Non-Linear Models for Time Series Using Mixtures of Autoregressive Models

Assaf Zeevi* Ron Meir† Robert J. Adler‡

First version: 1998; Last revised: October 2000

Abstract

We consider a novel class of non-linear models for time series analysis based on *mixtures of local autoregressive models*, which we call MixAR models. MixAR models are constructed so that at any given point time, one of a number of alternative AR models describes its dynamics. The driving AR model is randomly selected from the set of m possible models via according to a state (lag vector) dependent probability distribution. Thus, the MixAR process is a Markov chain with a transition kernel that takes the form of a mixture distribution with non-constant (state dependent) weights. This structure gives MixAR models considerable flexibility, as will be indicated both theoretically and via example. The theoretical aspects of MixAR models that we examine include stochastic stability of MixAR processes, parameter estimation algorithms, and approximation of quite general underlying prediction functions, when the true process is not of the MixAR family. We complement this study with some numerical examples, which seem to indicate that the out-of-sample performance is competitive, in spite of the fairly large number of parameters in MixAR models. Prediction results on benchmark time series are compared to linear and non-linear models.

KEY WORDS: Mixtures of autoregressions, non-linear models, prediction, mixtures of distributions, EM algorithm, Markov chains, stochastic stability.

1 Introduction

The general framework of linear autoregressive moving average (ARMA) models has dominated much of the research in time series analysis in the statistics and engineering communities. However, the need to model non-linear features of processes commonly encountered in nature, as well as utilize the models for forecasting purposes, have led to increasing interest in non-linear modeling techniques. We will review some of these below, in Section 3.

*Corresponding author: Information Systems Lab, Stanford University, Stanford CA. 94305; e-mail: assaf@isl.stanford.edu

†Faculty of Electrical Engineering, Technion, Haifa 32000, Israel. e-mail: rmeir@dumbo.technion.ac.il Research supported in part by a grant from the Israel Science Foundation. Support from the Ollendorff center of the Faculty of EE at the Technion is also acknowledged.

‡Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel. e-mail: robert@ieadler.technion.ac.il Research supported in part by NSF grant 9625753, and MSRI, Berkeley, which provided peace and quiet, and time to think.

In this work we introduce a new class of models, substantially extending the classic AR models, yet retaining much of their structural simplicity. The main idea is, at first sight, not that different from the threshold autoregressive (TAR) models of Tong [see, e.g., Tong (1983, 1990) for an overview]. The key assumption is that locally in the state space the time series follows a linear model. The simplest version of this structure is an AR model. Tong, and later others, developed the idea that the state space can be partitioned by comparing a lagged variable to a fixed threshold, or to several thresholds. Each region has an associated AR model, which determines the dynamics of the process while it remains in that region. Thus, conditional on the value of the lagged variable (or variables), the process follows the corresponding AR model. Our approach is somewhat different.

Let $(X_t)_{t\geq 0}$ denote a real valued time series, i.e., a discrete time stochastic process. Fix $d \in \mathbb{N}$, and write X_{t-d}^{t-1} for the lag vector $(X_{t-1}, \ldots, X_{t-d})$. Throughout we use capital letters to denote random variables, while Corresponding realizations are written using lower case, e.g., x_t. Vector valued quantities are distinguished from scalars using boldface notation. Without further background and definitions, we proceed to give an informal presentation of the *mixture of autoregressions* (MixAR) model in a way that is meant to clarify its structure. A more rigorous definition will be given in Section 2. In the sequel, d denotes the dimension of the lag-vector, and m is used to denote the number of AR components in a given MixAR model. Then, with obvious terminology, we say that a time series follows a MixAR$(m; d)$ process if

$$X_t = \begin{cases} \boldsymbol{\theta}_1^T X_{t-d}^{t-1} + \theta_{1,0} + \varepsilon_t^{(1)} & \text{w.p. } g_1(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \\ \boldsymbol{\theta}_2^T X_{t-d}^{t-1} + \theta_{2,0} + \varepsilon_t^{(2)} & \text{w.p. } g_2(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \\ \quad \vdots & \quad \vdots \\ \boldsymbol{\theta}_m^T X_{t-d}^{t-1} + \theta_{m,0} + \varepsilon_t^{(m)} & \text{w.p. } g_m(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \end{cases} \tag{1}$$

where $(\varepsilon_t^{(j)})_{t\geq 0}$ $j = 1, 2, \ldots, m$ are mutually independent i.i.d., zero mean, Gaussian noise processes with (possibly) different variances σ_j^2. The $g_j(\cdot; \boldsymbol{\theta}_g)$ are chosen so that $\sum_j g_j(\cdot; \boldsymbol{\theta}_g) \equiv 1$, and $g_j(\cdot; \boldsymbol{\theta}_g) \geq 0$, so that they can be interpreted as a probability distribution over the m component AR models.

Implicit in (1) is a randomization process at each time step, which determines the governing AR equation for the process at the next time step. It is this aspect of the model that is most novel, generates its non-linearity, and gives rise to most of its interesting properties, while at the same time allowing it to remain theoretically and computationally amenable. The precise functional form of the g_j's will be specified in the sequel.

The main ideas underlying this model were first introduced in the neural network literature by Jacobs *et al.* (1991) [see also Jordan and Jacobs (1994) for related work] in the context of a regression problem. The model used there was termed *mixtures of experts*. Some numerical results were obtained by Waterhouse and Robinson (1996) for the prediction of acoustic vectors, and by Peng *et al.* (1996) for speech recognition. Both use a variant of mixtures of experts known as a hierarchical mixture of experts. Recently, some theoretical aspects of the mixtures of experts in the context of non-linear regression were studied by Zeevi *et al.* (1998).

The main contributions of this paper are the following:

1. The derivation of conditions that ensure stochastic stability (i.e., existence and uniqueness of a stationary version, and rates of convergence to stationarity) for MixAR processes; see Theorems 5.1 and 5.2.

2. Some discussion of the generality of MixAR processes, from the point of view of approximating arbitrary prediction functions; see Theorem 6.1.

3. The study of a computationally efficient parameter estimation scheme via a generalized expectation–maximization algorithm, and conditions for convergence of this algorithm; see Algorithm 7.1 and Proposition 7.1.

4. A numerical study that validates the effectiveness of MixAR models in practice; see, e.g., Tables 1 and 2. This study highlights the benefits of retaining local linear structure both in terms of interpretation as well as model specification; see the discussion in Section 8.2 and Figures 4 and 5.

We also include a fairly broad discussion of where this model fits into the existing literature, and indicate in which ways it fundamentally differs from existing non-linear models. Since MixAR models, by construction, involve a large number of parameters, we add a discussion on the issue 'parameters and parsimony'. The purpose of this discussion is to indicate, mainly via reference to recent literature, why MixAR out-of-sample performance is not impaired by apparent over-parameterization. We do not have a complete explanation at this point, and consider this important topic a main avenue for further investigation.

Here is a road map to the various sections in the paper. Section 1 is devoted mainly to presenting the MixAR model in detail, while Section 3 discusses its relationship with other non-linear models in the literature. Section 4 discusses the issue of apparent 'over-parameterization' in MixAR models. Stochastic stability of MixAR processes is discussed in Section 5. The issue of generality, by which we mean the ability to approximate a wide class of prediction functions, is discussed in Section 6. In Section 7 we present an efficient algorithm for parameter estimation, and briefly study some of its properties. Numerical results are given in Section 8, and Section 9 wraps up the main part of the paper discussing the main significance of the results. All proofs are relegated to Appendix A, while Appendix B contains the specifics of the various models used in the numerical study.

Acknowledgement: We are extremely grateful to two excellent referees, whose comments and criticisms did a lot to improve the final version of the paper. One referee went above and beyond the call of duty and checked some of the numerics in Section 8, thus pinpointing a systematic error in some of our computations. We are particularly grateful to him/her. Any remaining errors are, of course, our responsibility.

2 Model Formulation

Let $\mathbb{X} = (X_t)_{t\geq 0}$ be a discrete time, real valued, stochastic process. As above, let $X_{t-d}^{t-1} = (X_{t-1}, \ldots, X_{t-d})$ denote the vector of its lagged values. Then, $\mathbb{X}$ is said to be *Markov of order* d if

$$\mathbb{P}(X_t \in B | X_0^{t-1}) = \mathbb{P}(X_t \in B | X_{t-d}^{t-1})$$

almost surely, for all Borel sets B, and all $t \geq d$.

