p_1) - U^n g(p_2) = & \sum_{i_1,\ldots,i_n \in S} p_1^{i_1} q^{i_1,i_2}\cdots q^{i_{n-1},i_n} \int_{\mathbb{R}^{nd}} b^{i_1}(y_1)\cdots b^{i_n}(y_n) \\ & \times \left(g(f[y_n,\ldots,y_1,p_1]) - g(f[y_n,\ldots,y_1,p_2])\right) \lambda^{(n)}\left(dy^{(n)}\right) + \sum_{i_1,\ldots,i_n \in S} (p_1^{i_1} - p_2^{i_1}) \\ & \times q^{i_1,i_2}\cdots q^{i_{n-1},i_n} \int_{\mathbb{R}^{nd}} b^{i_1}(y_1)\ldots b^{i_n}(y_n) g(f[y_n,\ldots,y_1,p_2])\lambda^{(n)}\left(dy^{(n)}\right). \end{aligned}$$

Let T_1 and T_2 denote the first and the second term, respectively. Using (4.17), we

get:

$$|T_1| \leq 2s(g)c_2^{n-r}. \tag{4.18}$$

For any positive integer $l \leq n-1$ and any sequence $z_1, \ldots, z_l \in \mathbb{R}^d$ decomposing $g(f[y_n, \ldots, y_1, p_2])$, we obtain:

$$|T_2| \leq \sum_{k=2}^{l} \sum_{i_k,\ldots,i_n \in S} \left| \sum_{i_1,\ldots,i_{k-1} \in S} (p_1^{i_1} - p_2^{i_1}) q^{i_1,i_2} \cdots q^{i_{k-1},i_k} \right| q^{i_k,i_{k+1}} \cdots q^{i_{n-1},i_n}$$
$$\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} |g(f[y_n, \ldots, y_k, z_{k-1}, \ldots, z_1, p_2]) - g(f[y_n, \ldots, y_{k+1}, z_k, \ldots, z_1, p_2])|$$
$$b^{i_k}(y_k) \cdots b^{i_n}(y_n) \lambda(dy_k) \ldots \lambda(dy_n) + \sum_{i_1,\ldots,i_n \in S} |p_1^{i_1} - p_2^{i_1}| q^{i_1,i_2} \cdots q^{i_{n-1},i_n}$$
$$\int_{\mathbb{R}^{nd}} b^{i_1}(y_1) \cdots b^{i_n}(y_n) \, |g(f[y_n, \ldots, y_1, p_2]) - g(f[y_n, \ldots, y_2, z_1, p_2])| \, \lambda^{(n)}\left(dy^{(n)}\right)$$
$$+ \sum_{i_{l+1},\ldots,i_n \in S} \left| \sum_{i_1,\ldots,i_l \in S} (p_1^{i_1} - p_2^{i_1}) q^{i_1,i_2} \cdots q^{i_l,i_{l+1}} \right| q^{i_{l+1},i_{l+2}} \cdots q^{i_{n-1},i_n}$$
$$\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} |g(f[y_n, \ldots, y_{l+1}, z_l, \ldots, z_1, p_2])| \, b^{i_{l+1}}(y_{l+1}) \cdots b^{i_n}(y_n) \lambda(dy_{l+1}) \ldots \lambda(dy_n).$$

The inequality (4.17) and $p_1, p_2 \in W$ yields

$$|T_2| \leq 2s(g) \sum_{k=2}^{l} c_2^{n-k-r} \sum_{i_k \in S} \left| \sum_{i_1 \in S} (p_1^{i_1} - p_2^{i_1})(q_{k-1}^{i_1,i_k} - \pi^{i_k}) \right| + 2s(g)c_2^{n-1-r} \|p_1 - p_2\|_1$$
$$+ \|g\| \sum_{i_{l+1} \in S} \left| \sum_{i_1 \in S} (p_1^{i_1} - p_2^{i_1})(q_l^{i_1,i_{l+1}} - \pi^{i_{l+1}}) \right|$$

Using (4.13) and $l = n - r$, we get

$$|T_2| \leq \|p_1 - p_2\|_1 \left(2s(g) M a_1 (n-r-1) c^{n-r-1} + 2s(g) c_2^{n-1-r} + \|g\| M a_1 c_1^{n-r} \right). \tag{4.19}$$

Since $\|p_1 - p_2\|_1 \leq 2$, adding (4.18) and (4.19), we get the conclusion. □

PROOF OF THEOREM 4.2. Using Lemma 4.1 and proceedings as in the proof of Corollary 3.6 in [12], it can be shown that there exists a unique invariant probability distribution Q^∞ for the Markov chain $\{p_n, n \geq 1\}$. Moreover, for any positive integer

$n \geq 1$, all $g \in L(W)$, and any $z \in W$, we have

$$|U^n g(z) - U^\infty g| \leq K\|g\|_{BL} n c^n / (1-c)^{-2}.$$

If $\Delta < \infty$, the properties of the associated RSCC allow us to state stronger results concerning the convergence of U^n. As we have already mentioned, the Markov chain $\{p_n, n \geq 1\}$ is compact and by Theorem 3.2.2, in [8], the Ionescu Tulcea-Marinescu ergodic Theorem applies ([6]).

Furthermore, Lemma 4.1 implies that any eigenfunction $g \in L(W)$ of U, corresponding to an eigenvalue γ with $|\gamma| = 1$, is a constant function. Hence, $\gamma = 1$ is the only eigenvalue of modulus 1 of U and the subspace $E(1) = \{g \in L(W) : Ug = g\}$ is one dimensional. Thus, we obtain the stated conclusions from the Ionescu Tulcea-Marinescu ergodic Theorem and Theorem 3.2.4, in [8]. □

4.4 The hybrid models

Now consider a model formed by the unobserved random sequence $\{S_n, n \geq 1\}$, and the observed sequences $\{Y_n, n \geq 1\}$ and $\{X_n, n \geq 0\}$, defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P}_\bullet)$ with values in $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively. Let λ_d and λ_m be two non-negative and σ- finite measures on $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right)$ and $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$, respectively. We suppose that the conditional probability distribution of the observation Y_n given that $\{S_n = i, X_n = x\}$ is absolutely continuous with respect to λ_d and it has a positive and continuous density

$$\mathbf{P}_\bullet(Y_n \in A_y | S_n = i, X_n = x) = \int_{A_y} b^i(y,x)\lambda_d(dy),$$

for any integer $n \geq 1$, $i \in S$, $A_y \in \mathcal{B}\left(\mathbb{R}^d\right)$, and $x \in \mathbb{R}^m$. Moreover, the observations $\{Y_n, n \geq 1\}$ are mutually independent given the sequences $\{S_n, n \geq 1\}$ and $\{X_n, n \geq 0\}$:

$$\mathbf{P}_\bullet(Y_n \in A_n, \ldots, Y_1 \in A_1 | S_n = i_n, X_n = x_n \ldots, S_1 = i_1, X_1 = x_1,$$

$$X_0 = x_0) = \prod_{k=1}^{n} \mathbf{P}_\bullet(Y_k \in A_k | S_k = i_k, X_k = x_k),$$

for any integer $n \geq 1$, any $i_1, \ldots i_n \in S$, all $x_0, \ldots x_n \in \mathbb{R}^m$, and any sets $A_n, \ldots A_1 \in \mathcal{B}\left(\mathbb{R}^d\right)$. We also suppose that $\{X_n, n \geq 0\}$ is a Markov chain, and there exists a positive, and continuous function $a(\cdot, \cdot)$ such that for all $x \in \mathbb{R}^m$, and $A_x \in \mathcal{B}\left(\mathbb{R}^m\right)$

$$\mathbf{P}_\bullet\left(X_{n+1} \in A_x | X_n = x\right) = \int_{A_x} a(y, x)\lambda_m(dy), \quad n \geq 0,$$
$$\mathbf{P}_\bullet\left(X_{n+1} \in A_x | X_n = x, S_{n+1}, Y_n, S_n, \ldots, Y_1, S_1\right) = \mathbf{P}_\bullet\left(X_{n+1} \in A_x | X_n = x\right),$$

for any $n > 0$. Let $P_X^n(x, A_x) = \mathbf{P}_\bullet(X_n \in A_x | X_0 = x)$.

Remark 4.5 The Bayesian network corresponding to this model is displayed in Fig. 1.3.

We define the prediction filter $w_n = (w_n^i)$, $n \geq 1$, where $w_1 = \mathbf{p}_\bullet$, and for any $n \geq 2$, $w_n^i = \mathbf{P}_\bullet(S_n = i | Y_{n-1}, X_{n-1}, \ldots, Y_1, X_1)$. For any $y \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$, let $b_\bullet(y, x) = (b^i(y, x))$ and $B_\bullet(y, x) = \mathrm{diag}(b^i(y, x))$. It is easy to prove that

$$w_{n+1} = \frac{Q^t B_\bullet(Y_n, X_n) w_n}{b_\bullet^t(Y_n, X_n) w_n}, \quad n \geq 1. \tag{4.20}$$

We now write an equation similar to (4.7):

$$w_{n+1} = F[Y_n, X_n \ldots, Y_1, X_1, w_1] = \frac{K_{n,1} w_1}{e^t K_{n,1} w_1} = K_{n,1} \cdot w_1, \tag{4.21}$$

where for any integers $i \geq j$, $K_{i,j} = Q^t B_\bullet(Y_i, X_i) \cdots Q^t B_\bullet(Y_j, X_j)$.

In practice, the matrix Q, the initial probability $\mathbf{p}_\bullet$ and the vector $b_\bullet$ might be unknown, and only some estimations $\hat{Q}$, $\hat{\mathbf{p}}_\bullet$, $\hat{b}_\bullet$ might be available. Thus, corresponding to (4.20) we have

$$\hat{w}_{n+1} = \frac{\hat{Q}^t \hat{B}_\bullet(Y_n, X_n) \hat{w}_n}{\hat{b}_\bullet^t(Y_n, X_n) \hat{w}_n}, \quad n \geq 1, \tag{4.22}$$

and $\hat{w}_1 = \hat{\mathbf{p}}_\bullet$. Similarly to (4.21) we can write $\hat{w}_{n+1} = \hat{F}[Y_n, X_n \ldots, Y_1, X_1, \hat{w}_1]$.

Remark 4.6 The model (4.1) - (4.2) in example 4.2 is a mixture model, and the number of components increases exponentially with n. The Gibbs sampler can be used for inference. Hence, we draw samples $\{X_n(l), n \geq 0\}$, $l = 1, \ldots, L$ and then, for each sample, we calculate the prediction filter $w_n(l) = (w_n^i(l))$, $n \geq 1$. At any step l, the model formed with $\{S_n, Y_n, X_n(l)\}$ satisfies all the previous assumptions. For any $i \in S$, we have

$$b^i(y,x) = (2\pi)^{-d/2}[\det R(i)]^{-1/2} \exp[-(y - C(i)x)^t R(i)^{-1}(y - C(i)x)/2],$$
$$a(y,x) = (2\pi)^{-m/2}[\det H]^{-1/2} \exp[-(y - Ax)^t H^{-1}(y - Ax)/2],$$

where λ_d and λ_m are the Lebesgue measures on $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively. Here,

$$w_1 = \mathbf{p}_\bullet, \quad w_n^i(l) = \mathbf{P}_\bullet(S_n = i | Y_{n-1}, X_{n-1}(l), \ldots, Y_1, X_1(l)), \quad n \geq 2.$$

Let us denote $V = \mathbb{R}^m \times W$, and $\hat{V} = S \times \mathbb{R}^m \times \mathbb{R}^d \times W$, where W is given in (4.6). Then (V, d) and $(\hat{V}, \hat{d})$ are metric spaces, where

$$d((x_1, p_1), (x_2, p_2)) = \|x_1 - x_2\| + \|p_1 - p_2\|_1,$$
$$\hat{d}((i, x_1, y_1, p_1), (j, x_2, y_2, p_2)) = \|x_1 - x_2\| + \|y_1 - y_2\| + |i - j| + \|p_1 - p_2\|_1.$$

It is easy to show that the topology induced by the metric d on V is the the topology induced by the Euclidian space $\mathbb{R}^{m+M}$, and the topology induced by the metric $\hat{d}$ on $\hat{V}$ is the topology induced by the Euclidian space $\mathbb{R}^{d+m+M+1}$. Let $\mathcal{V}$ and $\hat{\mathcal{V}}$ be the collection of the Borel sets on V and $\hat{V}$, respectively. For any positive integer n, let $\lambda_{n(m+d)} = (\lambda_d \otimes \lambda_m) \otimes \ldots \otimes (\lambda_d \otimes \lambda_m)$ be the product measure on $\mathbb{R}^{n(m+d)}$. We use the notation $(y, x)^{(n)}$ for the sequence $(y_1^t, x_1^t, \ldots, y_n^t, x_n^t)^t \in \mathbb{R}^{n(m+d)}$. We denote by $C^\infty(V)$ the real-valued , bounded and continuous functions on V, and by $L(V)$ the set of real-valued, bounded and Lipschitz continuous functions defined on V. Then, $L(V)$ is a Banach space for the norm $\|\cdot\|_{BL}$ defined by $\|g\|_{BL} = \|g\| + s(g)$, where

$$\|g\| = \sup_{(x,p) \in V} |g(x,p)|, \quad s(g) = \sup_{(x,p) \neq (x',p')} \frac{|g(x,p) - g(x',p')|}{\|p - p'\|_1 + \|x - x'\|}.$$

Similarly, let $C^{\infty}(\hat{V})$ denote the real-valued, bounded and continuous functions on $\hat{V}$, and $L(\hat{V})$ the set of real-valued, bounded and Lipschitz continuous functions defined on $\hat{V}$. Then $L(\hat{V})$ is a Banach space for the norm $\|\cdot\|_{BL}$ with

$$\|g\| = \max_{i\in S} \sup_{(x,y,p)\in\mathbb{R}^{d+m}\times W} |g(i,x,y,p)|,$$
$$s(g) = \max_{i\in S} \sup_{(x,y,p)\neq(x',y',p')} \frac{|g(i,x,y,p) - g(i,x',y',p')|}{\|p-p'\|_1 + \|x-x'\| + \|y-y'\|}.$$

4.4.1 Ergodicity of the Markov chain $\{w_{n+1}, X_n, n > 0\}$

In this case, $\{w_n, n \geq 1\}$ is not a Markov chain, but $\{w_{n+1}, X_n, n \geq 0\}$ is a Markov chain with the transition probability

$$\Pi(w_{n+2} \in Y, X_{n+1} \in X | w_{n+1} = p, X_n = x') = \sum_{j\in S} p^j \int_X \int_{\mathbb{R}^d} b^j(y,x) \times a(x,x') 1_Y(F[y,x,p]) \lambda_d(dy) \lambda_m(dx), \tag{4.23}$$

for any $n \geq 0$, $Y \in \mathcal{W}$, $X \in \mathcal{B}(\mathbb{R}^m)$ and $(x',p) \in V$. We have

$$\begin{aligned} Ug(x_0,p) &= \mathbf{E}_\bullet [g(X_{n+1}, w_{n+2}) | X_n = x_0, w_{n+1} = p] \\ &= \sum_{j\in S} p^j \int_{\mathbb{R}^{d+m}} b^j(y,x) a(x,x_0) g\left(x, F[y,x,p]\right) \lambda_{d+m}\left(d(y,x)\right), \end{aligned} \tag{4.24}$$

$$\begin{aligned} U^n g(x_0,p) &= \mathbf{E}_\bullet [g(X_n, w_{n+1}) | X_0 = x_0, w_1 = p] \\ &= \sum_{i_1,\ldots,i_n \in S} p^{i_1} q^{i_1,i_2} \ldots q^{i_{n-1},i_n} \int_{\mathbb{R}^{n(d+m)}} b^{i_1}(y_1,x_1) a(x_1,x_0) \cdots b^{i_n}(y_n,x_n) \\ &\times a(x_n,x_{n-1}) g\left(x_n, F[y_n, x_n \ldots, y_1, x_1, p]\right) \lambda_{n(d+m)}\left(d(y,x)^{(n)}\right), \end{aligned} \tag{4.25}$$

for any real-valued, bounded and $(\mathcal{V}, \mathcal{B}(R))$ measurable function g defined on V, any $(x_0,p) \in V$, and any $n > 1$.