Let $n(x; \mu, \sigma)$ denote the density of a Gaussian r.v. with mean μ and variance σ^2, and $\nu(\cdot)$ the corresponding probability measure over $\mathcal{B}(\mathbb{R})$.

Definition 2.1 $\mathbb{X}$ *is said to follow a mixture of autoregressions, or* MixAR*$(m; d)$, model if it is Markov of order d and is equipped with transition probability kernel*

$$\mathbb{P}\left(X_t \in B | X_{t-d}^{t-1} = \mathbf{x}; \boldsymbol{\theta}\right) = \sum_{j=1}^{m} g_j(\mathbf{x}; \boldsymbol{\theta}_g) \nu(B; \boldsymbol{\theta}_j^T \mathbf{x} + \theta_{j,0}, \sigma_j), \tag{2}$$

for every Borel set $B \in \mathcal{B}(\mathbb{R})$*, and* $\mathbf{x} \in \mathbb{R}^d$*. Here* $\boldsymbol{\theta}_j := [\theta_{j,1}, \ldots, \theta_{j,d}]^T \in \mathbb{R}^d$*,* $\boldsymbol{\theta}_g = [\mathbf{a}_1^T, b_1, \ldots, \mathbf{a}_m^T, b_m]^T \in \mathbb{R}^{m(d+1)}$*, while* $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1^T, \theta_{1,0}, \ldots, \boldsymbol{\theta}_m^T, \theta_{m,0}, \boldsymbol{\theta}_g^T, \sigma_1^2, \ldots, \sigma_m^2]^T \in \mathbb{R}^{2m(d+1)+m}$ *denotes the full parameter vector specifying the model. The weighting function* g_j *is a multinomial logit, specifically,*

$$g_j(\mathbf{x}; \boldsymbol{\theta}_g) \equiv \frac{\exp(\mathbf{a}_j^T \mathbf{x} + b_j)}{\sum_{k=1}^m \exp(\mathbf{a}_k^T \mathbf{x} + b_k)}, \qquad j = 1, \ldots, m \quad . \tag{3}$$

Remark 2.1 The specification of the weighting functions g_j as multinomial logits follows the original presentation of *mixtures of experts*, in the context of non-linear regression, by Jacobs *et al.* (1991). Obviously, other functions will also work here (e.g., radial basis functions).Our results on the generality of the MixAR models do however require that these functions obey some more technical restrictions [for more details see Zeevi *et al.* (1998), in particular their Assumption 2 specifies such conditions]. The results we obtain in Section 8 indicate that from a practical standpoint this choice works quite well. Nevertheless, the basic approach and results of this paper will also work for other prudent choices.

It is immediate from (2) that the conditional density of an observation of X_t, given the current state (lag vector) $X_{t-d}^{t-1} = \mathbf{x}$, is a mixture of Gaussians with means $\boldsymbol{\theta}_j^T \mathbf{x} + \theta_{j,0}$ and variances σ_j^2. The Gaussian components have a straightforward interpretation; each is the conditional density associated with a local Gaussian AR(d) process

$$X_t = \boldsymbol{\theta}_j^T X_{t-d}^{t-1} + \theta_{j,0} + \varepsilon_t^{(j)}$$

where $\varepsilon_t^{(j)} \sim \mathcal{N}(0, \sigma_j^2)$. Note that the lags of the AR models need not be identical, but for the purpose of writing out the model one can always choose the largest lag and 'pad out' all the corresponding parameter vectors with zeros. Note also that, despite our choice of terminology, it is clear that our choice of transition kernel does not follow standard mixture model formulations [cf. Titterington (1985)] in that the distribution over classes is a function of the *state vector* $\mathbf{x}_{t-d}^{t-1}$, rather than being constant. As mentioned above, it is precisely this aspect of the model that endows the MixAR model with non-linearity, concurrently the mixture structure retains tractability. The process of course is no longer Gaussian (as opposed to linear models), although the local structure is linear.

3 Relation to Other Models

Since the MixAR model described above bears at least superficial similarity to many others in the literature, we will attempt to indicate the precise nature of this relation.

The first attempts to incorporate non-linearity and non-stationarity arose in the statistics literature mainly through the contribution of two workers. The first, Maurice Priestley, introduced what he called 'state dependent models' in which processes were treated as locally linear [for an account of these models see Priestley (1988)]. Later, Tong (1983,1990) also modified the linear methodology in this spirit by introducing *threshold autoregressive* (TAR) models. There a threshold variable controls the switching between different autoregressive models. These models are conceptual antecedents of MixAR, and can formally be written as

$$X_t = \boldsymbol{\theta}_j^T X_{t-d}^{t-1} + \theta_{j,0} + \varepsilon_t^{(j)} \quad \text{if} \quad X_{t-d}^{t-1} \in \mathcal{R}^{(j)} \tag{4}$$

where $j = 1, 2, \ldots, m$, and $\{\mathcal{R}^{(j)}\}_{1 \leq j \leq m}$ is a partition of $\mathbb{R}^d$. This model is denoted TAR$(m; d)$, though to allow for different AR orders often the notation TAR$(m; d_1, d_2, \ldots, d_m)$ is used. Typically, the general assignment rule $\mathbf{x}_{t-d}^{t-1} \in \mathcal{R}^{(j)}$ is implemented via the much simplified $X_{t-k} \in [r_{j-1}, r_j)$ with $\{r_j\}$ the so called *threshold variables* and k the *delay parameter* [for more details see Tong (1990, pp. 98-103)].

In order to compare the MixAR and TAR models, assume for the moment that $\{\varepsilon_t^{(j)}\}$ in the TAR model are i.i.d. Gaussian with mean 0 and variance σ_j^2. Then the conditional probability measure for the TAR$(m; d)$ model can be written as

$$\mathbb{P}(X_t \in B | X_{t-d}^{t-1} = \mathbf{x}) = \sum_{j=1}^{m} \mathbb{I}_{\{\mathbf{x} \in \mathcal{R}^{(j)}\}} \nu(B; \boldsymbol{\theta}_j^T \mathbf{x} + \theta_{j,0}, \sigma_j) \tag{5}$$

with $\mathbb{I}_A$ the indicator function of the set A, and ν as before. Comparing this to (2), we see that the TAR model can be framed as a degenerate mixture model, with class probabilities that are state dependent but which, for each state, put all their mass on only one of the classes. Consequently, the conditional TAR transition kernel remains Gaussian, unlike the MixAR case.

Nevertheless, at least in the case of threshold variables inducing interval partitions of $\mathbb{R}$, the MixAR processes can easily be made to approximate TAR processes. For example, consider the simplest TAR process with only one threshold. All that is required is to note that logit functions can be made to approximate indicator functions; i.e. $[1 + \exp\{\alpha(r - x)]^{-1} \to \mathbb{I}_{(-\infty, r]}(x)$ as $\alpha \to \infty$. Details as to how

to extend this fact to more than two regimes are left to the reader. However, if one considers a TAR model with an arbitrary partitions of the real line into measurable sets, then there may not be any MixAR process of finite complexity (i.e. finite m) that can approximate it. Consequently, neither of these classes of processes is a strict subset of the other.

Numerous other models have also been introduced with the underlying idea of preserving local linearity. Some of the more prominent in recent years have been the functional coefficient autoregressive model (FAR) of Chen and Tsay (1993), which in some sense generalizes the TAR model, and some of its variants such as EXPAR of Haggan and Ozaki (1981), STAR and SETAR [cf. Tong (1983, 1990)], and the adaptive splines threshold autoregressive model (ASTAR) of Lewis and Stevens (1991). The recently proposed FAR and ASTAR models consider more general schemes of partitioning the state space, with the inherent expense of complicating the specification and estimation tasks. None of these, however, exhibit the same type of non-linearity as in the MixAR model.

Another related model is the NEAR(d) model of Chan (1988), developed initially in the setting of time series with non-negative values and exponential marginal distributions. In this model, X_t is generated by one of d different AR models, with each linear model assigned a fixed probability. That is,

$$X_t = \beta_{I_t} X_{t-I_t} + \varepsilon_t$$

with I_t a discrete random variable taking values in the set $\{0, 1, \ldots, d\}$, independently of the previous values of X. While the two models are not completely comparable, it is obviously that the MixAR model is considerably richer, in part since it allows the component choice to be dependent on the lagged variables, and allows each AR component to be of full order.

A somewhat similar idea was introduced in the Econometrics literature, under the name of 'regime switching models'. The philosophy underlying these models is that there may be occasional shifts in the mean, variance, or autoregressive dynamics of the time series. Hamilton (1990) proposes to model a time series as Gaussian AR(d) with mean that depends on the state of an independent (hidden) Markov chain. For this model, a simple algorithm for maximum likelihood estimation exits, namely the expectation–maximization (EM) algorithm. We will see in Section 7 that the same algorithm is also applicable in our setting, although we consider a more general construction (at least for the case of scalar valued time series). The idea of regime switching models is also part of Tjøstheim's (1986) work,

introducing a class of 'doubly stochastic models'. One such model is

$$X_t = \sum_{j=1}^{d} \Theta_t^{(j)} X_{t-j} + \varepsilon_t$$

with $\{\Theta_t^{(j)}\}$ a stochastic process independent of ε_t. For instance, one may take a hidden Markov chain with the vector $(\Theta^1, \Theta^2, \ldots, \Theta^d)$ a stochastic functional defined on the Markov chain. For further references the reader is referred to the monograph of Tong (1990), Granger and Teräsvirta (1993) who emphasize economic applications as well, and the review paper by Tjøstheim (1994).

4 On Parameters and Parsimony

One of the fundamental principles of time series analysis over the past few decades has been that of *parsimony*, which seems to have originated in Tukey (1961). In the words of Box and Jenkins (1976), this demands 'employing the smallest possible number of parameters for adequate representation'. One of the strengths of ARMA models among the class of linear time series modeling has indeed been their long acknowledged achievement of parsimony, and considerable experience has led to a central limit theorem based rule of thumb saying that more than 30 data points are needed for each estimated parameter.

While parsimony seems to be also a feature that would be desirable in non-linear models, it is not at all clear that the same rule of thumb should apply here. In fact, recent investigations of wide classes of quite different non-linear structures indicate that the 'old' parameter/data ratio might not be the right thing to be looking at. In particular, over the last few years several general statistical models have been proposed which attempt to construct a complex model by forming a convex linear combination of simple 'base' models. Three examples that follow this approach are the mixtures of experts of Jacobs *et al.* (1991), and in particular its incarnation in terms of the MixAR models that we are proposing, the method of 'bagging' introduced by Breiman [cf. Breiman (1998) for a detailed discussion], and the 'boosting' technique reviewed in detail in Freund *et. al.* (1998). These models are currently under vigorous investigation within the statistics community. While the specific methodology of each is slightly different, the overall structure of the final estimation is the same, namely a convex combination of 'simple' estimators. In *bagging*, a set of bootstrap samples is constructed, an estimator is formed from each and a convex combination of these estimators is then formed. In *boosting*, an initial

estimator is formed based on the data, the data is then re-sampled, giving larger weight to data points for which the error of the previous estimator is large, and a new estimator is then formed based on the re-sampled data. Finally, a convex combination is constructed where each estimator is weighted according to its empirical performance. In the mixture of experts situation, the basic model itself is a convex combination in the sense that the underlying conditional density assumes the form of a mixture distribution, with the novelty that the coefficients in the combination are themselves data-dependent. The associated predictor is then a convex (data-dependent) combination of linear (AR) predictors.

While very little theory has yet been developed, considerable empirical investigation has unquestionably demonstrated that both bagging and boosting lead to excellent performance, in spite of the fact that in many cases the *number of parameters is of the same order as the number of data points*. To be specific, consider for example Breiman's (1998) study of the performance of 'arcing' (a variant of boosting) on several classification problems. In one scenario, he uses a training set of size 300 to construct a convex combination of 50, 100, 250 and even 500 (!!) classification trees. Clearly the number of parameters here, for any reasonable definition of 'parameter', exceeds the number of samples in the training set. Quite surprisingly, Breiman finds that aggregating 500 trees drives the test set error down to nearly the Bayes risk. Moreover, on several benchmark data sets he observes that the test set (out-of-sample) error decreases with the number of trees in the aggregated classifier. In other words, *increasing the number of parameters far beyond the cardinality of the training set results in superior out-of-sample performance* [for exact details see Breiman (1998, §3.3 and §3.4)]. Similar results have been reported by Freund and Schapire (1996), experimenting with their boosting algorithm.

Recent theory also seems to indicate that counting parameters may not be adequate for characterizing the complexity and parsimony of certain classes of models, particularly when convex combinations of 'simple' models are used. An interesting and important theoretical contribution is the work of Bartlett (1998) [see also Bartlett *et al.* (1996)], where examples of estimators possessing an *infinite* number of parameters are presented, for which excellent convergence rates can be established. It turns out that in these situations there is an alternative quantity, termed the 'fat-shattering dimension', conceptually related to Vapnick-Cervonenkis dimension, which controls the complexity of the model. In this setting, empirical evidence is even backed up by theoretical results [e.g., Lee *et al* (1996), and the recent book by Anthony and Bartlett (1999)].

While this situation is anathema to the conventional wisdom, it seems that the type of procedures described above possess some inherent feature which avoids overfitting. To quote Friedman, Hastie and Tibshirani (1998) "One fascinating issue not covered in this paper is the fact that boosting, whatever flavor, seldom seems to overfit, no matter how many terms are included in the additive expansion". Since the number of terms in their system is directly proportional to the number of parameters, it seems that in this case the number of parameters is a very poor indicator of the complexity of the model. Friedman *et al.* conclude their article with the statement: "Whatever the explanation, the empirical evidence is strong; the introduction of boosting by Schapire, Freund and colleagues has brought an exciting and important set of new ideas to the table".

Overall, the bottom line of all this literature is that when dealing with non-linear models, estimation involving convex combinations of simple estimators seems remarkably free of the classical problems of overfitting, and specific algorithms and models (bagging, boosting, arcing, and mixtures of experts) play a much more significant rôle than is captured by a simple count of parameters.

The connection between this literature and the MixAR model is via the mixture structure of the latter. Specifically, one defines a Markov structure by taking a transition kernel that is a mixture (or convex combination) of Gaussian densities, corresponding to local AR models. The important additional ingredient is to allow the coefficients of the mixture to depend on the data. Consequently, there is some hope, and partial evidence (but no proofs), that MixAR may also be 'free' of the usual concern of over-parameterisation, as described above.

The proof of the pudding is in the eating, however, and that is the purpose of Section 8, where we look at a number of examples. In all of these, we split our data set into two parts. The first part, the *training set*, was used for model fitting and parameter estimation, and the second for assessing prediction quality. In all the examples the number of parameters used was large, in one case over half the number of data points, and so one would expect the fit for a MixAR model to be significantly better on the *training set* than that of competing models with fewer parameters. That this is the case is therefore not a surprise. However, one would also expect 'overfitting' to lead to an immediate degradation of predictive capabilty. This was simply not the case as Tables 1 and 2 in Section 8 indicate.

5 Probabilistic Properties

In this section we study some basic probabilistic properties of MixAR processes. As in the study of the simple AR processes, the most fundamental issues involve finding sufficient conditions to ensure the stochastic stability of the process. In particular, we need to know when a MixAR process will be stationary, or, equivalently, when a stationary probability measure exists. Since MixAR$(m;1)$ processes are Markov, the existence and uniqueness of such a measure imply that the process is ergodic. This is the minimal regularity we require in order to make useful inferences about the process, since it ensures, for example, that we can expect to find consistent parameter estimates, and also imply central limit theorems for estimators.

For MixAR$(m;d)$ processes with $d > 1$, we use the standard trick of replacing $\mathbb{X}$ by the vector process $\mathbf{Y}_t := X_{t-d}^{t-1}$. The fact that $\mathbb{X}$ is Markov of order d implies that $\mathbb{Y} = (\mathbf{Y}_t)$ is a Markov chain; i.e. Markov of order 1. We now need to find a stationary distribution for $(\mathbf{Y}_t)$, a problem which is equivalent to finding an 'initial distribution' for $X_1, \ldots, X_{d-1}$. Since the first coordinate of $\mathbf{Y}_t$ follows the sample path of X_t, ergodic and rate theorems for $\mathbf{Y}_t$ imply equivalent results for X_t.