For any $E \in \mathcal{B}(\mathbb{R}^{m+d})$, $(x,p) \in V$, and $(y,x') \in \mathbb{R}^{m+d}$, we define the mapping u and the transition probability function P;

$$u((x,p),(y,x')) = (x', F[y,x',p]) = \left(x', \frac{Q^t B_\bullet(y,x')p}{b^t_\bullet(y,x')p}\right), \tag{4.26}$$

$$P((x,p),E)) = \sum_{j=1}^{M} p^j \int_E b^j(y,x^{'})a(x^{'},x)\lambda_{m+d}(d(x^{'},y)). \tag{4.27}$$

It is easy to verify that the quadruple $\left\{(V,\mathcal{V}),(\mathbb{R}^{m+d},\mathcal{B}(\mathbb{R}^{m+d})),u,P\right\}$ is a RSCC. For the maps $u^{(n)}: V\times\mathbb{R}^{n(m+d)}\to V$ we have the following formula

$$u^{(n)}\left((x_0,p),(y_1,x_1,\ldots,y_n,x_n)\right) = (x_n, F[y_n,x_n,\ldots,y_1,x_1,p])\,. \tag{4.28}$$

Throughout this section $(x_0,p)(y,x)^{(n)}$ denotes $u^{(n)}\left((x_0,p),(y,x)^{(n)}\right)$. For the model considered here, the transition probability functions P_n are

$$\begin{aligned} P_n((x_0,p),E) = &\sum_{i_1,\ldots,i_n\in S} p^{i_1}q^{i_1,i_2}\ldots q^{i_{n-1},i_n}\int_E b^{i_n}(y_n,x_n)a(x_n,x_{n-1}) \\ &\cdots b^{i_1}(y_1,x_1)a(x_1,x_0)\lambda_{n(d+m)}(d(y,x)^{(n)}), \end{aligned} \tag{4.29}$$

for any $n\geq 1$, $E\in\mathcal{B}\left(\mathbb{R}^{n(d+m)}\right)$ and $(x_0,p)\in V$. Hence, for any integers $n,k\geq 1$,

$$\begin{aligned} P_k^n((x_0,p),E) = P_{n+k-1}\left((x_0,p),\mathbb{R}^{(n-1)(d+m)}\times E\right) = &\sum_{i_1,\ldots,i_{k+1}\in S} p^{i_1}q_{n-1}^{i_1,i_2}q^{i_2,i_3}\cdots \\ \times q^{i_k,i_{k+1}}\int_{\mathbb{R}^{(n-1)m}}\Bigg(\int_E b^{i_{k+1}}(y_k,x_{k+n-1})a(x_{k+n-1},x_{k+n-2})\cdots b^{i_2}(y_1,x_n)a(x_n,x_{n-1}) \\ \cdots a(x_1,x_0)\lambda_{kd}\left(dy^{(k)}\right)\lambda_{km}\left(d(x_{k+n-1},\ldots,x_n)\right)\Bigg)\lambda_{(n-1)m}\left(dx^{(n-1)}\right), \end{aligned} \tag{4.30}$$

where $(x_0,p)\in V$, and $E\in\mathcal{B}\left(\mathbb{R}^{k(d+m)}\right)$. Comparing (4.27) with (4.24) and (4.29) with (4.25), we notice that

$$\begin{aligned} Ug(x_0,p) &= \int_{\mathbb{R}^{d+m}} g\left(x,F[y,x,p]\right)P\left((x_0,p),d(y,x)\right), \\ U^n g(x_0,p) &= \int_{\mathbb{R}^{n(d+m)}} g\left(x_n,F[y_n,x_n\ldots,y_1,x_1,p]\right)P_n\left((x_0,p),d(y,x)^{(n)}\right). \end{aligned}$$

Moreover, using (4.26)-(4.30), it is easy to verify that for any $(x_0,\mathbf{p}_\bullet)\in V$, the random sequences $\{(X_n,w_{n+1}),n\geq 0\}$, with $(X_0,w_1)=(x_0,\mathbf{p}_\bullet)$, and $\{Y_n,n\geq 1\}$ are associated with the previously defined RSCC, on the probability space $(\Omega,\mathcal{F},\mathbf{P}_\bullet(\cdot|w_1=\mathbf{p}_\bullet,X_0=x_0)$ (see Theorem 4.1).

Now we make the following assumptions.

ASSUMPTION A The Matrix Q is positive.

ASSUMPTION B The Markov chain $\{X_n, n > 0\}$ is geometrically ergodic such that there exists a probability distribution π_X, a positive real number $\rho_x < 1$, and a real valued, non-negative, $(\mathcal{B}(\mathbb{R}^m), \mathcal{B}(\mathbb{R}))$ measurable, π_X-integrable function G with properties:

1. $\|P_X^n(x_0, \cdot) - \pi_X(\cdot)\| \leq G(x_0)\rho_X^n$, for all $x_0 \in \mathbb{R}^m$, and $n \geq 1$;

2. For all $x^{'} \in \mathbb{R}^m$:

$$\mu(x^{'}) = \int_{\mathbb{R}^m} G(x)a(x, x^{'})\lambda_m(dx) < \infty;$$

3. The function $G_1(x) = G(x)+\mu(x)$ is $\Pi(\cdot, |x^{'}, p')$ integrable, for any $(x^{'}, p') \in V$.

Since a positive matrix is a primitive matrix with index of primitivity $r = 1$, the Markov chain $\{S_n, n \geq 1\}$ is geometrically ergodic and the inequality (4.13) is true. Using Theorem 3.5 in [11], it is easy to show that

$$\left\| F[y_n, x_n, \ldots, y_1, x_1, p] - F[y_n, x_n, \ldots, y_1, x_1, p^{'}] \right\|_1 \leq 2(1-\epsilon)^n, \tag{4.31}$$

for any positive integer $n \geq 1$ and any sequence $(x, y)^{(n)} \in \mathbb{R}^{n(d+m)}$.

Proposition 4.4 *There exists a positive constant C such that for any function $g \in L(V)$, any positive integers n, h, and $k \geq n$*

$$\left| U^h g\left((x_0, p)(y, x)^{(k)}\right) - U^h g\left((x_0^{'}, p^{'})(y, x)^{(k)}\right) \right| \leq C(1-\epsilon)^n \|g\|_{BL},$$

for all (x_0, p), $(x_0^{'}, p^{'}) \in V$ and all sequences $(y, x)^{(k)} \in \mathbb{R}^{k(d+m)}$.

PROOF. For any positive integer n, let

$$a_n = \sup \left| P\left((x_0, p)(y, x)^{(k)}, X\right) - P\left((x_0^{'}, p^{'})(y, x)^{(k)}, X\right) \right|,$$

where the supremum is taken over all $k \geq n$, (x_0, p), $(x_0', p') \in V$, $(y, x)^{(k)} \in \mathbb{R}^{k(d+m)}$, and $X \in \mathcal{B}(\mathbb{R}^{d+m})$. Then, replacing in (4.27) and using (4.31), we get

$$a_n \leq 2(1-\epsilon)^n, \tag{4.33}$$

and then

$$\sum_{n\geq 1} a_n \leq 2\frac{1-\epsilon}{\epsilon} < \infty.$$

Using (4.31), we can easily show that

$$\left|g((x_0,p)(y,x)^{(k)}) - g((x_0',p')(y,x)^{(k)})\right| \leq 2s(g)(1-\epsilon)^n.$$

Applying Proposition 4.1 with $A_0 = \mathbb{R}^{m+d}$ and $\nu = 1$ we get

$$\left|U^h g\left((x_0,p)(y,x)^{(k)}\right) - U^h g\left((x_0',p')(y,x)^{(k)}\right)\right| \leq 4\|g\| \sum_{j=n}^{\infty} a_j + 2s(g)(1-\epsilon)^n.$$

Hence, the conclusion of the proposition is just a direct consequence of the previous inequality and (4.33). □

Proposition 4.5 *If $g \in C^\infty(V)$, then $U(g) \in C^\infty(V)$.*

PROOF. Obviously for any $g \in C^\infty(V)$

$$\|Ug\| \leq \|g\| < \infty.$$

Now we prove that Ug is continuous. Let us arbitrary fix $(x_0, p) \in V$ and consider any sequence $\{(x_n, p_n), n \geq 1\} \subset V$, such that

$$\lim_{n\to\infty} \|p_n - p\|_1 = 0, \quad \lim_{n\to\infty} \|x_n - x_0\| = 0.$$

Replacing in (4.24), we get

$$Ug(x_n,p_n) - Ug(x_0,p) = \sum_{j\in S} p_n^j \int_{\mathbb{R}^{d+m}} b^j(y,x)a(x,x_n)g(x,F[y,x,p_n])\lambda_{d+m}(d(y,x))$$

$$-\sum_{j\in S} p^j \int_{\mathbb{R}^{d+m}} b^j(y,x)a(x,x_0)g(x,F[y,x,p])\lambda_{d+m}(d(y,x)).$$

Hence,

$$\begin{aligned}
&|Ug(x_n,p_n) - Ug(x_0,p)| \leq \sum_{j\in S} |p_n^j - p^j| \int_{\mathbb{R}^{d+m}} b^j(y,x)a(x,x_n)\,|g(x,F[y,x,p_n])| \\
&\lambda_{d+m}\left(d(y,x)\right) + \sum_{j\in S} p^j \left| \int_{\mathbb{R}^{d+m}} b^j(y,x)(a(x,x_n) - a(x,x_0))g(x,F[y,x,p_n]) \right. \\
&\left. \lambda_{d+m}\left(d(y,x)\right) \right| + \sum_{j\in S} p^j \left| \int_{\mathbb{R}^{d+m}} b^j(y,x)a(x,x_0)g(x,F[y,x,p_n])\lambda_{d+m}\left(d(y,x)\right) \right. \\
&\left. - \int_{\mathbb{R}^{d+m}} b^j(y,x)a(x,x_0)g\left(x,F[y,x,p]\right)\lambda_{d+m}\left(d(y,x)\right) \right|
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\sum_{j\in S} |p_n^j - p^j| \int_{\mathbb{R}^{d+m}} b^j(y,x)a(x,x_n)\,|g(x,F[y,x,p_n])|\,\lambda_{d+m}\left(d(y,x)\right) \\
&\leq \|g\|\|p_n - p\|_1 \underset{n\to\infty}{\to} 0.
\end{aligned}$$

Let $\delta_n(x) = a(x,x_0) - a(x,x_n)$ and $\delta_n^+(x) = \max\{a(x,x_0) - a(x,x_n), 0\}$, for any $x \in \mathbb{R}^m$. Notice that $\lim_{n\to\infty} \delta_n^+(x) = 0$, and $\delta_n^+(x) \leq a(x,x_0)$, for any $x \in \mathbb{R}^m$. Hence, by the Lebesgue's theorem of dominated convergence [1]

$$\|P_X(x_n,\cdot) - P_X(x_0,\cdot)\| = \int_{\mathbb{R}^m} |\delta_n(x)|\,\lambda_m\left(dx\right) = 2\int_{\mathbb{R}^m} \delta_n^+(x)\lambda_m\left(dx\right) \underset{n\to\infty}{\to} 0.$$

Thus, using Lemma 2.1.1 in [8]), for the second term, we get

$$\begin{aligned}
&\sum_{j\in S} p^j \left| \int_{\mathbb{R}^{d+m}} b^j(y,x)\delta_n(x)g(x,F[y,x,p_n])\lambda_{d+m}\left(d(y,x)\right) \right| \\
&\leq \|g\|\|P_X(x_n,\cdot) - P_X(x_0,\cdot)\| \underset{n\to\infty}{\to} 0.
\end{aligned}$$

Since $g \in C^\infty(V)$, the Lebesgue's theorem of dominated convergence implies that the third term also goes to 0 when $n \to \infty$. Thus $\lim_{n\to\infty} |Ug(x_n,p_n) - Ug(x_0,p)| = 0$, and, since $(x_0,p) \in V$ was arbitrarily chosen, Ug is continuous. □

Proposition 4.6 *For any positive integers k, n and any (x_0, p_1), $(x_0', p_2) \in V$ we have*

$$\|\Lambda_k^n((x_0,p_1),(x_0',p_2))\| \leq 4a_1 c_1^{n-1} + 2\rho_X^{n-1}\left(G(x_0)+G(x_0')\right),$$

where c_1, a_1 are defined in (4.13), ρ_X is defined in assumption B, and for any $X \in \mathcal{B}(\mathbb{R}^{k(m+d)})$

$$\Lambda_k^n\left((x_0,p_1),(x_0',p_2)\right)(X) = P_k^n\left((x_0,p_1),X\right) - P_k^n\left((x_0',p_2),X\right).$$

PROOF. For any positive integers k, n and any (x_0,p_1), $(x_0',p_2) \in V$, $\Lambda_k^n((x_0,p_1),(x_0',p_2))$ is a signed measure on $\left(\mathbb{R}^{k(m+d)}, \mathcal{B}(\mathbb{R}^{k(m+d)})\right)$. Let $\mathbb{R}^{k(m+d)} = X^+ \cup X^-$ be a Hahn decomposition of $\mathbb{R}^{k(m+d)}$ with respect to $\Lambda_k^n((x_0,p_1),(x_0',p_2))$.