To start, we first recall some basic definitions related to the study of Markov chains. We focus on the scalar case for simplicity, but the setting generalizes directly to vector valued processes. Let $\mathbb{X}$ be a MixAR$(m;1)$ process, and let π be a probability measure satisfying

$$\pi(B) = \int_{\mathbb{R}} p(x, B)\pi(dx), \qquad \forall B \in \mathcal{B}(\mathbb{R}),$$

with

$$\begin{aligned} p(x, B) &\equiv p(X_1 \in B | X_0 = x) \\ &= \sum_{j=1}^{m} g_j(x; \boldsymbol{\theta}_g)\nu(B; \theta_j x + \theta_{j,0}, \sigma_j). \end{aligned}$$

We also write $p^n(x, B)$ for the n-step transition probability $P(X_{n-1} \in A | X_0 = x)$. If such a measure π exists, it is called the invariant probability measure, and we can construct a stationary probability measure P for the process $\mathbb{X}$ using the standard bottom–up method to obtain marginals, followed with an application of the Kolmagorov extension theorem [for a standard application see Meyn and Tweedie (1993, pp. 66)]. If π is unique then the process is also ergodic. Under some further conditions, the rate at which the process settles down to its stationary behavior can also be ascertained. The above

are the main two measures of stochastic stability that will be employed in the sequel; the following two theorems summarize sufficient conditions. We treat MixAR$(m;1)$ processes separately, since a more detailed analysis can be performed here, and the conditions are more concise. Unfortunately, in the analysis of the general MixAR process we have only been able to obtain rather strong sufficient conditions.

For the MixAR$(m;1)$ case we assume, without loss of generality, that the AR components have been ordered so that $a_1 \leq a_2 \leq \cdots \leq a_m$, where the a_j parameterize g_j as in (3). The following theorem (as well as its extension to general MixAR$(m;d)$ processes), is based on sufficient conditions for stochastic stability of Markov chains on general state space. Our main reference for this is Meyn and Tweedie (1993). The proof, which is deferred to Appendix A, provides the details, definitions and quotes auxiliary results from Meyn and Tweedie (1993). Here, and in the sequel, we use the notation $x_n = O(a_n)$ if $\limsup_{n\to\infty} x_n/a_n < \infty$.

Theorem 5.1 *Let* $\mathbb{X}$ *follow a* MixAR$(m;1)$ *model, with* $m > 1$. *Suppose* $\theta_1 < 1$, $\theta_m < 1$, *and* $\theta_1\theta_m < 1$. *Then,* π *exists, is unique, and the process is geometrically ergodic. In particular, there exists* $r > 1$ *such that* $\sup_{B\in\mathcal{B}(\mathbb{R})} |p^n(x,B) - \pi(B)| = O(r^{-n})$.

We pause to make several comments on this result.

Remark 5.1 1. The theorem tells us that it suffices to restrict attention to the two local AR(1) models in the mixture that are 'dominant' in terms of the behavior of $g_1(x;\boldsymbol{\theta}_g)$ and $g_m(x;\boldsymbol{\theta}_g)$ for sufficiently large positive x values and negative x values respectively. Let us call them the 'extreme' models. A priori, it seems plausible that if *all* local autoregressive models are stable (in the sense that $|\theta_j| < 1$), then the MixAR$(m;1)$ model would be stable as well. This is in fact correct, but the theorem asserts a more refined result. Roughly, if the two 'extreme' AR models cause an average 'drift' towards 0, then the MixAR$(m;1)$ model is stable. Note that it suffices to have this so called 'drift', and in fact the two 'extreme' AR models need not be stable simultaneously. We note in passing that the conditions in Theorem 5.1 are similar to the ones obtained for TAR processes by Tong and co-workers [cf. Tong (1990), §4.1 for more details].

2. The ergodicity of $\mathbb{X}$ is sufficient to obtain strong laws of large numbers (via the ergodic theorem), and

a central limit theorem (via, say, martingale methods). Both limit theorems are crucial in establishing large sample properties of estimators, e.g., maximum likelihood. These implications are briefly discussed in Section 7. Geometric ergodicity, i.e., the property that the n-step transition probabilities converge in total variation (and weakly) to the marginal stationary distribution with a uniform geometric rate, also implies that the process is exponentially β-mixing [cf. Bradley (1984), for a discussion of mixing conditions]. The latter is useful in deriving further probabilistic properties of the process, such as uniform strong laws for empirical processes driven by MixAR processes.

3. The conditions in Theorem 5.1 are not meant to be exhaustive, but they do cover the obvious cases. For example, it is easy to check from the proof of the Theorem that if $|\theta_j| \leq 1$ and the global parameter $\boldsymbol{\theta} \in T_1 \cap T_2$, where $T_1 = \{(\theta_m = 1, \theta_{m,0} < 0), (\theta_m = -1, \theta_{m,0} > 0)\}$ and $T_2 = \{(\theta_1 = 1, \theta_{1,0} > 0), (\theta_1 = -1, \theta_{1,0} < 0)\}$, then there exists a unique invariant probability measure π and so the process is ergodic.

We now turn to the general MixAR$(m; d)$ process, with $(d \geq 1)$. Set

$$\alpha_k \equiv \max_{j=1,2,\ldots,m} |\boldsymbol{\theta}_j(k)|, \qquad k = 1, 2, \ldots, m,$$

where $\boldsymbol{\theta}_j(k)$ is the k-th component of the parameter vector $\boldsymbol{\theta}_j$ (defining the j-th autoregression in the MixAR model). Let $\|\cdot\|$ denote the usual Euclidean norm. The following theorem, the proof of which is also in Appendix A, establishes sufficient conditions for geometric ergodicity of the MixAR process.

Theorem 5.2 *Let $\mathbb{X}$ follow a* MixAR*$(m; d)$ model. Assume that the polynomial*

$$\mathcal{P}(z) = z^d - \sum_{k=1}^{d} \alpha_k z^{d-k}, \qquad z \in \mathbf{C}$$

has all its zeros in the open unit disk, $z < 1$. Then, the vector process $\mathbf{Y}_t = X_{t-d}^{t-1}$ has a unique stationary probability measure, and is geometrically ergodic in the sense of of Theorem 5.1.

Note that if $d = 1$, Theorem 5.2 requires that $\alpha_1 < 1$, i.e. $\max_{1 \leq i \leq m} |\theta_i| < 1$. While this is consistent with Theorem 5.1 (a), which is close to sharp, it is clear that for $d > 1$ our results are rather stringent. Further investigation is needed here.

6 Approximating Non-linear Prediction Functions

In this section we will focus on a basic approximation property of the MixAR model; *viz.* its ability to provide high quality prediction for general stationary processes. Our discussion is meant to be indicative of the flexibility of the model, and not comprehensive.

Consider a strictly stationary, real valued, discrete time process $\mathbb{X}$ with $\mathbb{E}|X_t|^2 < \infty$. Then it is well known that the optimal (in mean squared error sense) predictor (generally non-linear) of X_t given the past (i.e., its past filtration) is the conditional mean. If X is Markov order d, then this is given by the conditional expectation

$$f(\mathbf{x}) = \mathbb{E}[X_t | X_{t-d}^{t-1} = \mathbf{x}] \quad . \tag{6}$$

Suppose we propose a MixAR$(m; d)$ model for the underlying process $\mathbb{X}$, which may not, in fact, follow such a model. One of a plethora of measures of closeness for such an approximation is closeness of predictions. That, given a sufficiently large m, we can always do well on this count is the content of Theorem 6.1 following, the proof of which is in Appendix A. We require a little notation first. Let

$$\begin{aligned} f_m(\mathbf{x}; \boldsymbol{\theta}) &\stackrel{\triangle}{=} \int_{\mathbb{R}} x \, P(dx | X_{t-d}^{t-1} = \mathbf{x}; \boldsymbol{\theta}) \\ &= \sum_{j=1}^{m} g_j(\mathbf{x}; \boldsymbol{\theta}_g)[\boldsymbol{\theta}_j^T \mathbf{x} + \theta_{j,0}] \end{aligned} \tag{7}$$

where $P(\cdot | X_{t-d}^{t-1} = \mathbf{x}; \boldsymbol{\theta})$ is the transition kernel of the MixAR$(m; d)$ process, as in (2). It is clear that *if* $\mathbb{X}$ is a MixAR process, then f_m is the optimal (non-linear) mean square predictor. In general, however, this will not be the case (i.e., when $\mathbb{X}$ is not a MixAR$(m; d)$ process).

Let μ denote the d-dimensional stationary distribution of $(X_t, \ldots, X_{t+d})$, and let $L^q(\mathbb{R}^d, \mu)$, $q \in [0, \infty)$ be the space of measurable functions, for which $\|f\|^q \equiv \int |f|^q d\mu < \infty$. Then we have the following result, which we state for general q, but which is of primary statistical interest when $q = 2$.

Theorem 6.1 *Let $\mathbb{X}$ be a strictly stationary order d Markov process. Assume that the d order marginal μ has a density with respect to Lebesgue measure, continuous and bounded on $\mathbb{R}^d$. Then, for every $q \in [1, \infty)$ for which $\mathbb{E}|X_t|^q < \infty$, and all $\epsilon > 0$, there exists an m sufficiently large, and a stationary* MixAR*$(m; d)$ process, for which*

$$\|f - f_m\|_{L^q(\mathbb{R}^d, \mu)} < \epsilon \quad .$$

where f and f_m are defined in (6) and (7).

Remark 6.1 1. While the result of Theorem 6.1 seems standard at first glance, the proof is a little tricky. In similar situations, such as the corresponding results for TAR models and their variants [cf. Tong (1983)], the fact that the state space is being partitioned into clearly defined sets via indicator (simple) functions makes the approximation of L^q functions technically easier. In contrast, the partition induced by the weighting functions g_j in the MixAR model raises some technical difficulties. In fact, it is not at all clear whether one generates a 'rich' enough partition in this manner.

2. Under stronger assumptions on the underlying true process, one can go beyond Theorem 6.1 and actually obtain degree of approximation results. For example, suppose that $\mathbb{P}(X_t \in [-K, K]) = 1$, for some finite K, and that f (the prediction function) is 'smooth enough', in that f is in a Sobolev ball $W_r^q(L)$; i.e. f has r continuous derivatives whose L^q norms are uniformly bounded over $[-K, K]^d$. In this case Zeevi *et. al.* (1998) show that there is a constant c, dependent only on r, d, q and K, such that

$$\sup_{f \in W_r^q} \inf_{f_m} \|f - f_m\|_{L^q([-T,T]^d, \mu)} \leq \frac{c}{m^{r/d}}.$$

7 Parameter Estimation

We now turn to the issues of parameter estimation and statistical inference for MixAR processes. We will concentrate exclusively on maximum likelihood (ML) methods, and focus on efficient implementation of the estimation algorithm. Asymptotics are only briefly discussed, as they are based on well known results on estimation in Markov processes [cf. Billingsley (1961)], and recent results pertaining to ML estimation in mixtures of experts [cf. Jiang and Tanner (1999, 2000)].

Let $\mathcal{D}_N = \{X_t\}_{t=1}^N$ be a sample from a stationary, ergodic, MixAR$(m; d)$ process. The likelihood equations are

$$\begin{aligned} L(\mathcal{D}_N; \boldsymbol{\theta}) &= p(X_1, X_2 \ldots, X_N; \boldsymbol{\theta}) \\ &= p(X_1, X_2, \ldots, X_d; \boldsymbol{\theta}) \prod_{t=d+1}^{N} p(X_t | X_{t-1}, X_{t-2}, \ldots, X_{t-d}; \boldsymbol{\theta}) \end{aligned} \tag{8}$$

with $p(x_1, x_2, \ldots, x_d; \boldsymbol{\theta})$ the d-dimensional marginal of the stationary distribution.

We first tackle the algorithmic problem of obtaining a ML estimator $\hat{\boldsymbol{\theta}}_N$ of $\boldsymbol{\theta}$ from (8). As a non-linear model, it is not, *a priori*, clear whether efficient and robust procedures exist for estimating the parameters of the MixAR$(m;d)$ via the maximum likelihood (ML) method. It turns out however that a very efficient estimation procedure, the so-called Expectation–Maximization (EM) algorithm [see, Dempster *et al.* (1977)] can be applied. This algorithm has been extensively studied and utilized in related contexts. [See Titterington *et al.* (1985), for a comprehensive summary.] The algorithm was also formulated by Jordan and Jacobs (1994) in the context of a regression problem (for i.i.d. data), using the mixture of experts model. In Jordan and Xu (1996), theoretical convergence results were obtained. Based on the two cited references, we adapt the algorithm to our situation.

Before discussing the details of the estimation algorithm we make the standard step of neglecting the terms involving the first d observations in (8), and working with the *reduced likelihood* equation. On substituting the transition kernel of the MixAR (2) we have

$$L(\mathcal{D}_N;\boldsymbol{\theta}) = \prod_{t=d+1}^{N} \sum_{j=1}^{m} g_j(X_{t-d}^{t-1};\boldsymbol{\theta}_g) n(X_t;\boldsymbol{\theta}_j^T X_{t-d}^{t-1} + \theta_{j,0}, \sigma_j) \tag{9}$$

with $n(x;\mu,\sigma)$ the Gaussian density. Set

$$\hat{\boldsymbol{\theta}}_N \equiv \arg\max_{\boldsymbol{\theta}} L(\mathcal{D}_N;\boldsymbol{\theta}) \quad .$$

Direct maximization of the reduced likelihood (9), or equivalently its logarithm, needs to be avoided, essentially because of the complexity arising from the mixture structure of the MixAR transition kernel which leads to logarithms of sums. To avoid this difficulty, we turn to the EM algorithm.

The essence of the EM algorithm lies in augmenting the original data with a set of *missing variables*, which in our set up are chosen to be indicator variables, indicating which of the AR mixture components generated a given sample. The *complete data* likelihood then assumes the simple double product form

$$P(\mathcal{Z}|\boldsymbol{\theta}) = \prod_{t=d+1}^{N} \prod_{j=1}^{m} \Big[g_j(X_{t-d}^{t-1};\boldsymbol{\theta}_g) n(X_t;\boldsymbol{\theta}_j^T X_{t-d}^{t-1} + \theta_{j,0}, \sigma_j) \Big]^{I_t^j}, \tag{10}$$

where

$$I_t^j = \begin{cases} 1 & \text{if } X_t \text{ is generated by the } j\text{-th mixture component,} \\ 0 & \text{otherwise.} \end{cases}$$

and $\mathcal{Z} = \{X_t, \{I_t^j\}_{j=1}^m\}_{t=d+1}^N$ denotes the set of observed and missing variables. The logarithm of (10) is

$$\log P(\mathcal{Z}|\boldsymbol{\theta}) = \sum_{t=d+1}^{N} \sum_{j=1}^{m} I_t^j \log \Big[g_j(X_{t-d}^{t-1};\boldsymbol{\theta}_g) n(X_t;\boldsymbol{\theta}_j^T X_{t-d}^{t-1} + \theta_{j,0}, \sigma_j) \Big] \tag{11}$$

This is a random variable in the missing data, and so in the next step we take an expectation with respect to them. Thus, we have the complete expected log-likelihood

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &\equiv \mathbb{E}\log P(\mathcal{Z}|\boldsymbol{\theta}) \\ &= \sum_{t=d+1}^{N}\sum_{j=1}^{m}\tau_{t,j}^{(k)}\log[g_j(X_{t-d}^{t-1};\boldsymbol{\theta}_g] \\ &+ \sum_{t=d+1}^{N}\sum_{j=1}^{m}\tau_{t,j}^{(k)}\log\left(n(X_t;\boldsymbol{\theta}_j X_{t-d}^{t-1}+\theta_{j,0},\sigma_j)\right) \end{aligned} \tag{12}$$

where the superscript (k) denotes the iteration step of the algorithm, $\boldsymbol{\theta}^{(k)}$ denotes the parameter vector for that iteration, and $\mathbb{E}(\cdot)$ is the expectation operator with respect to the distribution of the *missing data*.

The defining equation for $\tau_{t,j}^{(k)}$ is

$$\tau_{t,j}^{(k)} = \frac{g_j(X_{t-d}^{t-1};\boldsymbol{\theta}_g^{(k)})n(X_t;\boldsymbol{\theta}_j^{(k)}X_{t-d}^{t-1}+\theta_{j,0}^{(k)},\sigma_j^{(k)})}{\sum_{i=1}^{m} g_i(X_{t-d}^{t-1};\boldsymbol{\theta}_g^{(k)})n(X_t;\boldsymbol{\theta}_j^{(k)}X_{t-d}^{t-1}+\theta_{j,0}^{(k)},\sigma_j^{(k)})}. \tag{13}$$

Thus $\tau_{t,j}^{(k)}$ may be interpreted as the posterior probability of classifying X_t to the j-th class in the mixture, having observed the lag vector X_{t-d}^{t-1} and given the current parameter fit $\boldsymbol{\theta}^{(k)}$.