For any $x \in \mathbb{R}^m$ and any sequence $i_1,\ldots,i_k \in S$, let

$$g^{(i_1,\ldots,i_k)}(x) = \int_{X^+} b^{i_k}(y_k,x_k)a(x_k,x_{k-1})\cdots b^{i_1}(y_1,x_1)a(x_1,x)\lambda_{k(d+m)}\left(d(y,x)^{(k)}\right).$$

The Fubini's theorem [1] implies that g is a measurable function. Moreover, we have

$$0 \leq g^{(i_1,\ldots,i_k)}(x) \leq 1, \quad x \in \mathbb{R}^m. \tag{4.34}$$

Replacing in (4.30), we get

$$\|\Lambda_k^n((x_0,p_1),(x_0',p_2))\| = 2\Lambda_k^n((x_0,p_1),(x_0',p_2))(X^+) = 2\Big|\sum_{i_1,\ldots,i_{k+1}\in S} p_1^{i_1} q_{n-1}^{i_1,i_2} q^{i_2,i_3}\cdots$$
$$q^{i_k,i_{k+1}}\int_{\mathbb{R}^{(n-1)m}} g^{(i_2,\ldots,i_{k+1})}(x_{n-1})a(x_{n-1},x_{n-2})\cdots a(x_1,x_0)\lambda_{(n-1)m}\left(dx^{(n-1)}\right)$$
$$-\sum_{i_1,\ldots,i_{k+1}\in S} p_2^{i_1} q_{n-1}^{i_1,i_2} q^{i_2,i_3}\ldots q^{i_k,i_{k+1}}\int_{\mathbb{R}^{(n-1)m}} g^{(i_2,\ldots,i_{k+1})}(x_{n-1})a(x_{n-1},x_{n-2})\cdots$$
$$a(x_1,x_0')\lambda_{(n-1)m}\left(dx^{(n-1)}\right)\Big| \leq 2\sum_{i_1,\ldots,i_{k+1}\in S} p_1^{i_1}|q_{n-1}^{i_1,i_2} - \pi^{i_2}|q^{i_2,i_3}\ldots q^{i_k,i_{k+1}}$$
$$\int_{\mathbb{R}^{(n-1)m}} g^{(i_2,\ldots,i_{k+1})}(x_{n-1})a(x_{n-1},x_{n-2})\cdots a(x_1,x_0)\lambda_{(n-1)m}\left(dx^{(n-1)}\right)$$
$$+2\sum_{i_1,\ldots,i_{k+1}\in S} p_2^{i_1}|q_{n-1}^{i_1,i_2} - \pi^{i_2}|q^{i_2,i_3}\ldots q^{i_k,i_{k+1}}\int_{\mathbb{R}^{(n-1)m}} g^{(i_2,\ldots,i_{k+1})}(x_{n-1})a(x_{n-1},x_{n-2})$$

$$\cdots a(x_1, x_0^{'})\lambda_{(n-1)m}\left(dx^{(n-1)}\right) + 2\sum_{i_2,\ldots,i_{k+1}\in S}\pi^{i_2}q^{i_2,i_3}\ldots q^{i_k,i_{k+1}}$$
$$\left|\int_{\mathbb{R}^{(n-1)m}} g^{(i_2,\ldots,i_{k+1})}(x_{n-1})a(x_{n-1},x_{n-2})\cdots a(x_1,x_0)\lambda_{(n-1)m}\left(dx^{(n-1)}\right)\right.$$
$$\left.-\int_{\mathbb{R}^{(n-1)m}} g^{(i_2,\ldots,i_{k+1})}(x_{n-1})a(x_{n-1},x_{n-2})\cdots a(x_1,x_0^{'})\lambda_{(n-1)m}\left(dx^{(n-1)}\right)\right|$$

As a consequence of the inequalities (4.13) and (4.34), the first two terms are less than $2a_1c_1^{n-1}$. The third term can be written as

$$T_3 = 2\sum_{i_2,\ldots,i_{k+1}\in S}\pi^{i_2}q^{i_2,i_3}\ldots q^{i_k,i_{k+1}}\left|\int_{\mathbb{R}^m} g^{(i_2,\ldots,i_{k+1})}(x_{n-1})\left(P_X^{n-1}(x_0,dx_{n-1})\right.\right.$$
$$\left.\left.-\pi_X(dx_{n-1})+\pi_X(dx_{n-1})-P_X^{n-1}(x_0^{'},dx_{n-1})\right)\right|.$$

Hence, Lemma 2.1.1 in [8], the assumption B and (4.34) imply

$$T_3 \le 2\rho_X^{n-1}\left(G(x_0)+G(x_0^{'})\right). \square$$

Theorem 4.3 *1. There exists positive constants K and $\rho < 1$, such that for any function $g \in L(V)$, (x_0,p), $(x_0^{'},p^{'}) \in V$ and any positive integer $n > 2$, we have*

$$|U^n g(x_0,p) - U^n g(x_0^{'},p^{'})| \le K\rho^n\left(1+G(x_0)+G(x_0^{'})\right)\|g\|_{BL}. \tag{4.35}$$

2. There exists a positive constant K_1 independent of $g \in L(V)$ and $(x_0,p) \in V$, and a constant $\Gamma(g)$, such that for any positive integer $n > 2$

$$|U^n g(x_0,p) - \Gamma(g)| \le K_1\frac{\rho^n}{1-\rho}\|g\|_{BL}\left(1+G(x_0)+\mu(x_0)\right). \tag{4.36}$$

Moreover, $u = \sum_{n\ge 0}(U^n g - \Gamma(g))$ is a solution of the equation $u - Uu = g - \Gamma(g)$.

PROOF. 1. Let us choose any positive integers $1 \le n_0 < n_1 < n$ and define $X = \mathbb{R}^{n_1(m+d)}$ and $Y = \mathbb{R}^{(n_1-n_0+1)(m+d)}$. For any $g \in L(V)$, $(x_0,p) \in V$, we have

$$U^n g(x_0,p) = \int_X P_{n_1}\left((x_0,p),d(y,x)^{(n1)}\right)U^{n-n_1}g\left((x_0,p)(y,x)^{(n_1)}\right),$$

$$\int_X P_{n_1}\left((x_0,p),d(y,x)^{(n1)}\right)U^{n-n_1}g\left((x_0,p)(y_{n_0},x_{n_0},\ldots,y_{n_1},x_{n_1})\right)$$
$$=\int_Y P^{n_0}_{n_1-n_0+1}\left((x_0,p),d(y,x)^{(n1-n_0+1)}\right)U^{n-n_1}g\left((x_0,p)(y,x)^{(n_1-n_0+1)}\right).$$

Using the previous two equations, we write

$$U^ng(x_0,p)-U^ng(x_0^{'},p^{'})=\int_X P_{n_1}\left((x_0,p),d(y,x)^{(n1)}\right)U^{n-n_1}g\left((x_0,p)(y,x)^{(n_1)}\right)$$
$$-\int_X P_{n_1}\left((x_0^{'},p^{'}),d(y,x)^{(n1)}\right)U^{n-n_1}g\left((x_0^{'},p^{'})(y,x)^{(n_1)}\right)=\int_X P_{n_1}((x_0,p),$$
$$d(y,x)^{(n1)})\left[U^{n-n_1}g\left((x_0,p)(y,x)^{(n_1)}\right)-U^{n-n_1}g\Big((x_0,p)(y_{n_0},x_{n_0},\ldots,y_{n_1}x_{n_1})\Big)\right]$$
$$+\int_Y P^{n_0}_{n_1-n_0+1}\left((x_0,p),d(y,x)^{(n1-n_0+1)}\right)\left[U^{n-n_1}g\Big((x_0,p)(y,x)^{(n1-n_0+1)}\Big)-\right.$$
$$\left.U^{n-n_1}g\left((x_0^{'},p^{'})(y,x)^{(n1-n_0+1)}\right)\right]+\int_Y P^{n_0}_{n_1-n_0+1}\left((x_0,p),d(y,x)^{(n1-n_0+1)}\right)$$
$$U^{n-n_1}g\Big((x_0^{'},p^{'})(y,x)^{(n1-n_0+1)}\Big)-\int_Y P^{n_0}_{n_1-n_0+1}\left((x_0^{'},p^{'}),d(y,x)^{(n1-n_0+1)}\right)$$
$$U^{n-n_1}g\Big((x_0^{'},p^{'})(y,x)^{(n1-n_0+1)}\Big)-\int_X P_{n_1}\left((x_0^{'},p^{'}),d(y,x)^{(n1)}\right)\Big[U^{n-n_1}g\Big((x_0^{'},p^{'})$$
$$(y,x)^{(n_1)}\Big)-U^{n-n_1}g\Big((x_0^{'},p^{'})(y_{n_0},x_{n_0},\ldots,y_{n_1}x_{n_1})\Big)\Big]$$

From Proposition 4.4, we get

$$|U^ng(x_0,p)-U^ng(x_0^{'},p^{'})|\leq 3C(1-\epsilon)^{n_1-n_0+1}\|g\|_{BL}+\left|\int_Y U^{n-n_1}g(\right.$$
$$\left.(x_0^{'},p^{'})(y,x)^{(n1-n_0+1)})\Lambda^{n_0}_{n_1-n_0+1}\left((x_0,p),(x_0^{'},p^{'}),d(y,x)^{(n1-n_0+1)}\right)\right| \tag{4.37}$$

Obviously, for any $(y,x)^{n_1-n_0+1}\in Y$

$$\left|U^{n-n_1}g\left((x_0,p)(y,x)^{(n1-n_0+1)}\right)\right|\leq\|g\|.$$

Hence, Lemma 2.1.1 in [8] and Proposition 4.6 yield

$$\left|\int_Y \Lambda^{n_0}_{n_1-n_0+1}\left((x_0,p),(x_0^{'},p^{'}),d(y,x)^{(n1-n_0+1)}\right)U^{n-n_1}g\big((x_0,p)(y,x)^{(n1-n_0+1)}\big)\right|$$

$$\leq \|g\| \left(4a_1 c_1^{n_0-1} + 2\rho_X^{n_0-1} \left(G(x_0) + G(x_0')\right)\right).$$

For $n > 2$, taking $n_0 = \lfloor n/2 \rfloor$, $n_1 = n - 1$, $\rho = \max\{c_1^{1/2}, \rho_X^{1/2}, (1-\epsilon)^{1/2}\} < 1$, and $K = \max\{3C + 4a_1/c_1^2, 2/\rho_X^2\}$, the previous inequality and (4.37) imply

$$|U^n g(x_0, p) - U^n g(x_0', p')| \leq K\rho^n \left(1 + G(x_0) + G(x_0')\right) \|g\|_{BL}.$$

2. Proceeding as in the proof of Lemma 1, page 252 in [2], we get the conclusion of the theorem using the inequality (4.35) and the assumption B. □

Corollary 4.1 *Under the probability measure* $\mathbf{P}_\bullet$, *the Markov chain* $\{w_{n+1}, X_n\}$ *has a unique invariant probability distribution* ν, *and for any* $g \in L(V)$, *we have*

$$\Gamma(g) = \int_V g(z)\nu(dz).$$

PROOF. The proof follows the ideas in [17] and in the proof of Proposition 2, page 253 in [2]. From the previous theorem, we can see that $g \to \Gamma(g)$ defines a positive linear functional on $L(V)$. Let $L_c(V) = \{g \in L(V) : g \text{ has compact support}\}$. Then, by the Riesz representation theorem [1], there exists ν, a positive Radon measure on V, such that for all $g \in L_c(V)$, we have

$$\Gamma(g) = \int_V g(z)\nu(dz). \tag{4.38}$$

Now we extend (4.38) to $L(V)$. For all positive integers $s \geq 1$ and all non-negative real number t, let define

$$\Phi_s(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq s/2 \\ -2t/s + 2 & \text{if } s/2 < t < s \\ 0 & \text{if } t \geq s. \end{cases}$$

Hence, for any t_1, $t_2 \in \mathbb{R}_+$, we have $|\Phi_s(t_1) - \Phi_s(t_2)| \leq 2|t_1 - t_2|/s$. Choose any $g \in L(V)$, $g \geq 0$, and for any positive integer $s \geq 1$ and all $(x, p) \in V$, define

$h_s(x,p) = g(x,p)f_s(x,p)$, where $f_s(x,p) = \Phi_s(\|x\| + \|p\|_1)$. Then h_s is a bounded continuous function with $\|h_s\| \leq \|g\|$. Moreover, for any (x_1,p_1), $(x_2,p_2) \in V$, we have

$$|f_s(x_1,p_1) - f_s(x_2,p_2)| \leq 2\,|\|x_1\| - \|x_2\| + \|p_1\|_1 - \|p_2\|_1|\,/s$$
$$\leq 2\,(|\|x_1| - \|x_2\|\| + |\|p_1\|_1 - \|p_2\|_1|)\,/s \leq 2\,(\|x_1 - x_2\| + \|p_1 - p_2\|_1)\,/s.$$

Hence, if $(x_1,p_1) \neq (x_2,p_2)$, then

$$\frac{|h_s(x_1,p_1) - h_s(x_2,p_2)|}{\|p_1 - p_2\|_1 + \|x_1 - x_2\|} \leq \frac{|f_s(x_1,p_1)||g(x_1,p_1) - g(x_2,p_2)|}{\|p_1 - p_2\|_1 + \|x_1 - x_2\|}$$
$$+ \frac{|g(x_2,p_2)||f_s(x_1,p_1) - f_s(x_2,p_2)|}{\|p_1 - p_2\|_1 + \|x_1 - x_2\|} \leq s(g) + \frac{2}{s}\|g\|$$

Thus, $h_s \in L_c(V)$ for any positive integer $s \geq 1$, and $\|h_s\|_{BL} \leq 3\|g\|_{BL}$. From (4.36), for any $(x_0,p) \in V$, we get

$$\left|U^n h_s(x_0,p) - \int_V h_s(z)\nu(dz)\right| \leq K_1 \frac{3\rho^n}{1-\rho}\|g\|_{BL}(1 + G(x_0) + \mu(x_0)).$$

Since h_s converges non-decreasingly to g, as $s \to \infty$, passing to the limit in the previous inequality, for any $(x_0,p) \in V$, we obtain

$$\left|U^n g(x_0,p) - \int_V g(z)\nu(dz)\right| \leq K_1 \frac{3\rho^n}{1-\rho}\|g\|_{BL}(1 + G(x_0) + \mu(x_0)).$$

Together with (4.36), this proves that for any $g \in L(V)$, $g \geq 0$, we have

$$\int_V g(z)\nu(dz) = \Gamma(g) < \infty.$$

Choosing $g(z) = 1$ for all $z \in V$, we get $\nu(V) = 1$, and thus ν is a probability. Any function $g \in L(V)$ can be written as a difference of two positive functions $g = g^+ - g^-$ with

$$g^+(z) = (|g(z)| + g(z))/2, \quad g^-(z) = (|g(z)| - g(z))/2, \quad z \in V.$$

If $g \in L(V)$, then $|g| \in L(V)$, and as a consequence g^+ and $g^- \in L(V)$. Hence, for any $g \in L(V)$

$$\int_V g(z)\nu(dz) = \Gamma(g).$$

Now we prove that ν is an invariant probability for the Markov chain $\{w_{n+1}, X_n, n \geq 0\}$. Let $\Pi^n(\cdot|\cdot)$ be the $n-$ step transition probability function (for $n = 1$, $\Pi(\cdot|\cdot)$ is given in (4.23)). For any $g \in L(V)$, we have

$$\lim_{n\to\infty} U^n g(\cdot) = \lim_{n\to\infty} \int_V g(z)\Pi^n(dz|\cdot) = \int_V g(z)\nu(dz).$$

The previous equality implies that, for any $(x,p) \in V$ the probabilities $\Pi^n(\cdot|(x,p))$ converge weakly to ν. Hence, for any $g \in C^\infty(V)$, we have

$$\lim_{n\to\infty} \int_V g(z)\Pi^n(dz|\cdot) = \int_V g(z)\nu(dz) = \lim_{n\to\infty} U^n g(\cdot).$$

From Proposition 4.5, we know that for any function $g \in C^\infty(V)$, we have $Ug \in C^\infty(V)$. Hence, for any function $g \in C^\infty(V)$ and any $z \in V$, we can write

$$\int_V g(z)\nu(dz) = \lim_{n\to\infty} U^{n+1} g(z) = \lim_{n\to\infty} U^n(Ug)(z) = \int_V Ug(z)\nu(dz).$$

Thus, the probability ν is an invariant probability for the Markov chain $\{w_{n+1}, X_n\}$.

Suppose that there exists also other probability measure ν', such that for all $g \in L(V)$ and any positive integer $n \geq 1$

$$\int_V g(z)\nu'(dz) = \int_V U^n g(z)\nu'(dz).$$

Hence, by the Lebesgue's theorem of dominated convergence, we have

$$\int_V g(z)\nu'(dz) = \lim_{n\to\infty} \int_V U^n g(z)\nu'(dz) = \int_V \Gamma(g)\nu'(dz) = \Gamma(g) = \int_V g(z)\nu(dz).$$

We arbitrarily choose a compact set $\mathbf{K} \subset V$ and for any positive integer $s \geq 1$, let define the functions $\phi_s(z) = 1 - \min(1, sd(z,\mathbf{K}))$, $z \in V$. It is easy to prove that

$\phi_s \in L(V)$ and non-increasingly $\lim_{s\to\infty} \phi_s = 1_{\mathbf{K}}$. Hence

$$\nu(\mathbf{K}) = \lim_{s\to\infty} \int_V \phi_s(z)\nu(dz) = \lim_{s\to\infty} \int_V \phi_s(z)\nu'(dz) = \nu'(\mathbf{K}).$$

From the uniqueness theorem [1], we get $\nu = \nu'$. □

4.4.2 The model with misspecified parameters

For any vectors $y \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$, we define

$$\hat{\delta}(y,x) = \frac{\max_{j\in S} \hat{b}^j(y,x)}{\min_{j\in S} \hat{b}^j(y,x)}, \quad \hat{\Delta}_{-1}(x) = \min_{i\in S} \int_{\mathbb{R}^d} \hat{\delta}^{-1}(y,x)\hat{b}^i(y,x)\lambda_d(dy) \quad \hat{\epsilon} = \min_{i,j\in S}{}^{+} \hat{q}^{i,j}$$

$$\hat{R} = \hat{\epsilon}^r \int_{\mathbb{R}^m} \cdots \int_{\mathbb{R}^m} \prod_{k=2}^{r} \hat{\Delta}_{-1}(x_k) a(x_k, x_{k-1})\, \lambda_m(dx_r) \cdots \lambda_m(dx_2)\pi_X(dx_1).$$

Proposition 4.7 *If the stochastic matrix $\hat{Q}$ is primitive with index of primitivity r, then for any $p_1, p_2 \in W$, and any integers k, l such that $k \geq l + r - 1$, and any sequence $y_k, x_k, \ldots, y_l, x_l$, we have*

$$\left\| \hat{F}[y_k, x_k, \ldots, y_l, x_l, p_1] - \hat{F}[y_k, x_k, \ldots, y_l, x_l, p_2] \right\|_1 \leq \hat{\epsilon}^{-r} \hat{\delta}(y_l, x_l) \cdots \hat{\delta}(y_{l+r-1},$$

$$x_{l+r-1}) \prod_{i=1}^{[k,l]} \left(1 - \hat{\epsilon}^r [\hat{\delta}(y_{l+(i-1)r+1}, x_{l+(i-1)r+1}) \cdots \hat{\delta}(y_{l+ir-1}, x_{l+ir-1})]^{-1}\right) \|p_1 - p_2\|_1,$$

$$\left\| \hat{F}[y_k, x_k, \ldots, y_l, x_l, p_1] - \hat{F}[y_k, x_k, \ldots, y_l, x_l, p_2] \right\|$$

$$\leq 2 \prod_{i=1}^{[k,l]} \left(1 - \hat{\epsilon}^r [\hat{\delta}(y_{l+(i-1)r+1}, x_{l+(i-1)r+1}) \cdots \hat{\delta}(y_{l+ir-1}, x_{l+ir-1})]^{-1}\right),$$

where $[k,l]$ is the maximum number of disjoints blocks of length r in the set $\{l, \ldots, k\}$.

Proof. For any $p \in W$, we have

$$\hat{F}[y_k, x_k, \ldots, y_l, x_l, p] = \hat{Q}^t \hat{B}(y_k, x_k) \cdots \hat{Q}^t \hat{B}(y_l, x_l) \cdot p.$$

Using this equation, the proof is a consequence of Theorem 3.5 in [11]. □

Remark 4.7 If $r = 1$, i.e. $\hat{Q}$ is positive, then we have

$$\left\| \hat{F}[y_k, x_k, \ldots, y_l, x_l, p_1] - \hat{F}[y_k, x_k, \ldots, y_l, x_l, p_2] \right\| \leq 2(1 - \hat{\epsilon})^{k-l+1}. \tag{4.39}$$

The results stated in Proposition 4.7 allow us to obtain an upper bound for the a.s. exponential rate of forgetting of the initial condition for (4.22).

ASSUMPTION A′ The Matrix Q is primitive.

Theorem 4.4 *Under the assumptions A′ and B, if the stochastic matrix $\hat{Q}$ is primitive with index of primitivity r, then for any p_1, $p_2 \in W$ and any positive integer l we have*

$$\limsup_{k\to\infty} \frac{1}{k} \log \|\hat{F}[Y_k, X_k, \ldots, Y_l, X_l, p_1] - \hat{F}[Y_k, X_k, \ldots, Y_l, X_l, p_2]\|_1$$
$$\leq r^{-1} \log(1 - \hat{R}) < 0, \quad \mathbf{P}_\bullet a.s.$$

Proof. Under the stated assumptions, the Markov chain $\{S_k, X_k, Y_k\}$ is geometrically ergodic, with a unique invariant probability distribution $\sigma_\bullet$ on $S \times \mathbb{R}^d \times \mathbb{R}^m$. For any $i \in S$, we have $\sigma^i(dydx) = \mu^i b^i(x, y)\lambda_d(dy)\pi_X(dx)$.

As in the proof of Theorem 2.2 in [12], using Proposition 4.7, we get

$$\limsup_{k\to\infty} \frac{1}{k} \log \|\hat{F}[Y_k, X_k, \ldots, Y_l, X_l, p_1] - \hat{F}[Y_k, X_k, \ldots, Y_l, X_l, p_2]\|_1 \leq \limsup_{k\to\infty} \frac{1}{k}$$
$$\times \sum_{i=1}^{[k,l]} \log\left(1 - \hat{\epsilon}^r [\hat{\delta}(Y_{l+(i-1)r+1}, X_{l+(i-1)r+1}) \cdots \hat{\delta}(Y_{l+ir-1}, X_{l+ir-1})]^{-1}\right) = r^{-1}$$
$$\times \limsup_{L\to\infty} L^{-1} \sum_{i=1}^{L} \log\left(1 - \hat{\epsilon}^r [\hat{\delta}(Y_{l+(i-1)r+1}, X_{l+(i-1)r+1}) \cdots \hat{\delta}(Y_{l+ir-1}, X_{l+ir-1})]^{-1}\right).$$

Moreover, the ergodic theorem yield

$$\lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \log(1 - \hat{\epsilon}^r [\hat{\delta}(Y_{l+(i-1)r+1}, X_{l+(i-1)r+1}) \cdots \hat{\delta}(Y_{l+ir-1}, X_{l+ir-1})]^{-1}) = \sum_{i_1=1}^{M}$$

$$\cdots\sum_{i_r=1}^{M}\int_{\mathbb{R}^m}\cdots\int_{\mathbb{R}^m}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}\log(1-\hat{\epsilon}^r[\hat{\delta}(y_2,x_2)\cdots\hat{\delta}(y_r,x_r)]^{-1})\mu^{i1}q^{i_1,i_2}\cdots q^{i_{r-1},i_r}$$

$$b^{i_1}(y_1,x_1)\cdots b^{i_r}(y_r,x_r)\prod_{i=2}^{r}a(x_i,x_{i-1})\lambda_d(dy_1)\lambda_m(dx_2)\cdots\lambda_d(dy_r)\lambda_m(dx_r)\pi_X(dx_1)$$

From the Jensen inequality, we obtain

$$\lim_{L\to\infty}\frac{1}{L}\sum_{i=1}^{L}\log\left(1-\hat{\epsilon}^r[\hat{\delta}(Y_{l+(i-1)r+1},X_{l+(i-1)r+1})\cdots\hat{\delta}(Y_{l+ir+1},X_{l+ir+1})]^{-1}\right)\leq\sum_{i_1=1}^{M}$$

$$\cdots\sum_{i_r=1}^{M}\mu^{i_1}q^{i_1,i_2}\cdots q^{i_{r-1},i_r}\log\int_{\mathbb{R}^m}\cdots\int_{\mathbb{R}^m}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}(1-\hat{\epsilon}^r[\hat{\delta}(y_2,x_2)\cdots\hat{\delta}(y_r,x_r)]^{-1})$$

$$b^{i_1}(y_1,x_1)\cdots b^{i_r}(y_r,x_r)\lambda_d(dy_1)\cdots\lambda_d(dy_r)\prod_{i=2}^{r}a(x_i,x_{i-1})\,\lambda_m(dx_r)\ldots\lambda_m(dx_2)\pi(dx_1)$$

$$=\sum_{i_1=1}^{M}\cdots\sum_{i_r=1}^{M}\mu^{i1}q^{i_1,i_2}\cdots q^{i_{r-1},i_r}\log\Bigg(1-\hat{\epsilon}^r\int_{\mathbb{R}^m}\cdots\int_{\mathbb{R}^m}\prod_{t=2}^{r}\Bigg(\int_{\mathbb{R}^d}\hat{\delta}^{-1}(y_t,x_t)b^{i_t}(y_t,x_t)$$

$$\lambda_d(dy_t)\Bigg)\prod_{i=2}^{r}a(x_i,x_{i-1})\,\lambda_m(dx_r)\ldots\lambda_m(dx_2)\pi_X(dx_1)\Bigg)\leq\sum_{i_1=1}^{M}\cdots\sum_{i_r=1}^{M}\mu^{i1}q^{i_1,i_2}\cdots$$

$$q^{i_{r-1},i_r}\log\Bigg(1-\hat{\epsilon}^r\int_{\mathbb{R}^m}\cdots\int_{\mathbb{R}^m}\prod_{i=2}^{r}\hat{\Delta}_{-1}(x_i)a(x_i,x_{i-1})\,\lambda_m(dx_r)\ldots\lambda_m(dx_2)\pi_X(dx_1)\Bigg)$$

$$=\log(1-\hat{R})<0.\quad\square$$

ASSUMPTION C The Matrix $\hat{Q}$ is positive.

Now we study the ergodic properties of the stochastic process $\{\hat{w}_n, n\geq 1\}$ defined in (4.22). It is easy to show that under the true probability measure $\mathbf{P}_\bullet$ corresponding to the true values Q, $\mathbf{p}_\bullet$ and $b_\bullet$, the Markov chain $\{S_n, X_n, Y_n, \hat{w}_n, n\geq 1\}$ has the transition probability

$$\hat{\Pi}(j,X,Y,Z|i,x',y',p')=\mathbf{P}_\bullet(S_{n+1}=j,X_{n+1}\in X,Y_{n+1}\in Y,\hat{w}_{n+1}\in Z|S_n=i,X_n$$

$$=x',Y_n=y',\hat{w}_n=p')=q^{i,j}\int_X\int_Y b^j(y,x)a(x,x')1_Z(\hat{F}[y',x',p'])\lambda_d(dy)\lambda_m(dx),$$

for any $n\geq 1$, $j\in S$, $Z\in\mathcal{W}$, $X\in\mathcal{B}(\mathbb{R}^m)$, $Y\in\mathcal{B}(\mathbb{R}^d)$ and $(i,x',y',p')\in\hat{V}$. For any real-valued, bounded and $\left(\hat{V},\mathcal{B}(R)\right)$ measurable function g defined on $\hat{V}$, any

$n > 1$, and any (i_0, x_0, y_0, p), $(i, x^{'}, y^{'}, p^{'}) \in \hat{V}$, we have

$$\hat{U}g(i, x^{'}, y^{'}, p^{'}) = \mathbf{E}_{\bullet}\left[g(S_{n+1}, X_{n+1}, Y_{n+1}, \hat{w}_{n+1})\Big| S_n = i, X_n = x^{'}, Y_n = y^{'}, \hat{w}_n = p^{'}\right]$$

$$= \sum_{j \in S} q^{i,j} \int_{\mathbb{R}^{d+m}} b^j(y, x) a(x, x^{'}) g\left(j, x, y, \hat{F}[y^{'}, x^{'}, p^{'}]\right) \lambda_{d+m}\left(d(y, x)\right),$$

$$\hat{U}^n g(i_0, x_0, y_0, p) = \mathbf{E}_{\bullet}\Big[g(S_{n+1}, X_{n+1}, Y_{n+1}, \hat{w}_{n+1}) | S_1 = i_0, \hat{w}_1 = p, X_1 = x_0, Y_1 = y_0\Big]$$

$$= \sum_{i_1, \ldots, i_n \in S} q^{i_0, i_1} q^{i_1, i_2} \ldots q^{i_{n-1}, i_n} \int_{\mathbb{R}^{n(d+m)}} b^{i_1}(y_1, x_1) a(x_1, x_0) \cdots b^{i_n}(y_n, x_n) a(x_n, x_{n-1})$$

$$g(i_n, x_n, y_n, \hat{F}[y_{n-1}, x_{n-1}, \ldots, y_1, x_1, x_0, y_0, p]) \lambda_{n(d+m)}\left(d(y, x)^{(n)}\right). \tag{4.40}$$

Proposition 4.8 *Under the assumptions $A^{'}$, B and C, if $g \in C^{\infty}(\hat{V})$, then $\hat{U}(g) \in C^{\infty}(\hat{V})$.*

PROOF. The proof is similar to the proof of Proposition 4.5 .□

Theorem 4.5 *Under the assumptions $A^{'}$, B and C, we have:*

1. There exists positive constants $\hat{K}$ and $\hat{\rho} < 1$, such that we have

$$|\hat{U}^n g(i, x_0, y_0, p) - \hat{U}^n g(i^{'}, x_0^{'}, y^{'}, p^{'})| \leq \hat{K}\hat{\rho}^n \left(n + G(x_0) + G(x_0^{'})\right) \|g\|_{BL},$$

for any function $g \in L(\hat{V})$, any (i, x_0, y_0, p), $(i^{'}, x_0^{'}, y_0^{'}, p^{'}) \in \hat{V}$ and any positive integer $n > 2$;

2. There exists a positive constant $\hat{K}_1$ independent of $g \in L(\hat{V})$ and $(i, x_0, y_0, p) \in \hat{V}$, and a constant $\hat{\Gamma}(g)$ depending only on g, such that

$$|\hat{U}^n g(i, x_0, y_0, p) - \hat{\Gamma}(g)| \leq \hat{K}_1 \frac{\hat{\rho}^n}{(1 - \hat{\rho})^2} \|g\|_{BL} \left(n + G(x_0) + \mu(x_0)\right),$$

for any positive integer $n > 2$. Moreover, $u = \sum_{n \geq 0}(\hat{U}^n g - \hat{\Gamma}(g))$ is a solution of the Poisson equation $u - \hat{U}u = g - \hat{\Gamma}(g)$.