The process of formulating (12) and (13) is called the *expectation step* (**E**). With the complete likelihood Q at hand, we now turn to the second step of the EM algorithm, namely the *maximization step* (**M**). The objective of the maximization step is to maximize $Q(\cdot|\boldsymbol{\theta}^{(k)})$, and consequently [see the proofs in Dempster *et al.* (1977) and Wu (1983)], the likelihood (9) is maximized as well. Differentiating with respect to $\boldsymbol{\theta}$ and going through some algebra yields the parameter update equations

$$\begin{aligned} \sigma_j^{(k+1)} &= \left(\frac{1}{\sum_{t=d+1}^{N}\tau_{t,j}^{(k)}}\sum_{t=d+1}^{N}\tau_{t,j}^{(k)}[X_t-\boldsymbol{\theta}_j^{(k)}X_{t-d}^{t-1}-\theta_{j,0}^{(k)}][X_t-\boldsymbol{\theta}_j^{(k)}X_{t-d}^{t-1}-\theta_{j,0}^{(k)}]^T\right)^{1/2} \\ \left[(\boldsymbol{\theta}_j^{(k+1)})^T,\theta_{j,0}\right]^T &= (R_j^{(k)})^{-1}c_j^{(k)} \\ \boldsymbol{\theta}_g^{(k+1)} &= \boldsymbol{\theta}_g^{(k)} + \alpha\left.\frac{\partial Q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_g}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} \end{aligned} \tag{14}$$

where the matrix R_j and the vector c_j are given by

$$\begin{aligned} R_j^{(k)} &= \sum_{t=d+1}^{N}\tau_{t,j}^{(k)}\mathcal{X}_{t-1}\mathcal{X}_{t-1}^T \\ c_j^{(k)} &= \sum_{t=d+1}^{N}\tau_{t,j}^{(k)}\mathcal{X}_{t-1}x_t \quad , \end{aligned} \tag{15}$$

and the $(d+1)$-dimensional vector $\mathcal{X}_{t-1}$ is given by $\mathcal{X}_{t-1}^T = \left[(X_{t-d}^{t-1})^T, 1\right]$. Note that, in principle, the update of $\boldsymbol{\theta}_g$ is not given by an explicit equation, and so requires numerical optimization. In this work we use the steepest ascent algorithm, as evident in the update equation, where α is the step size obtained from a line search. By defining this step of steepest ascent we are actually in the framework of the GEM (generalized EM) algorithm [cf. Dempster *et al.* (1977)], although for consistency we continue refering to it simply as EM.

One of the main advantages of the above estimation algorithm is its straightforward implementation. The global convergence result given in Proposition 7.1 asserts that under quite general conditions, the algorithm will terminate. As far as our practical experience with the algorithm is concerned, we have not experienced any problems in quite a few test cases which we investigated, some of which are reported in Section 8. Slow convergence has been noted, however. As in all non-linear optimization problems the usual 'rule of thumb' applies, namely, one should experiment with multiple initial conditions.

In summary, the algorithm is as follows.

Algorithm 7.1

1. **Initialize**: Fix $\delta > 0$ (a tolerance parameter), set $k = 0$ (counter), and set an initial parameter value $\boldsymbol{\theta}^{(0)}$.

   - **E** step: Compute the $\tau_{t,j}^{(k)}$ in (13).
   - **M** step:

     M1. Compute $\sigma_j^{(k+1)}, \boldsymbol{\theta}_j^{(k+1)}, \theta_{j,0}$ for $j = 1, 2, \ldots, m$ using the re-estimation equations in (14).

     M2. Update the value of $\boldsymbol{\theta}_g^{(k)}$ by performing one step of steepest ascent as given in the last equation in (14).

2. **Stopping rule**: Compute the new value of $Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k+1)})$ and inspect the ascent condition

$$\left| Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) \right| > \delta \quad .$$

   If this holds then increment k and goto **E**, else terminate with the current value of $\boldsymbol{\theta}$.

Having cast the optimization problem as an EM algorithm one can utilize the theoretical results concerning its convergence properties. In particular, it is well known that the local convergence rate of the EM algorithm is linear [cf. Redner and Walker (1984)]. A detailed discussion of convergence properties of the particular variant presented above is beyond the scope of this paper. Some details can be found in the work of Jordan and Xu (1996) on the EM algorithm applied to mixtures of experts. However, for completeness of presentation we give the following proposition, which summarizes the key characteristic. A sketch of the proof, which is quite standard, is given in Appendix A.

Proposition 7.1 *Let* $\mathcal{L}(\boldsymbol{\theta}) = \log P\left(X_{d+1}^{n}; \boldsymbol{\theta}\right)$*, and let* $\{\boldsymbol{\theta}^{(k)}\}$ *be the sequence produced by Algorithm 7.1. Suppose that that the sequence* $\{\boldsymbol{\theta}^{(k)}\}$ *is contained in a compact subset of* $\mathbb{R}^{2m(d+1)}$*. Then,*
(1) All limit points of $\{\boldsymbol{\theta}^{(k)}\}$ *are stationary points of* $\mathcal{L}(\boldsymbol{\theta})$
(2) $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ *converges monotonically to* $\mathcal{L}^* := \mathcal{L}(\boldsymbol{\theta}^*)$*, for some stationary point* $\boldsymbol{\theta}^*$*.*
If in addition $\{\boldsymbol{\theta} : \mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^*\} = \{\boldsymbol{\theta}^*\}$*, then*

$$\boldsymbol{\theta}^{(k)} \longrightarrow \boldsymbol{\theta}^*$$

as $k \rightarrow \infty$*.*

Note, that in order to obtain convergence of the sequence $\boldsymbol{\theta}^{(k)}$ one needs to impose the more strict condition of the limit set of stationary points being a singleton. Consequently, under the conditions of Proposition 7.1, we have that $\boldsymbol{\theta}^{(k)}$ converges to $\hat{\boldsymbol{\theta}}_N$, the ML estimator, if this maximizer is the unique stationary point of the likelihood.

We complete this section with a brief discussion of the properties of the ML estimator, starting with the issue of identifiability. In mixture models, this is a rather subtle point which requires special interpretation. In what follows, we adopt the convention of distinguishing between parameterizations, only if they are not permutations on the index class $j = 1, 2, \ldots, m$ (i.e. a simple relabeling among classes). For more details see the review paper by Redner and Walker (1984). The only point that does require some care arises from the fact that in our case the means of the Gaussian components are affine transformations of the lag vector, $\mu_j = \boldsymbol{\theta}_j^T \mathbf{x}_{t-d}^{t-1} + \theta_{j,0}$, and not fixed parameters. However,

distinct parameterizations $(\boldsymbol{\theta}_j, \theta_{j,0})$ will necessarily result in $\mu_j = \mu_i$ only on sets of Lebesgue measure 0 (hyperplanes in $\mathbb{R}^d$). In addition, we require that the parametrization of the weight functions g_j be defined via the first $m-1$ (in lexicographical order, say) vectors $\boldsymbol{\theta}_g$. This is due to the normalization of the g_j's. In what follows, we assume the model is identifiable in the above sense. For more details on the issue of identifiability see the recent work by Jiang and Tanner (2000, Theorem 1 and Corollary 1).

Assume that the MixAR process is both stationary and ergodic, sufficient conditions for which are given in Theorems 5.1 and 5.2. Let $\mathbf{Y}_t = (X_t, \ldots, X_{t-d+1})$ be the usual vector valued Markov process, and set

$$u(\mathbf{Y}_{t-1}, \mathbf{Y}_t; \boldsymbol{\theta}) \equiv \log p(\mathbf{Y}_t | \mathbf{Y}_{t-1}; \boldsymbol{\theta}).$$

Let u_k and $u_{k\ell}$ denote the first and second order derivatives of $u(w, z; \boldsymbol{\theta})$ with respect to the components of $\boldsymbol{\theta}$. Let $J = (J_{k\ell})$ be the Fisher information matrix, with $J_{k\ell} \equiv \mathbb{E}[u_k u_\ell]$. The empirical counterpart is easily obtained by taking the expectation w.r.t. the empirical distribution. Then , under mild side conditions, Theorems 2.1 and 2.2 of Billingsley (1961, pp. 10-14) apply, and establish that the MLE $\hat{\boldsymbol{\theta}}_N$ is a consistent, and asymptotically normal estimator of $\boldsymbol{\theta}_0$ (the underlying parameter vector of the MixAR process), so that

$$\hat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0, \quad \text{and} \quad \sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \Rightarrow \mathcal{N}(0, J^{-1}), \quad \text{as } N \to \infty,$$

where $\xrightarrow{p}$ denotes convergence in probability, and $\Rightarrow$ denotes convergence in distribution. We do not pursue the technical details of these arguments, which are somewhat tedious, but the essential arguments follow as in the i.i.d. case which is treated nicely in Jiang and Tanner (1999).

8 Numerical Examples

We will now illustrate the power and versatility of the MixAR model by applying it to simulated data and some benchmark time series. In our study, linear (ARMA) models, as well as threshold (TAR) models, will in general constitute the reference point; in doing so we quote the best results cited in the literature when available.

In the past few years, neural networks have become a popular tool in both modeling and forecasting in many fields. As a parametric model, they have been applied to problems of signal processing, time

series analysis and prediction. [For a review and recent contributions see Weigend and Gershenfeld (1994).] Although neural nets are not mainstream statistical models for time series analysis, their recent popularity and extensive use has led us to include them as yet another reference point. When referring to a neural network model NN$(m; d)$ we mean a feed-forward network, with an input layer of dimension d, a hidden layer with m sigmoidal units, and one linear output unit. This network realizes the function $f_m(\mathbf{x}) = \sum_{j=1}^{m} \alpha_j \sigma(\mathbf{w}_j^T \mathbf{x} + b_j)$ with $\sigma(x)$ being the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$, and $\alpha_j, b_j \in \mathbb{R}$, $\mathbf{w}_j \in \mathbb{R}^d$. For more details see Weigend *et al.* (1990), and Granger and Teräsvirta (1993).

8.1 Preliminaries

The following applies to the time series analyzed in Section 8.4 and 8.5 respectively. Each data set was partitioned into two (or more) subsets. The initial portion was used for estimation purposes, and so is referred to as the *training* set. The second subset was not revealed during the estimation phase, and was subsequently used to evaluate the performance of the fitted model. This subset will be referred to as the *prediction* set. This is particularly important to keep in mind when reading the results, since in both benchmark time series the fitted MixAR model involves a number of parameters that is of the order of 1/3–1/2 of the number of observations in the *training set*.

In one case (the Lynx trapping sequence) we applied a logarithmic variance stabilizing transformation; all reference to the series will henceforth be understood to be refering to the transformed series. Also, sequences were normalized using a linear transformation to the interval $[0, 1]$, so as to avoid standard numerical problems associated with the non-linear programming part of the estimation algorithm. Prediction values were transformed back to the original scale. The estimation algorithm that was implemented was the EM variant discussed in Section 7, using `MATLAB` code. In the sequel we will use the normalized mean squared error which is denoted NMSE for brevity. This accuracy measure is defined as follows. For any subset of the time series $\mathcal{S}$, and any predictor $\hat{X}$,

$$\begin{aligned} \text{NMSE}(\mathcal{S}, \mathcal{T}) &= \frac{\sum_{t \in \mathcal{S}} [X_t - \hat{X}_t(X_{t-d}^{t-1}; \mathcal{T})]^2}{\sum_{t \in \mathcal{S}} [X_t - \bar{X}(\mathcal{S})]^2} \\ &= \frac{1}{\hat{\sigma}_{\mathcal{S}}^2} \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} [X_t - \hat{X}_t]^2 \quad . \end{aligned} \tag{16}$$

Where $\mathcal{T}$ denotes the training set, $\bar{X}(\mathcal{S})$ is the empirical mean over $\mathcal{S}$, and $\hat{X}_t(X_{t-d}^{t-1}; \mathcal{T})$ and is the

predictor based on the parameters estimated using the set $\mathcal{T}$. The notation $|\cdot|$ means the cardinality of a set. It is quite obvious that this global measure is rather coarse, in particular in our setting where the transition kernel of the MixAR Markov chain is a mixture, and thus not unimodal. However the NMSE is such a standard and frequently used measure of fit, that we will use it here as well. In particular, it enables an easy comparison of the performance of MixAR with other models, as presented in Sections 8.4 and 8.5.

To shed some light on the dynamics of the MixAR model, and in particular the weighting function $g_j(\cdot)$ which determines the local AR behavior, we present for one test case the entropy of the probability weights $\{g_j\}$ defined as

$$\mathcal{H}_{\{g\}}(X_{t-d}^{t-1} = \mathbf{x}) = -\sum_{j=1}^{m} g_j(\mathbf{x}; \boldsymbol{\theta}_g) \log_2 g_j(\mathbf{x}; \boldsymbol{\theta}_g),$$

for each value of the lag-vector, and is a measure of the extent to which the AR models are 'mixed'. The extreme points are the assignment of probability one to a particular AR model (entropy 0), and a uniform distribution over all m models (entropy $\log_2 m$).

8.2 MixAR Model Specification

To specify a MixAR$(m; d)$ model, one must determine the number of linear AR models involved (m) and the maximal lag size (d).

1. **Determining the lag size** (d): Since the MixAR model has local AR models, we anticipate that the order of these models can be roughly determined in the same manner as in the case of linear models (e.g, via spectral analysis). In this respect, the local linear structure of MixAR is essential to make this 'leap of faith'. In practice, our numerical results seem to confirm that this is not a bad approximation. A careful look at the dynamics of the MixAR process (see for example Figures 3 and 5) seem to indicate that this may not be as big a 'leap' as it may seem at first glance. Note that without the local linear structure we would not be able to approach the problem in this straightforward way and would have to resort to methods from non-linear dynamical systems.

2. **Determining the number of AR models** (m): We propose a rather simple minded algorithm for model specification. Again, our approach makes use of the fact that the MixAR is a combination of local linear models. Schematically, we propose

```
Set m=2
Set precision level delta > 0
Call estimation_algorithm(m)
While all |parameters_AR_j - parameters_AR_i|> delta   do
  m=m+1
  call estimation_algorithm(m)
End
```

This amounts to gradually increasing the number of local models until the parameters of two local AR models 'degenerate', in the sense that the differences between parameter values are below a specified threshold δ. In practice, this degeneracy occurs rather dramatically, so that the issue of setting δ does not seem to pose a difficulty.

Remark 8.1 The 'model selection' criterion we propose for setting m is subject to immediate concerns. It is heuristic, and does not build on a rigorous procedure such as hypothesis testing. Moreover, given the structure of the MixAR model, (see the discussion in Section 4), one can expect that for moderate length time series, the fitted model will end up with a number of parameters that is a sizable fraction of the number of observations. This is what we have observed in practice, and in particular in both of the benchmark time series we analyze below. It thus seems unreasonable to expect that the asymptotic considerations behind the usual variable/model selection methods could be justified and applied in an automated fashion. At this point we do not have an alternative solid theory to offer. However, we believe strongly that until such a theory emerges a simple minded heuristic is preferable to false use of more 'rigorous' procedures. Our practical experience with the MixAR model, the estimation algorithm and the crude 'model selection' procedure, indicates that this 'rule of thumb' seems to work quite well.

8.3 Simulated TAR Data

We start with an example of data generated by a TAR model. As explained in Section 3, the MixAR model should be capable of reproducing TAR dynamics, and so the fitted parameters from a simulated TAR data set should be consistent with the local linear structure of the that model.To see how this works, consider a time series constructed from two linear regimes as follows:

$$X_t = \begin{cases} 0.8X_{t-1} + 1 + \varepsilon_t & X_{t-1} < 0 \\ -0.5X_{t-1} - 0.6 + \varepsilon_t & X_{t-1} \geq 0 \end{cases}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.2$. This is an example of a first order TAR model, TAR(2;1,1), with the same noise level in each region. We simulated 100 data points from this model, and used them as a training set for a MixAR$(2;1)$, composed of two local AR(1) models. Figure 1 summarizes the performance results of the MixAR$(2;1)$. The local patterns may be best understood from Figure 1(b),