Proof. 1. Notice that the following decomposition holds

$$\hat{U}^n g(i, x_0, y_0, p) - \hat{U}^n g(i^{'}, x_0^{'}, y^{'}, p^{'}) = \left(\hat{U}^n g(i, x_0, y_0, p) - \hat{U}^n g(i, x_0^{'}, y_0, p)\right)$$

$$+\left(\hat{U}^n g(i,x_0^{'},y_0,p)-\hat{U}^n g(i,x_0^{'},y_0^{'},p^{'})\right)+\left(\hat{U}^n g(i,x_0^{'},y_0^{'},p^{'})-\hat{U}^n g(i^{'},x_0^{'},y_0^{'},p^{'})\right)$$

To estimate the first term, let us take $1 \leq n0 = \lfloor n/2 \rfloor < n$. We have

$$T_1 = \left|\hat{U}^n g(i,x_0,y_0,p)-\hat{U}^n g(i,x_0^{'},y_0,p)\right| \leq \left|\hat{U}^n g(i,x_0,y_0,p) - \sum_{i_1,\ldots,i_n \in S} q^{i,i_1}q^{i_1,i_2}\right.$$

$$\ldots q^{i_{n-1},i_n}\int_{\mathbb{R}^{n(d+m)}} b^{i_1}(y_1,x_1)a(x_1,x_0)\cdots b^{i_n}(y_n,x_n)a(x_n,x_{n-1})g\Big(i_n,x_n,y_n,$$

$$\hat{F}[y_{n-1},x_{n-1}\ldots,y_{n0},x_{n0},x_0,y_0,p]\Big)\lambda_{n(d+m)}\left(d(y,x)^{(n)}\right)\Big| + \Big|\sum_{i_1,\ldots,i_n \in S} q^{i,i_1}q^{i_1,i_2}\ldots$$

$$q^{i_{n-1},i_n}\int_{\mathbb{R}^{n(d+m)}} b^{i_1}(y_1,x_1)a(x_1,x_0)\cdots b^{i_n}(y_n,x_n)a(x_n,x_{n-1})g\Big(i_n,x_n,y_n,\hat{F}[y_{n-1},$$

$$x_{n-1}\ldots,y_{n0},x_{n0},x_0,y_0,p]\Big)\lambda_{n(d+m)}\left(d(y,x)^{(n)}\right) - \sum_{i_1,\ldots,i_n \in S} q^{i,i_1}q^{i_1,i_2}\ldots q^{i_{n-1},i_n}$$

$$\int_{\mathbb{R}^{n(d+m)}} b^{i_1}(y_1,x_1)a(x_1,x_0^{'})\cdots b^{i_n}(y_n,x_n)a(x_n,x_{n-1})g\Big(i_n,x_n,y_n,\hat{F}[y_{n-1},$$

$$x_{n-1}\ldots,y_{n0},x_{n0},x_0,y_0,p]\Big)\lambda_{n(d+m)}\left(d(y,x)^{(n)}\right)\Big| + \Big|\sum_{i_1,\ldots,i_n \in S} q^{i,i_1}q^{i_1,i_2}\ldots q^{i_{n-1},i_n}$$

$$\int_{\mathbb{R}^{n(d+m)}} b^{i_1}(y_1,x_1)a(x_1,x_0^{'})\cdots b^{i_n}(y_n,x_n)a(x_n,x_{n-1})g\Big(i_n,x_n,y_n,\hat{F}[y_{n-1},x_{n-1}$$

$$\ldots,y_{n0},x_{n0},x_0,y_0,p]\Big)\lambda_{n(d+m)}\left(d(y,x)^{(n)}\right) - \hat{U}^n g(i,x_0^{'},y_0,p)\Big|$$

For any $x \in \mathbb{R}^m$ and any sequence $i_1,\ldots,i_{n-n0+1} \in S$, let

$$g^{i_1,\ldots,i_{n-n0+1}}_{x_0,y_0,p}(x) = \int_{\mathbb{R}^{(n-n0+1)(d+m)}} b^{i_1}(y_1,x_1)a(x_1,x)\cdots b^{i_{n-n0+1}}(y_{n-n0+1},x_{n-n0+1})$$

$$a(x_{n-n0+1},x_{n-n0})g\Big(i_{n-n0+1},x_{n-n0+1},y_{n-n0+1},\hat{F}[y_{n-n0},x_{n-n0}\ldots,y_1,x_1,x_0,$$

$$y_0,p]\Big)\lambda_{(n-n0+1)(d+m)}\left(d(y,x)^{(n-n0+1)}\right)$$

Using (4.39), we obtain

$$T_1 \leq 4s(g)(1-\hat{\epsilon})^{n-n0} + \Big|\sum_{i_{n0},\ldots,i_n \in S} q^{i,i_{n0}}_{n0}q^{i_{n0},i_{n0+1}}\ldots q^{i_{n-1},i_n}\int_{\mathbb{R}^{(n0-1)m}} g^{i_{n0},\ldots,i_n}_{x_0,y_0,p}($$

$$x_{n0-1})a(x_{n0-1},x_{n0-2})\cdots a(x_1,x_0)\lambda_{(n-n0+1)m}\left(d(x)^{(n-n0+1)}\right) - \sum_{i_{n0},\ldots,i_n \in S} q^{i,i_{n0}}_{n0}$$

$$q^{i_{n0},i_{n0+1}}\dots q^{i_{n-1},i_n}\int_{\mathbb{R}^{(n0-1)m}} g^{i_{n0},\dots,i_n}_{x_0,y_0,p}(x_{n0-1})a(x_{n0-1},x_{n0-2})\cdots a(x_1,x_0^{'})$$
$$\lambda_{(n-n0+1)m}\left(d(x)^{(n-n0+1)}\right)\Big| \le 4s(g)(1-\hat{\epsilon})^{n-n0}+\sum_{i_{n0},\dots,i_n\in S} q^{i,i_{n0}}_{n0}q^{i_{n0},i_{n0+1}}\dots q^{i_{n-1},i_n}$$
$$\left|\int_{\mathbb{R}^m} g^{i_{n0},\dots,i_n}_{x_0,y_0,p}(x_{n0-1})P_X^{n0-1}(x_0,dx_{n0-1})-\int_{\mathbb{R}^m} g^{i_{n0},\dots,i_n}_{x_0,y_0,p}(x_{n0-1})P_X^{n0-1}(x_0^{'},dx_{n0-1})\right|$$

Since for any $x\in\mathbb{R}^m$,

$$|g^{i_1,\dots,i_{n-n0+1}}_{x_0,y_0,p}(x)|\le\|g\|,$$

from Lemma 2.1.1 in [8] and assumption B, we have

$$T_1\le 4s(g)(1-\hat{\epsilon})^{n-n0}+\|g\|\rho_X^{n0-1}\left(G(x_0)+G(x_0^{'})\right).$$

Let $c_{max}=\max\{(1-\hat{\epsilon})^{1/2},(\rho_X)^{1/2}\}$ and $a_{max}=\max\{4,\rho_X^{-2}\}$. We get

$$T_1\le\|g\|_{BL}a_{max}c^n_{max}\left(1+G(x_0)+G(x_0^{'})\right). \tag{4.41}$$

Replacing in (4.40) and using (4.39), it is easy to show that for the second term

$$T_2=\left|\hat{U}^ng(i,x_0^{'},y_0,p)-\hat{U}^ng(i,x_0^{'},y_0^{'},p^{'})\right|\le 2s(g)(1-\hat{\epsilon})^{n-1}. \tag{4.42}$$

To estimate the third term, we use the same decomposition as in the proof of Theorem 3.5 in [12]. For any positive integer $n1<n-1$, and any $z_l\in\mathbb{R}^{d+m}$, $1\le l\le n1$, we have

$$T_3=\left|\hat{U}^ng(i,x_0^{'},y_0^{'},p^{'})-\hat{U}^ng(i^{'},x_0^{'},y_0^{'},p^{'})\right|\le\sum_{l=1}^{n1}\sum_{i_1,\dots,i_n\in S}\Big|(q^{i,i_1}-q^{i^{'},i_1})$$
$$q^{i_1,i_2}\dots q^{i_{n-1},i_n}\int_{\mathbb{R}^{n(d+m)}} b^{i_1}(y_1,x_1)a(x_1,x_0^{'})\cdots b^{i_n}(y_n,x_n)a(x_n,x_{n-1})$$
$$\Big(g\big(i_n,x_n,y_n,\hat{F}[y_{n-1},x_{n-1}\dots,y_l,x_l,z_{l-1},\dots,z_1,x_0^{'},y_0^{'},p^{'}]\big)-g\big(i_n,x_n,y_n,$$
$$\hat{F}[y_{n-1},x_{n-1}\dots,y_{l+1},x_{l+1},z_l,\dots,z_1,x_0^{'},y_0^{'},p^{'}]\big)\Big)\lambda_{n(d+m)}\left(d(y,x)^{(n)}\right)\Big|$$
$$+\sum_{i_1,\dots,i_n\in S}\Big|(q^{i,i_1}-q^{i^{'},i_1})q^{i_1,i_2}\dots q^{i_{n-1},i_n}\int_{\mathbb{R}^{n(d+m)}} b^{i_1}(y_1,x_1)a(x_1,x_0^{'})$$

$$\cdots b^{i_n}(y_n, x_n)a(x_n, x_{n-1})g\Big(i_n, x_n, y_n, \hat{F}[y_{n-1}, x_{n-1}\ldots, y_{n1+1}, x_{n1+1}, z_{n1},$$
$$\ldots, z_1, x_0^{'}, y_0^{'}, p^{'}]\Big)\lambda_{n(d+m)}\left(d(y,x)^{(n)}\right)\Big|.$$

Elementary calculations and the inequalities (4.39) and (4.14) yield

$$T_3 \leq 2s(g)\sum_{l=1}^{n1}\sum_{i_l,\ldots,i_n\in S}\left|q_l^{i,i_l} - q_l^{i^{'},i_l}\right| q^{i_l,i_{l+1}}\ldots q^{i_{n-1},i_n}(1-\hat{\epsilon})^{n-l-1} + \|g\|$$
$$\times \sum_{i_{n1+1},\ldots,i_n\in S} q^{i_{n1+1},i_{n1+2}}\ldots q^{i_{n-1},i_n}\left|q_{n1+1}^{i,i_{n1+1}} - q_{n1+1}^{i^{'},i_{n1+1}}\right| \leq 2s(g)\hat{a}_1\sum_{l=1}^{n1}\hat{c}_1^l$$
$$\times (1-\hat{\epsilon})^{n-l-1} + \|g\|\hat{a}_1\hat{c}_1^{n1+1}$$

Let $d_{max} = \max\{\hat{c}_1, 1-\hat{\epsilon}\}$ and $n_1 = n-2$. Then

$$T_3 = \left|\hat{U}^n g(i, x_0^{'}, y_0^{'}, p^{'}) - \hat{U}^n g(i^{'}, x_0^{'}, y_0^{'}, p^{'})\right| \leq 2(n-2)\hat{a}_1\|g\|_{BL}d_{max}^{n-1}. \qquad (4.43)$$

Adding (4.41), (4.42), and (4.43), we get the inequality in the first part of the theorem, with $\hat{\rho} = \max\{c_{max}, d_{max}\}$ and $\hat{K} = \max\{2\hat{a}_1/d_{max}, 2/(1-\hat{\epsilon}), a_{max}\}$.

2. Proceeding as in the proof of Lemma 1, page 252 in [2], and using the first part of the theorem and the assumption B, we can easily prove the second part of the theorem. □

Corollary 4.2 *Under the probability measure* $\mathbf{P}_\bullet$, *the Markov chain* $\{S_n, X_n, Y_n, \hat{w}_n\}$ *has a unique invariant probability distribution* $\hat{\tau}$ *on* $\hat{V}$, *and for any function* $g \in L(\hat{V})$ *we have*

$$\hat{\Gamma}(g) = \int_{\hat{V}} g(z)\hat{\tau}(dz).$$

Proof. Using Proposition 4.8, we can prove the corollary similarly to corollary 4.1.

Bibliography

[1] H. Bauer. *Probability Theory and Elements of Measure Theory.* Academic Press, London, 1981.

[2] M. Benveniste, A. Métivier and P. Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag, Berlin, 1990.

[3] C.K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81:541–553, 1994.

[4] Z. Ghahramani. An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1):9–42, 2001.

[5] Z. Ghahramani and G.E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–996, 2000.

[6] H. Hennion. Sur un théoréme spectral et son application aux noyaux lipchitziens. *Proc. Amer. Math. Soc.*, 118(2):627–634, 1993.

[7] M. Iosifescu. *Finite Markov Processes and Their Applications.* John Wiley, New York, 1980.

[8] M. Iosifescu and S. Grigorescu. *Dependence with Complete Connections and Its Applications.* Cambridge University Press, Cambridge, 1990.

[9] T. Kaijser. A limit theorem for partially observed Markov chains. *Ann. Probab.*, 3:677–696, 1975.

[10] H. Kunita. Asymptotic behavior of the nonlinear filtering errors of Markov processes. *J. Multivariate Anal.*, 1:365–393, 1971.

[11] F. Le Gland and L. Mevel. Basic properties of the projective product, with application to products of column-allowable non-negative matrices. *Math. Control Signals Systems*, 13:41–62, 2000.

[12] F. Le Gland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Math. Control Signals Systems*, 13:63–93, 2000.

[13] G.J. McLachlan and D. Peel. *Finite Mixture Models.* John Wiley, New York, 2000.

[14] C. A. Popescu and Y. S. Wong. Ergodicity in models with hidden Markov switching. submitted to Math. Control Signals Systems, 2004.

[15] R.H. Shumway and D.S. Stoffer. Dynamic linear models with switching. *J. Amer. Stat. Asoc.*, 86:763–769, 1991.

[16] L. Stettner. On invariant measures of filtering processes. In K. Christopeit, N. Helmes and M. Kohlmann, editors, *Stochastic Differential Systems*, Lecture Notes in Control and Inform. Sci., pages 279–292. Springer-Verlag, Berlin, 1989.

[17] C. Sunyach. Une classe de chaines récurrentes sur un espace métrique complet. *Ann. Inst. H. Poincaré Probab. Statist.*, 11:325–343, 1975.

Chapter 5

Monte Carlo approach for switching state-space models

5.1 Introduction

In recent years, various switching state-space models (SSMs) have been developed with applications across board ranges of disciplines from econometrics [8] to control engineering [18]. In this chapter, we are mainly concerned with the models introduced in [6]. They are represented mathematically by

$$X_k^m = A_m X_{k-1}^m + w_k^m, \quad w_k^m \sim N(0, H_m), \quad m = 1, \ldots, M \tag{5.1}$$

$$Y_k = C_{S_k} X_k^{S_k} + v_k, \quad v_k \sim N(0, R), \tag{5.2}$$

where $\{Y_k, k = 1, \ldots, T\}$ is the sequence of observations, $\{X_k^m, k = 1, \ldots, T,\ m = 1, \ldots M\}$ are M sequences of real valued hidden state vectors, and $\{S_k, k = 1, \ldots, T\}$ is a sequence of discrete hidden state vectors. The Bayesian network associated with these models is displayed in Fig. 1.4. Let denote $Y_{(t)} = \{Y_1, \ldots, Y_t\}$, $X_{(t)} = \{X_1^1, \ldots, X_1^m, \ldots, X_t^1, \ldots, X_t^m\}$ and $S_{(t)} = \{S_1, \ldots, S_t\}$. The discrete switching state

The material presented in this chapter was previously published in [15]

can take M values, $S_k \in \{1, \ldots, M\}$, and has a homogeneous Markovian structure specified by the initial probabilities π and the transition matrix Φ. The zero-mean Gaussian noise vectors w_k^m and v_k are uncorrelated with covariance matrices H_m, $m = 1, \ldots, M$, and R, respectively, and they are independent of the sequence $S_{(T)}$. A_m are the $s \times s$ transition matrices and C_m are the $d \times s$ output matrices for the state-space model m.

If the state vectors and the noise vectors are concatenated into large state and noise vectors $X_k = ((X_k^1)^t, \ldots, (X_k^M)^t)^t$ and $W_k = ((w_k^1)^t, \ldots, (w_k^M)^t)^t$, the model (5.1) - (5.2) is equivalent with the following model [18]:

$$\begin{aligned} X_k &= AX_{k-1} + W_k, \quad W_k \sim N(0, H), \\ Y_k &= C'_{S_k} X_k + v_k, \quad v_k \sim N(0, R), \end{aligned} \tag{5.3}$$

where $A = \text{diag}(A_m)$, $H = \text{diag}(H_m)$, $m = 1, \ldots, M$, and $C'_{S_k} = (0, \ldots, C^t_{S_k}, \ldots, 0)^t$.

Notice that for $M = 1$ (i.e. no discrete switching variable), we obtain a linear Gaussian state-space model. The Expectation Maximization (EM) algorithm, with the E step based on the Kalman filter and smoother can be applied for system identification [17]. On the other hand, if we keep only the observation sequence $\{Y_k\}$ and the discrete switching variables $\{S_k\}$, we are in the classical setting of hidden Markov chains. For learning the parameters, we can employ the Baum-Welch algorithm (see Chapter 4, Section 4.3.1).

The E step becomes intractable [6], if we attempt to implement the EM algorithm for the switching SSM (5.1)-(5.2). To overcome this difficulty, approximate inference methods are used (see [13] for a justification of the convergence of the variants of the EM algorithm). One possibility is to use a variational approach [6]: an approximate $Q(S_{(T)}, X_{(T)})$ of the posterior probability $P(S_{(T)}, X_{(T)}|Y_{(T)})$ of the hidden states is constructed, and the Kullback-Lieber divergence between Q and P is minimized in each cycle of the EM algorithm. Other methods based on the Generalized Pseudo Bayesian (GPB) algorithm ([8], pp. 99-109) have also been proposed. For example, a first order GPB algorithm merges a mixture of Gaussians into a single Gaussian

according to the formula:

$$E[X_k|Y_1,\ldots,Y_k] = \sum_{i=1}^{M} P(S_k = i|Y_1,\ldots,Y_k)E[X_k|S_k = i, Y_1,\ldots,Y_k].$$

A better approximation can be achieved by a second order algorithm which merges Gaussian according to their two-step history:

$$E[X_k|S_k = i, Y_1,\ldots,Y_k] = \sum_{i=1}^{M} P(S_k = i|S_{k-1} = j, Y_1,\ldots,Y_k)E[X_k|S_k = i,$$

$$S_{k-1} = j, Y_1,\ldots,Y_k].$$

Another closely related algorithm is the Interacting Multiple Models (IMM) algorithm described in [1].

However, these models suffer the well-known problems associated with the EM algorithm, namely the possibility to be trapped in a local minimum and the slow rate of convergence. To some extent, deterministic annealing [21] and the accelerated algorithms presented in [7], [11] and [12] have been proposed for solving these problems.

In this chapter, we apply a Monte Carlo EM (MCEM) algorithm [23] and we approximate the expectations in the E step using a multimove Gibbs sampler [2]. In a purely Bayesian approach, the Gibbs sampler was previously used for a slightly different model (see [8] pp.237-241). In the present study, the advantage of using the Gibbs sampler is that once the simulated values for the discrete variables $\{S_k\, k = 1,\ldots,T\}$ are obtained, we can apply the Kalman filter and smoother. This allows us to use the Rao-Blackwellised forms in the E step, and also to speed up the algorithm by nesting [22]. Once the system parameters are estimated, we can also solve a classification problem by assigning each observation Y_k to the class $i \in \{1,\ldots,M\}$ corresponding to the maximal probability $P(S_k = j|Y_{(T)})$ $j = 1,\ldots,M$. In the following section, we present the Gibbs sampler for the switching state-space models. The proposed nested Monte Carlo EM (NMCEM) algorithm is the main contribution of this chapter. We then discuss how the theoretical results presented in the previous

chapter can be applied for the model (5.1) - (5.2). At each particular step i of the Gibbs sampler, we need the prediction filter $\{p_k\}$, where $p_1 = \pi$ and

$$p_{k+1}^j = P\left(S_{k+1} = j | Y_k, X_k(i), \ldots, Y_1, X_1(i)\right), \tag{5.4}$$

for any $j = 1, \ldots M$, all $k > 0$ and any sample $X_{(T)}(i)$.

5.2 The Gibbs sampler

We define

$$f(S_{(T)}|Y_{(T)}) := P(S_{(T)}|Y_{(T)}), \tag{5.5}$$

$$g(X_{(T)}|Y_{(T)}) := P(X_{(T)}|Y_{(T)}). \tag{5.6}$$

$$f(S_{(T)}|X_{(T)}, Y_{(T)}) := P(S_{(T)}|X_{(T)}, Y_{(T)}) = P(S_T|X_{(T)}, Y_{(T)}) \prod_{k=1}^{T-1} P(S_k|S_{k+1},$$

$$X_{(k)}, Y_{(k)}), \tag{5.7}$$

$$g(X_{(T)}|S_{(T)}, Y_{(T)}) := P(X_{(T)}|S_{(T)}, Y_{(T)}) = \prod_{m=1}^{M} \Big[P(X_T^m|S_{(T)}, Y_{(T)}) \prod_{k=1}^{T-1} P(X_k^m|$$

$$X_{k+1}^m, S_{(k)}, Y_{(k)}) \Big], \tag{5.8}$$

The probabilities in (5.7) and (5.8) are calculated as in [2] (notice that in (5.8), we have only the Gaussian densities). The probabilities in (5.5) and (5.6) are in fact intractable, since they involve integration with respect to $X_{(T)}$ and $S_{(T)}$, respectively.

Starting with an initial guess $S_{(T)}(0)$, the multimove Gibbs sampler generates a sequence $X_{(T)}(1), S_{(T)}(1), X_{(T)}(2), S_{(T)}(2), \ldots$ such that $S_{(T)}(i)$ is drawn from $f(S_{(T)}|$ $X_{(T)}(i), Y_{(T)})$ and $X_{(T)}(i+1)$ is drawn from $g(X_{(T)}|S_{(T)}(i), Y_{(T)})$. This produces an homogeneous Markov chain $\{S_{(T)}(i), X_{(T)}(i), i > 0\}$, and under appropriate regularity conditions [19], we are eventually sampling from $P(S_{(T)}, X_{(T)}|Y_{(T)})$. We prefer to use a multimove Gibbs sampler, because of its faster convergence ([2]).

For any vectors $y \in \mathbb{R}^d$ and $x = (x^j, j = 1, \ldots, M) \in \mathbb{R}^{Ms}$, let $b(y, x) =$

$(b^j(y,x), j = 1,\ldots,M)$ and $B(y,x) = \mathrm{diag}(b^j(y,x), j = 1,\ldots,M)$, where

$$b^j(y,x) = (2\pi)^{-d/2}\det(R)^{-1/2}\exp(-(y - C_j x^j)^t R^{-1}(y - C_j x^j)/2).$$

The probabilities in (5.7) can be expressed in terms of the transition matrix Φ and the prediction filter $p_k = (p_k^j), j = 1,\ldots M$, given in (5.4):

$$P(S_k = i | S_{k+1} = j, X_{(k)}, Y_{(k)}) = \frac{\Phi^{i,j} b^i(Y_k, X_k) p_k^i}{p_{k+1}^j \sum_{l=1}^M b^l(Y_k, X_k) p_k^l}.$$

Following [2], for the prediction filter, we write

$$p_{k+1} = \frac{\Phi^t B(Y_k, X_k) p_k}{b^t(Y_k, X_k) p_k}, \quad k = 1,\ldots T-1.$$

If $\bullet$ denotes the projective product introduced in [10], then we have

$$p_{k+1} = \Phi^t B(Y_k, X_k) \bullet p_k, \quad k = 1,\ldots T-1. \tag{5.10}$$

The probabilities in (5.8) can be calculated using the Kalman filter as reported in [2]. Here, instead of sampling according to these probabilities, we prefer to use the computationally more efficient simulation smoother presented in [4].

Now we study the convergence of the Gibbs sampler. In this case, strong convergence results can be easily proved if the Markov chain $\{S_{(T)}(i)\}$ is aperiodic and irreducible [19] (for instance, if the transition matrix Φ is positive and $\pi^i > 0$, $i = 1,\ldots,M$). Under these conditions, we can state the following results:

Theorem 5.1 *(i) $\{X_{(T)}(i)\}$ is an ergodic Markov chain with the invariant distribution $g(X_{(T)}|Y_{(T)})$. For any initial value $X^{(0)}$, we have*

$$\int_{\mathbb{R}^{sMT}} \left| g^{(i)}(X_{(T)}|Y_{(T)}) - g(X_{(T)}|Y_{(T)}) \right| dX_{(T)} \leq Cq^i,$$

where $0 \leq q < 1$, $C > 0$, and $g^{(i)}(X_{(T)}|Y_{(T)})$ is the posterior distribution of $X_{(T)}(i)$. (ii) For any real-valued function $h(X_{(T)})$ with $E^g\left[\left|h(X_{(T)})\right| \middle| Y_{(T)}\right] < \infty$ and any

initial density $g^{(0)}(X_{(T)}|Y_{(T)})$ *there exists a constant* C_h *such that*

$$\left|E^{g^{(i)}}\left[h(X_{(T)})|Y_{(T)}\right] - E^{g}\left[h(X_{(T)})|Y_{(T)}\right]\right| \leq C_h q^i.$$

(iii) The Markov chain $\{X_{(T)}(i)\}$ *is geometrically* ϕ *mixing.*

Proof. The proof is very similar to the proof of Theorem 1 presented in [3] or [16]. Here we will outline only the main points.

First, we notice that if $\{S_{(T)}(i)\}$ is aperiodic and irreducible, since it has a finite support, it is uniformly geometrically ergodic and geometrically ϕ mixing.

Secondly, $\{S_{(T)}(i)\}$ and $\{X_{(T)}(i)\}$ are two Markov chains with kernels

$$H(S'_{(T)}, S_{(T)}) = \int_{\mathbb{R}^{sMT}} f(S'_{(T)}|X_{(T)}, Y_{(T)}) g(X_{(T)}|S_{(T)}, Y_{(T)}) dX_{(T)},$$
$$K(X'_{(T)}, X_{(T)}) = \sum_{S_1=1}^{M} \cdots \sum_{S_T=1}^{M} g(X'_{(T)}|S_{(T)}, Y_{(T)}) f(S_{(T)}|X_{(T)}, Y_{(T)}).$$

Hence, $f(S_{(T)}|Y_{(T)})$ and $g(X_{(T)}|Y_{(T)})$ are stationary distributions (in fact they are the unique invariant distributions) for the Markov chains $\{S_{(T)}(i)\}$ and $\{X_{(T)}(i)\}$. The posterior distributions at step i, $f^{(i)}$ and $g^{(i)}$ are derived from the kernels:

$$f^{(i)}(S'_{(T)}|Y_{(T)}) = \sum_{S_1=1}^{M} \cdots \sum_{S_T=1}^{M} H(S'_{(T)}, S_{(T)}) f^{(i-1)}(S_{(T)}|Y_{(T)}),$$
$$g^{(i)}(X'_{(T)}|Y_{(T)}) = \int_{\mathbb{R}^{sMT}} K(X'_{(T)}, X_{(T)}) g^{(i-1)}(X_{(T)}|Y_{(T)}) dX_{(T)}.$$

Finally, the following duality principle holds

$$g^{(i)}(X_{(T)}|Y_{(T)}) = \sum_{S_1=1}^{M} \cdots \sum_{S_T=1}^{M} g(X_{(T)}|S_{(T)}, Y_{(T)}) f^{(i)}(S_{(T)}|Y_{(T)}).$$

The main idea of the proof is to transfer some of the properties of $\{S_{(T)}(i)\}$ to $\{X_{(T)}(i)\}$ using this duality . □

Corollary 5.1 *For any function* h *such that* $E^{g}[|h(X_{(T)})|^2 \Big| Y_{(T)}] < \infty$, *we have the*

following convergence in distribution as $I \to \infty$*:*

$$\frac{1}{\sqrt{I}} \sum_{i=1}^{I} \left(h\left(X_{(T)}(i)\right) - E^g[h(X_{(T)})|Y_{(T)}] \right) \to \mathcal{N}(0, \sigma_h),$$

$$0 < \sigma_h = \mathrm{Var}^g\left(h(X_{(T)})\right) + 2\sum_{i=1}^{\infty} \mathrm{Cov}^g\left(h\left(X_{(T)}(0)\right), h\left(X_{(T)}(i)\right)\right) < \infty.$$

Proof. This is a direct consequence of the geometric ϕ-mixing property of the Markov chain $\{X_{(T)}(i))\}$. □

Remark 5.1 A similar result can be formulated for the Markov chain $\{S_{(T)}(i), X_{(T)}(i))\}$ with the stationary distribution $P(S_{(T)}, X_{(T)}|Y_{(T)}) = g(X_{(T)}|S_{(T)}, Y_{(T)})$ $f(S_{(T)}|Y_{(T)})$.

Remark 5.2 A very important consequence of the theorem is the fact that the ergodic theorem applies, and we can approximate

$$E^g\left[h(X_{(T)})|Y_{(T)}\right] \approx \frac{1}{I}\sum_{i=1}^{I} h\left(X_{(T)}(i)\right),$$

for I sufficiently large and $E^g\left[\left|h(X_{(T)})\right| \middle| Y_{(T)}\right] < \infty$. The central limit theorem in the corollary allows us to monitor the convergence by estimating the variance σ_h.

5.3 The learning algorithm

Let the parameters of the model (5.1) - (5.2) be $\theta = \{R, \Phi, \pi, A_m, C_m, H_m, Q_m, \mu_m, m = 1, \ldots, M\}$. Here, μ_m and Q_m are the mean and the covariance of the hidden Gaussian state variable X_1^m in the state-space model m.

5.3.1 The Monte-Carlo EM algorithm

The EM algorithm is particularly suitable for learning the parameters of the switching SSM (5.1)-(5.2), because it is easier to calculate the likelihood of the augmented data $\{S_{(T)}, X_{(T)}, Y_{(T)}\}$ than the likelihood of the observed data $\{Y_{(T)}\}$ (see [6] for

detailed calculations). The algorithm starts with an initial guess θ_0 of the unknown parameters and iteratively compute the estimation θ^*. Each iteration consists of two steps : the expectation (E) and the maximization (M) step.

Giving the current estimation θ_n of the parameters, the conditional expectation of the augmented data log-likelihood is computed in the E-step:

$$Q(\theta|\theta_n) = E[\log P(S_{(T)}, X_{(T)}, Y_{(T)}|\theta)|Y_{(T)}, \theta_n]. \tag{5.11}$$

For the mixture structure of the switching SSMs, any attempt to calculate the exact expectation in (5.11) becomes impractical [6]. To solve this problem, we propose an MCEM algorithm [23]. Instead of an exact E-step, we use an approximation

$$Q(\theta|\theta_n) \approx \frac{1}{L_n}\sum_{l=1}^{L_n} \log P\left(\{S_{(T)}(l), X_{(T)}(l), Y_{(T)}\}|\theta\right),$$

where $\{S_{(T)}(l), X_{(T)}(l)\}$, $l = 1, \ldots, L_n$ is a sample from $P\big(S_{(T)}, X_{(T)}, |Y_{(T)}, \theta_n\big)$.

The M-step performs a maximization with respect to the parameters θ:

$$\theta_{n+1} = \arg\max_{\theta} Q(\theta|\theta_n).$$

In our case, we maximize $Q(\theta|\theta_n)$ by setting

$$A_m(n+1) = \sum_{k=2}^{T} E\left[X_k^m \left(X_{k-1}^m\right)^t |Y_{(T)}, \theta_n\right] \left(\sum_{k=2}^{T} E\big[X_{k-1}^m \left(X_{k-1}^m\right)^t | Y_{(T)}, \theta_n\big]\right)^{-1}, \tag{5.14}$$

$$C_m(n+1) = \sum_{k=1}^{T} Y_k E\left[1_{\{m\}}(S_k)(X_k^m)^t|Y_{(T)}, \theta_n\right] \left(\sum_{k=1}^{T} E\big[1_{\{m\}}(S_k)X_k^m (X_k^m)^t \big|Y_{(T)}, \theta_n\big]\right)^{-1}, \tag{5.15}$$

$$Q_1^m(n+1) = E\left[X_1^m (X_1^m)^t |Y_{(T)}, \theta_n\right] - E\left[X_1^m|Y_{(T)}, \theta_n\right] E[(X_1^m)^t | Y_{(T)}, \theta_n], \tag{5.16}$$

$$H_m(n+1) = \frac{1}{T-1}\Big(\sum_{k=2}^{T} E\left[X_k^m (X_k^m)^t \,|Y_{(T)}, \theta_n\right] - A_m(n+1) \times \sum_{k=2}^{T} E\left[X_{k-1}^m (X_k^m)^t \,|Y_{(T)}, \theta_n\right]\Big), \tag{5.17}$$

$$R(n+1) = \frac{1}{T}\Big(\sum_{k=1}^{T} Y_k Y_k^t - \sum_{m=1}^{M} C_m(n+1) \sum_{k=1}^{T} E\left[1_{\{m\}}(S_k) X_k^m | Y_{(T)}, \theta_n\right] Y_k^t\Big), \tag{5.18}$$

$$\mu_{X_1}^m(n+1) = E\left[X_1^m | Y_{(T)}, \theta_n\right], \tag{5.19}$$

$$\pi^m(n+1) = P\left(S_1 = m | Y_{(T)}, \theta_n\right), \tag{5.20}$$

$$\Phi^{(q,m)}(n+1) = \left(\sum_{k=2}^{T} P\left(S_k = m, S_{k-1} = q | Y_{(T)}, \theta_n\right)\right) / \left(\sum_{k=1}^{T-1} P\left(S_k = q | Y_{(T)}, \theta_n\right)\right). \tag{5.21}$$

5.3.2 The nested MCEM algorithm

In each cycle of the MCEM, we repeatedly apply random generators. Hence, compared to any deterministic implementation of the EM algorithm, the MCEM is time consuming. Thus, any technique which is capable to improve the rate of convergence is very valuable. Here we propose a nested MCEM algorithm [22].

We consider two different augmented-data sets: $Y_{aug1} = \{Y_{(T)}, S_{(T)}\}$ and $Y_{aug2} = \{Y_{(T)}, S_{(T)}, X_{(T)}\}$. For acceleration, the nested MCEM algorithm will fix the augmented data Y_{aug1} and run several EM iterations conditional on these values. Notice that remark 5.2 allows us to use the following Rao-Blackwellised forms to estimating the expectations in (5.14)-(5.19):

$$E\left[1_{\{m\}}(S_k) X_k^m | Y_{(T)}\right] \approx \sum_{i=1}^{L_n} E\left[1_{\{m\}}(S_k(i)) X_k^m | S_{(T)}(i), Y_{(T)}\right] / L_n, \tag{5.22}$$

$$E\left[1_{\{m\}}(S_k) X_k^m (X_k^m)^t | Y_{(T)}\right] \approx \sum_{i=1}^{L_n} E\Big[1_{\{m\}}(S_k(i)) X_k^m (X_k^m)^t | S_{(T)}(i), Y_{(T)}\Big] / L_n, \tag{5.23}$$

$$E\left[X_k^m \left(X_k^m\right)^t | Y_{(T)}\right] \approx \sum_{i=1}^{L_n} E\left[X_k^m \left(X_k^m\right)^t | S_{(T)}(i), Y_{(T)}\right] / L_n, \tag{5.24}$$

$$E\left[X_k^m \left(X_{k-1}^m\right)^t | Y_{(T)}\right] \approx \sum_{i=1}^{L_n} E\left[X_k^m \left(X_{k-1}^m\right)^t | S_{(T)}(i), Y_{(T)}\right] / L_n. \tag{5.25}$$

We also have

$$P\left(S_k = m | Y_{(T)}\right) \approx \sum_{i=1}^{L_n} 1_{\{m\}}\left(S_k(i)\right) / L_n, \tag{5.26}$$

$$P\left(S_k = m, S_{k-1} = q | Y_{(T)}\right) \approx \sum_{i=1}^{L_n} 1_{\{m\}}\left(S_k(i)\right) 1_{\{q\}}\left(S_{k-1}(i)\right) / L_n. \tag{5.27}$$

Hence, the E-step for this inner EM algorithm is based only on the Kalman smoother and the previous Rao-Blackwellised forms. As a consequence, it is computationally more efficient. A formal proof of the faster rate of convergence for the nested EM is given in [22]. Now we present the algorithm in detail:

Initialize the parameters of the model.

Repeat until log likelihood has converged:

Draw L_n sets of values $S_{(T)}(i)$, $X_{(T)}(i)$ for $\{S_{(T)}, X_{(T)}\}$ using the Gibbs sampler;

Calculate the conditional probabilities using formulas (5.26)-(5.27);

For $t = 1$ to $K_{nesting}$ do

Inner E-step

Run Kalman smoothing recursions given the sample $\{S_{(T)}(i), i = 1, \ldots, L_n\}$;

Replace in the Rao-Blackwellised forms (5.22)-(5.25);

Inner M-step

Re-estimate the parameters of the model using the formulas (5.14)-(5.21).

Next, we discuss some issues concerning the implementation of the algorithm. Following the suggestion given in [23], we increase the number L_n of data in the Gibbs sampler at each new cycle of the MCEM algorithm. For the number of inner

steps, we have tried several values ($K_{nesting} = 3, 5, 7$) with comparable results. Since the log likelihood is intractable, we apply a harmonic average to estimate its values [14], and we apply the bridge sampling presented in [22] to monitor the convergence.

5.4 Asymptotic properties of the prediction filter

In the NMCEM algorithm, we apply the Gibbs sampler in every cycle of the EM algorithm. Since we do not know the real values of the parameters, we have to work with the estimates corresponding to each cycle of the EM algorithm. In this section, we consider a reduced set of parameters $\theta = \{\Phi, \pi\}$, and we apply the results obtained in Chapter 4 to study the impact due to the inaccuracies in the estimates of θ for a large number T of observations.

For every positive integer i and for every set of values for $X_{(T)}(i)$, the values of $S_{(T)}(i)$ are drawn from $f(S_{(T)}|X_{(T)}(i), Y_{(T)})$ given in (5.7). Since we are working with a fixed set of values $X_{(T)}(i)$, we omit the index i in order to simplify the notations. As we have already mentioned, the probabilities in (5.7) can be expressed in terms of the transition matrix Φ and the prediction filter $p_n = (p_n^j), j = 1, \dots M$, given in (5.4). The exact values of the parameters θ are in fact unknown, so that instead of (5.10), we consider an equation corresponding to the values $\theta(n)$ at the $n-$ step of the EM algorithm:

$$p_{k+1}(n) = \Phi^t(n) B(Y_k, X_k) \bullet p_k(n),$$

for any $k = 1, \dots T-1$, and $p_1(n) = \pi(n)$. To make explicit the dependency with respect to the initial condition and the observations, we denote

$$p_{k+1}(n) = f_n[Y_k, X_k, \dots, Y_l, X_l, p_l(n)],$$

for any positive integers k, l, such that $k \geq l$.

ASSUMPTION I. The Matrix Φ is primitive.

ASSUMPTION II. The matrix $A = \text{diag}(A_m)$, $m = 1, \dots, M$, has only eigenvalues σ with $|\sigma| < 1$.

As a consequence of the first assumption, the Markov chain $\{S_k\}$ is geometrically ergodic with a unique invariant probability distribution $\mu = (\mu^i)$, $i = 1, \ldots, M$. Under the second assumption, we study the properties of the Markov chain $\{X_k = (X_k^1, \ldots, X_k^M), k \geq 0\}$. From the normality and independence assumptions, we can easily prove that for any positive integer $k \geq 1$ and any $x_0 \in \mathbb{R}^{sM}$

$$P_X^k(x_0, \cdot) = P(X_k | X_0 = x) = N(A^k x_0, H_{(k)}), \tag{5.29}$$

where $H_{(k)} = H + AHA^t + \ldots + A^{k-1}H(A^{k-1})^t = H + F_k$.

Furthermore, Assumption II implies that the zero solution of the deterministic linear system $x_{k+1} = Ax_k$ is asymptotically stable (Theorem 4.9, in [5]). Since the system is autonomous, the zero solution is also exponentially stable (Corollary 4.6, in [5]). Moreover, by Theorem 4.3, page 128 in [20] the Markov chain $\{X_k, k \geq 0\}$ is geometrically ergodic with a unique invariant probability distribution ρ. From Theorem 4.5, in [5], there exists positive constants a_2 and $c_2 < 1$, such that, for all $k \geq 1$,

$$\|A^k\|_2 \leq a_2 c_2^k. \tag{5.30}$$

Lemma 5.1 *For any positive integer l and any normal distributions $N(m_1, Z)$ and $N(m_2, Z)$ on $(\mathbb{R}^l, \mathcal{B}(\mathbb{R}^l))$, with $m_1 \neq m_2$, and the same symmetric and positive definite covariance matrix Z, we have*

$$\|N(m_1, Z) - N(m_2, Z)\| \leq \rho(Z^{-1})^{1/2} \|m_1 - m_2\|.$$

Consequently, for a_2 and c_2 as given in (5.30) and any $x_0, x_0' \in \mathbb{R}^{sM}$, for the $k-$step transition probabilities $P_X^k(x_0, \cdot)$ defined in (5.29), we get

$$\|P_X^k(x_0, \cdot) - P_X^k(x_0', \cdot)\| \leq \sqrt{\rho(H^{-1})} a_2 c_2^k \|x_0 - x_0'\|. \tag{5.31}$$

PROOF. Let $\nu = N(m_1, Z) - N(m_2, Z)$ and

$$Q_1(x) = (2\pi)^{-l/2} \det(Z)^{-1/2} \exp\left(-(x - m_1)^t Z^{-1}(x - m_1)/2\right),$$

$$Q_2(x) = (2\pi)^{-l/2} \det(Z)^{-1/2} \exp\left(-(x-m_2)^t Z^{-1}(x-m_2)/2\right).$$

Proceedings as in [9], page 47, from the Schwartz inequality, we get

$$\begin{aligned}
\|\nu\| &= \int_{\mathbb{R}^l} |Q_1(x) - Q_2(x)| dx = \int_{\mathbb{R}^l} \left|\sqrt{Q_1(x)} - \sqrt{Q_2(x)}\right| \left(\sqrt{Q_1(x)} + \sqrt{Q_2(x)}\right) dx \\
&\leq \left(\int_{\mathbb{R}^l} \left(\sqrt{Q_1(x)} - \sqrt{Q_2(x)}\right)^2 dx\right)^{1/2} \left(\int_{\mathbb{R}^n} \left(\sqrt{Q_1(x)} + \sqrt{Q_2(x)}\right)^2 dx\right)^{1/2} \\
&= \left(2 - 2\int_{\mathbb{R}^l} \sqrt{Q_1(x)Q_2(x)} dx\right)^{1/2} \left(2 + 2\int_{\mathbb{R}^l} \sqrt{Q_1(x)Q_2(x)} dx\right)^{1/2}
\end{aligned}$$

Since

$$\int_{\mathbb{R}^l} \sqrt{Q_1(x)Q_2(x)} dx = \exp\left(-(m_1-m_2)^t Z^{-1}(m_1-m_2)/8\right),$$

we obtain

$$\|\nu\| \leq \left((m_1-m_2)^t Z^{-1}(m_1-m_2)\right)^{1/2}.$$

Let $Z^{-1} = Z_1^t Z_1$ be the Cholesky decomposition of the symmetric and positive definite matrix Z^{-1}. Thus $\|\nu\| \leq \|Z_1\|_2 \|m_1 - m_2\| = \rho(Z^{-1})^{1/2} \|m_1 - m_2\|$.

To prove the second part of the lemma, notice that for any positive integer k, $H_{(k)} = H + F_k$, with F_k being symmetric and non-negative definite. Hence, $\rho(H^{-1})I_k - H_{(k)}^{-1}$ is also symmetric and non-negative definite. As a consequence, $\rho(H_{(k)}^{-1}) \leq \rho(H^{-1})$. The inequality (5.31) is thus a consequence of the first part of the lemma and (5.30). □

Notice that from the previous lemma, it is obvious that assumptions I and II can replace assumptions A' and B in the previous chapter. Moreover, assumption II ensures the asymptotic exponential stability of the Kalman filters applied for drawing the values of $X_{(T)}(i)$ at each step i of the Gibbs sampler.

Let denote by $\lambda_k(\cdot)$ the Lebesgue measure on $\mathbb{R}^k$, $k \geq 1$. For any vectors $x_1, x_2, x \in \mathbb{R}^{Ms}$, $y \in \mathbb{R}^d$ we define:

$$a(x_2, x_1) = (2\pi)^{-sM/2} \prod_{j=1}^{M} \det(H_j)^{-1/2} \exp\left(-\left(x_2^j - A_j x_1^j\right)^t H_j^{-1} \left(x_2^j - A_j x_1^j\right)/2\right),$$

$$\delta(y,x) = \frac{\max\limits_{j=1,\ldots,M} b^j(y,x)}{\min\limits_{j=1,\ldots,M} b^j(y,x)}, \qquad \epsilon(n) = \min^{+}_{i,j=1,\ldots,M} \Phi^{i,j}(n),$$

$$\Delta_{-1}(x) = \min_{i=1,\ldots M} \int_{\mathbb{R}^d} \delta^{-1}(y,x) b^i(y,x) \lambda_d(dy),$$

$$R(n) = \epsilon^r(n) \int_{\mathbb{R}^{sM}} \cdots \int_{\mathbb{R}^{sM}} \prod_{k=2}^{r} \Delta_{-1}(x_k) a(x_k, x_{k-1})\, \lambda_{Ms}(dx_r) \ldots \lambda_{Ms}(dx_2) \rho(dx_1).$$

As a consequence of Theorem 4.4 in chapter 4, we can prove that the prediction filter has an exponential rate of forgetting the initial conditions.

Theorem 5.2 *Under the assumptions I and II, if the stochastic matrix $\Phi(n)$ is primitive with index of primitivity r, then for any positive integer l and any vectors p_1, $p_2 \in \mathbb{R}^M$ with positive entries that sum to one, we have*

$$\limsup_{k\to\infty} \frac{1}{k} \log \| f_n[Y_k, X_k, \ldots, Y_l, X_l, p_1] - f_n[Y_k, X_k, \ldots, Y_l, X_l, p_2] \|_1$$
$$\leq \frac{1}{r} \log(1 - R(n)) < 0, , \quad Pa.s..$$

Moreover, from Corollary 4.2, we have the following result.

Theorem 5.3 *Under assumptions I and II, if the matrix $\Phi(n)$ is positive, then the Markov chain $\{S_k, X_k, Y_k, p_k(n)\}$ has a unique invariant probability distribution* $\mathbf{m}$.

5.5 Applications

In order to asses the effectiveness of the proposed NMCEM, we compare our approach with the algorithms presented in [6] and [8].

5.5.1 Simulated data

First, we suppose that the model parameters are known and we apply the Gibbs sampler to solve a classification problem. As in [6], the SSMs and the switching

process are defined as

$$X_k^1 = 0.99X_{k-1}^1 + w_k^1, \quad w_k^1 \sim N(0,1),$$
$$X_k^2 = 0.9X_{k-1}^2 + w_k^2, \quad w_k^{(2)} \sim N(0,10),$$
$$Y_k = X_k^m + v_k, \quad v_k \sim N(0,0.1),$$

where the initial probabilities are $\pi^{(1)} = \pi^{(2)} = 0.5$, and the transition probabilities are $\Phi^{(1,1)} = \Phi^{(2,2)} = 0.95$ and $\Phi^{(1,2)} = \Phi^{(2,1)} = 0.05$. We generate 200 sequences of length 200 from this model. A set of data and its true segmentation are displayed in Fig. 5.1.

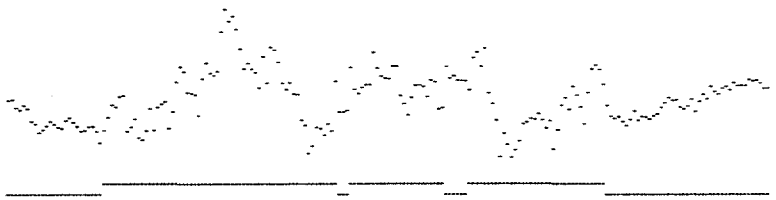

Figure 5.1: Data sequence of length 200 with its true segmentation below it: the upper dots represent the switch state 2, and the lower dots the switch state 1

For each sequence, we run the Gibbs sampler described in Section 5.2. A point Y_t is considered to be from class 1 or class 2 according to the values of the probabilities $P(S_t = m|Y_1, \ldots, Y_T)$, $m = 1, 2$. These conditional probabilities are estimated using (5.26). The histogram of the percent correct segmentations is displayed in Fig. 5.2(a). Compared with the results reported in [6], even with the annealed algorithm, we notice an significant increase of the percent correct segmentations. However, the Gibbs sampler is much slower than the variational approach and we could not use exactly the same data as in [6].

Now, for the same simulated data sets, we apply two merging methods: the IMM procedure, and the GPB algorithm. For the IMM approach, we use only the forward part of the algorithm and we classify according to $P(S_t = m|Y_1, \ldots, Y_t)$, $m = 1, 2$. The results are displayed in Fig. 5.2(b), and they are similar to the histogram presented in [6]. The average performance of the method based on the Gibbs sample is about 10% better than the IMM method. For the GPB approach,

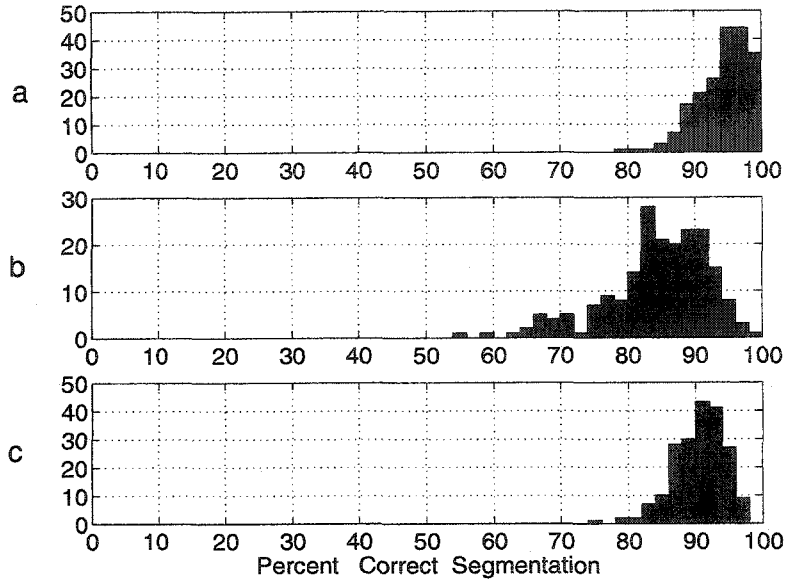

Figure 5.2: Histograms of percent correct segmentations: (a) inference based on the Gibbs sample (b) Gaussian merging, IMM (c) Gaussian merging with approximate smoothing

we approximate the conditional probabilities $P(S_t = m|Y_1, \ldots, Y_T)$, $m = 1, 2$ using the merging method and the smoother presented in [8] pp. 106-109. The results are shown in Fig. 5.2(c), and they are obviously better than the ones displayed in Fig. 5.2(b). However, the average of the percentage of the correct segmentations based on the Gibbs sampler is still about 3.8% greater compared to the results shown in Fig. 5.2(c).

We also tested other switching SSMs, and we have consistently obtained better results with the algorithm based on the Gibbs sampler than with the algorithm based on merging. Here we present the results obtained for the following SSM with $M = 3$:

$$
\begin{aligned}
X_k^1 &= 0.99X_{k-1}^1 + w_k^1, \quad w_k^1 \sim N(0,1), \\
X_k^2 &= 0.95X_{k-1}^2 + w_k^2, \quad w_k^2 \sim N(0,5), \\
X_k^3 &= 0.9X_{k-1}^3 + w_k^3, \quad w_k^3 \sim N(0,10), \\
Y_k &= X_k^m + v_k, \quad v_k \sim N(0,0.1).
\end{aligned}
$$

The initial probabilities are $\pi^{(1)} = \pi^{(2)} = \pi^{(3)} = 1/3$ and the transition probabilities are $\Phi^{(1,1)} = \Phi^{(2,2)} = \Phi^{(3,3)} = 0.95$ and $\Phi^{(1,2)} = \Phi^{(1,3)} = \Phi^{(2,1)} = \Phi^{(2,3)} = \Phi^{(3,1)} = \Phi^{(3,2)} = 0.025$. In Fig. 5.3, we display a set of data and its true segmentation.

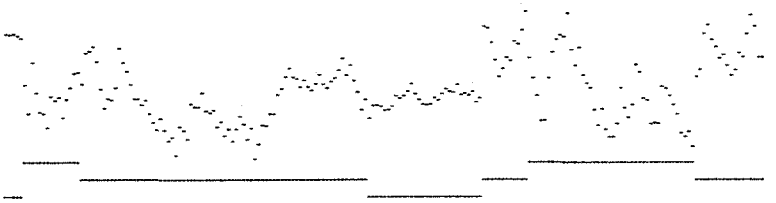

Figure 5.3: Data sequence of length 200 with its true segmentation below it: the upper, middle and lower dots represent the switch states 3, 2, 1, respectively

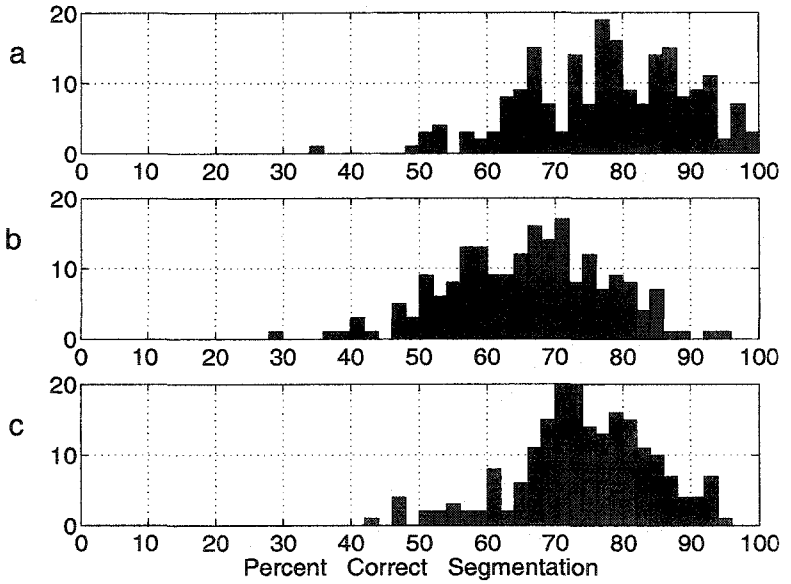

Figure 5.4: Histograms of percent correct segmentations: (a) inference based on the Gibbs sample (b) IMM (c) GPB with approximate smoothing

We generate 200 sequences of length 200 from this model. The histograms corresponding to the method based on the Gibbs sample, the IMM algorithm and the GPB algorithm are reported in Figs. 5.4(a), 5.4(b) and 5.4(c), respectively. For this classification problem, the results obtained with the method based on the Gibbs sample are in average about 3.2% better than the GPB algorithm and 11.5% better than the IMM algorithm.

5.5.2 Experimental medical data

We now consider the experimental physiological data displayed in Fig. 5.5(a), and we apply the NMCEM for system identification. This data correspond to a patient tentatively diagnosticated with sleeping apnea[1]. The respiration pattern is characterized by a succession of no breathing, gasping breathing and normal rhythmic breathing (see also [6] and the references therein).

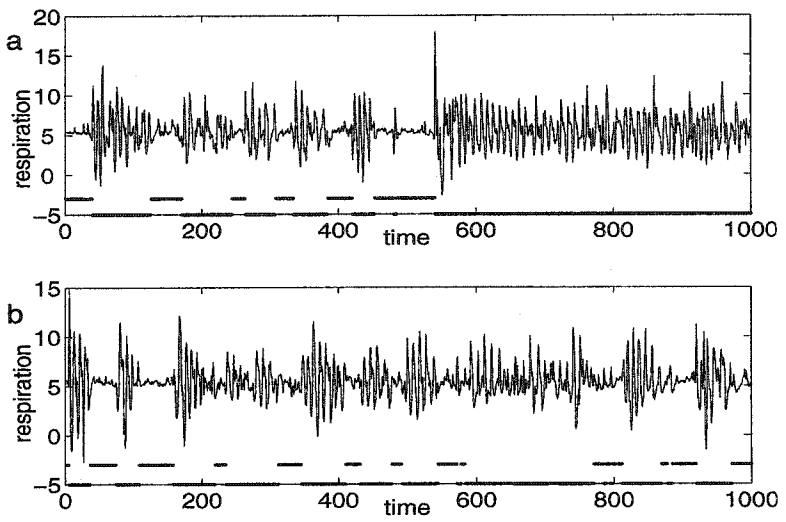

Figure 5.5: Chest volume of a patient with sleep apnea. (a) Training data (b) Test set

It is shown in [6] that a switching SSM with $M = 2$ components and the dimension of the state-space $K = 2$ is the most suitable model for this data set. One component is specialized for the gasping and normal breathing, and the other component models the data during periods of apnea. We use the same type of switching SSM and we apply the NMCEM for estimating the parameters. The results are very similar to those reported in [6]. The segmentation found is shown at the bottom of Fig. 5.5(a).

In Fig. 5.6, we compare the MCEM with the NMCEM for $K_{nested} = 3$. By introducing a nesting approach, we reduce the time at least by a factor of two. Since

[1]Data are available online at http://www.physionet.org/physiobank/database/santa-fe.

the likelihood is intractable, we can find only an estimation of the log likelihood for each cycle of the NMCEM. The final estimated value for the maximum likelihood is similar to the value of the lower bound obtained in [6]. .

The results for the test set are displayed in Fig. 5.5(b). Using the MCEM or the nested versions, we obtain estimations for the maximum log likelihood lower than the value of the bound reported in [6]. Our best values are around -0.75 nats per observation compared to -0.85 nats per observation reported in [6]. Also, the segmentation seems to be more accurate (see Fig. 5.5(b)), since it is capable of detecting possible apnea periods which were not noticed in [6].

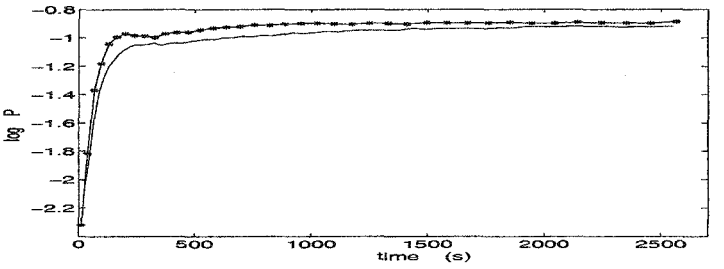

Figure 5.6: Learning curves for a MCEM (solid line) and a NMCEM (dotted line)

Bibliography

[1] Y. Bar-Shalom and X.-R. Li. *Estimation and Tracking.* Artech House, Boston, MA., 1993.

[2] C.K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81:541–553, 1994.

[3] J. Diebolt and C.P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *J.R. Statist. Soc. B*, 56(2):363–375, 1994.

[4] J. Durbin and S.J. Koopman. A simple and efficient simulation smoother for state-space time series analysis. *Biometrika*, 89(3):603–615, 2002.

[5] S.N. Elaydi. *An Introduction to Difference Equations.* Springer Verlag, New York, 1996.

[6] Z. Ghahramani and G.E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–996, 2000.

[7] M. Jamshidian and R.I. Jennrich. Conjugate gradient acceleration of the EM algorithm. *J. Amer. Stat. Asoc.*, 88(421):221–228, 1993.

[8] C.J. Kim and C.R. Nelson. *State Space Models with Regime Switching.* The MIT Press, 1999.

[9] L.M. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York, 1986.

[10] F. Le Gland and L. Mevel. Basic properties of the projective product, with application to products of column-allowable non-negative matrices. *Math. Control Signals Systems*, 13:41–62, 2000.

[11] I. Meilijson. A fast improvement of the EM algorithm in its own terms. *J.R. Statist. Soc. B*, 51(1):127–138, 1989.

[12] X.L. Meng and D. Van Dyk. The EM algorithm - an old folk-song to a fast new tune. *J.R. Statist. Soc.B*, 59(3):511–567, 1997.

[13] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, an other variants. In M.I. Jordan, editor, *Learning in Graphical Models*. 1998.

[14] M.A. Newton and A.E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *J.R. Statist. Soc. B*, 56(1):3–48, 1994.

[15] C. A. Popescu and Y. S. Wong. Monte Carlo approach for switching state-space models. In C. Orchard, B. Yang and M. Ali, editors, *Proceedings of the 17th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems*, volume 3029 of *Lecture Notes in Artificial Intelligence*, pages 945–955, Ottawa, 2004. Springer-Verlag.

[16] G. Robert, C.P. Celeux and J. Diebolt. Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statistics and Probability Letters*, 16:77–83, 1993.

[17] R.H. Shumway and D.S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.*, 3(4):253–264, 1982.

[18] R.H. Shumway and D.S. Stoffer. Dynamic linear models with switching. *J. Amer. Stat. Asoc.*, 86:763–769, 1991.

[19] L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–17028, 1994.

[20] H. Tong. *Nonlinear Time Series.* The Clarendon Press Oxford University Press, New York, 1990.

[21] N. Ueda and R. Nakano. Deterministic annealing variant of the EM algorithm. In D. Tesauro, G. Touretzky and J. Alspector, editors, *Advances in neural information processing systems*, pages 545–552. Morgan Kaufmann, 1995.

[22] D.A. Van Dyk. Nesting EM algorithms for computational efficiency. *Statist. Sinica*, 10:203–225, 2000.

[23] G.C.G. Wei and M.A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor men's data augmentation algorithm. *J. Amer. Stat. Asoc.*, 85:699–704, 1990.

Chapter 6

Conclusions

The work presented in this thesis is common to many fields such as time series, pattern recognition, machine learning, signal processing and system identification. The main contributions are the new approaches for predicting the nonlinear aeroelastic response, the study of the ergodic properties of the models with a Markov switching, and the improved stochastic algorithm for switching state-space models.

The illustrative case studies reported in Chapters 2 and 3 clearly demonstrate that the proposed approaches, based on nonlinear time series and the unscented filter, are capable of accurately predicting the limit cycle, damped, and unstable aeroelastic oscillations.

For the numerical simulated and the experimental aeroelastic data with polynomial restoring forces, the amplitude dependent exponential autoregressive models seems to be very appropriate. The nonlinear time series model provides long term predictions of the same accuracy or sometimes even more accurate than the unscented filter.

For a freeplay model, the predictions obtained using the unscented filter are more accurate than those obtained with nonlinear time series models. The unscented filter also performs more effective de-noising than the wavelet filters. However, the wavelet filters and the time series models do not require any information about the structure of the dynamical system associated with the input data. To implement the unscented filter method, even if no information about the system parameters is

given, the specific type of the nonlinearities associated with the aeroelastic system must be known. Thus, the extra information seems to be the reason leading to a better performance.

A method capable of accurately determining the freeplay structure is of great interest in nonlinear aeroelasticity. In Section 2.3, we propose a non-parametric method for finding the threshold structure for an aeroelastic model with a freeplay. As for the future work, we plan to test this method for experimental data.

The nonlinear time series models and the unscented filter can be used as the key components in the expert data mining system presented in Fig. 1.1. Other methods, such as neural networks can be easily incorporated in the processing module of the system. The present investigation also opens up the opportunity for other applications in flight dynamics and active vibration control systems.

In Chapters 4 and 5, we study models with Markov switching. We are mainly interested on switching state-space models, but we present also an extension of a classical result concerning hidden Markov models.

A nested Monte Carlo EM algorithm, based on the Gibbs sampler, for learning the parameters of switching state-space models is developed in Chapter 5. The proposed new algorithm is a good alternative when the execution time is not critical and high accuracy is required. Under very reasonable conditions, we show that convergence properties for the Gibbs sampler can be proved. The ergodic properties established in Chapter 4 allow us to study the effects of the possible misspecifications of some of the model parameters, and to find sufficient conditions that ensure strong asymptotic results. The efficiency of the new nested Monte Carlo EM algorithm can be improved by using parallel computation.

The new approach is compared with three existing alternative methods, all of them based on the EM algorithm but not involving random number generators. Although our method is slower than the others, the results reported here for the simulated and the experimental data are more accurate. We have studied several sets of data corresponding to different switching state-space models, and the proposed stochastic approach consistently provides a better classification when the parameter

of the system are known. For system identifications, more tests need to be done, since the performance of the EM algorithm depends on the initial guess and better results can be only obtained as a consequence of a more inspired initialization algorithm. The approach proposed here can also be applied to economical models with a Markov switching.