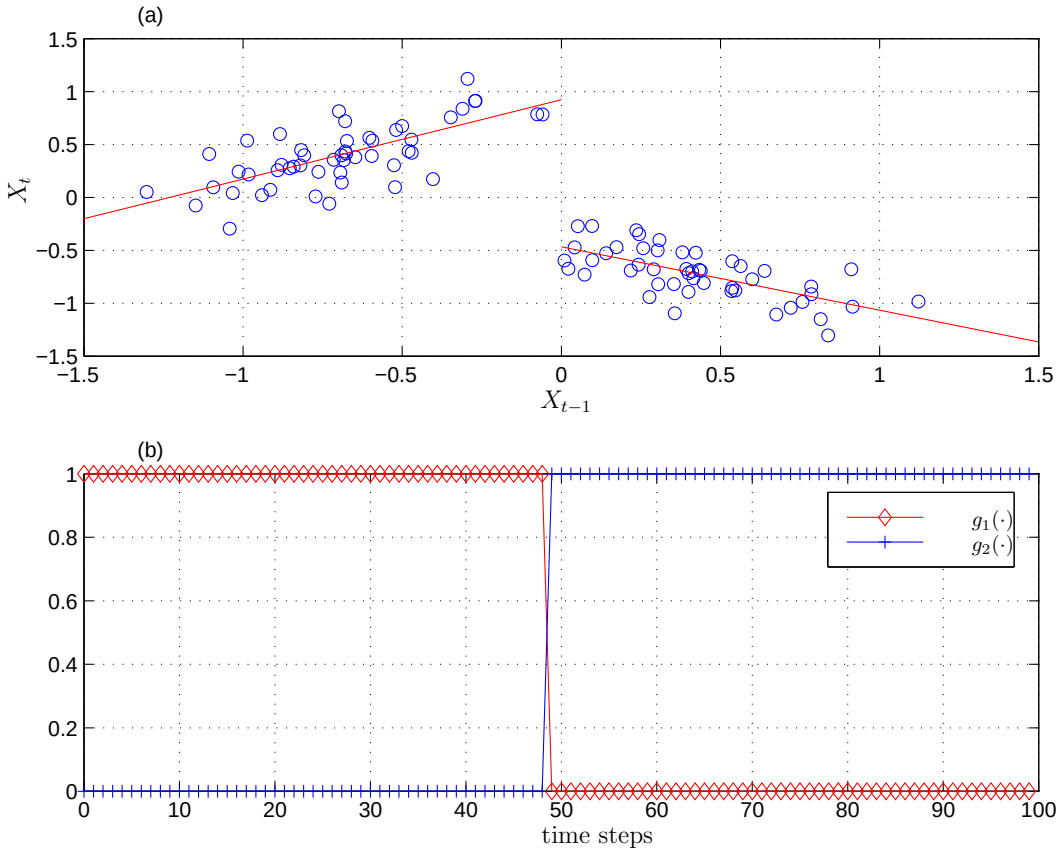

Figure 1: Analysis of TAR generated data. (a) Simulated data points and local AR models fitted by the MixAR$(2;1)$ model (solid line), observed in (X_t, X_{t-1}); (b) local AR model probabilities $g_1(\cdot)$ and $g_2(\cdot)$ respectively.

showing the two AR(1) models that had been fitted, superimposed on the TAR generated data (viewed in the lagged variable plane). The parameter estimates for the two AR models were respectively, $\theta_1 = 0.7501,\ \theta_{1,0} = 0.9244$ for the first, and $\theta_2 = -0.6004,\ \theta_{2,0} = -0.4655$ for the second. It is clear from viewing Figure 1(b) that the state-dependent probabilities (g_1 and g_2) degenerate in this case, thus best approximating the underlying threshold model generating the data. Indeed, the magnitude of the elements of $\boldsymbol{\theta}_g$ (parameterizing the g_j's) was very large (on the scale of 100), enabling the type of behavior seen in Figure 1(b). Together, the results shown in Figure 1 demonstrate that the MixAR(2;1) exactly captures the essence of the underlying stochastic process. We note in passing that the essence of this example is qualitative. Clearly, one observes that the estimated parameters of the local AR models have non-negligible bias. However, the point to be taken here is that the *essence* of the dynamics of the underlying process are well captured.

8.4 Canadian Lynx Time Series

Analysis of the Canadian lynx time series is summarized in Figure 2. Recall, we analyze the series after applying a logarithmic (base 10 logarithm) variance stabilizing transformation. Observations from the period of 1821–1920 constituted the *training set*, while 1921–1934 was used as the *prediction set*. The graphs display a normalized time scale; 1–93 training set, and 94–107 is the prediction set. Using the model specification procedure outlined in Section 8.2, we have chosen a MixAR$(3; 6)$ that utilizes three AR(6) models, thus having a total of 45 parameters. We should note that spectral analysis leads one to entertain two possibilities for the lag size: there is one dominant peak at ≈ 10 years, and the second at ≈ 5 [see the discussion in Tong (1990), p. 365, for further details]. We ended up considering a model with lag $d = 6$, since we found it to performs marginally better than the model with $d = 5$. The values of the estimated parameters are given in Appendix B.

Note that the realization of the log lynx data deviates from the anticipated reversibility structure of a Gaussian linear model, as it rises to, and falls from, its local maximum at different rates. Consequently, it is natural to consider a non-linear model. The MixAR captures this phenomenon as different AR models are 'assigned' to the rising and falling patterns accordingly, as shown in Figure 2(d). For the estimation portion of the sequence, the residual error sequence seems reasonably random; i.e., the residual ACF and PACF do not show serial correlation. The details are omitted.

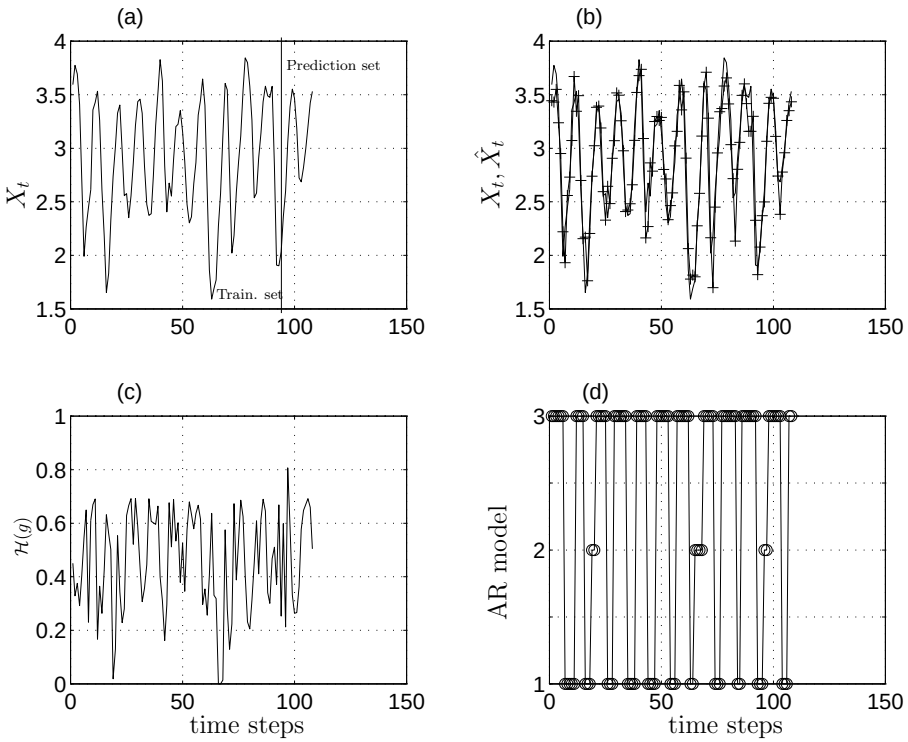

Figure 2: Analysis of the log transformed Canadian lynx time series (1821-1934). (a) The log lynx time series. (b) One step prediction of the MixAR(3; 6) model (+, line) for the training set (1-93) and prediction set (94-107), superimposed on the original data series. (c) Entropy $\mathcal{H}(g)$ of the AR component probabilities $\{g_j\}$. (d) Most probable AR model (largest value of $\{g_j\}_{j=1}^3$).

To contrast the performance of the MixAR with other models, we quote results that have been obtained in analyzing the log-lynx time series using the following models.

(i) A linear model: AR(12) given in Brockwell and Davis (1991, p. 550)

(ii) A threshold model: TAR(2;8,3) studied in Lim and Tong (1980) [see also Tong (1983), p. 190].

(iii) Two parsimonious non-linear models: a bilinear model considered by Subba and Gabr (1984, p. 204), and an AR(2) random coefficient varying model studied by Nichols and Quinn (1982, p. 144). The results are taken from Brockwell and Davis (1991, p. 552).

Model	Training set	Prediction set
AR(12)	0.1159	0.1392
TAR$(2; 8, 3)$	0.1320	0.1052
Bilinear	—	0.0964
Random coefficient varying AR(2)	—	0.0981
MixAR$(3; 6)$	0.1214	0.0732

Table 1: NMSE results for the log-lynx series training and prediction sets

The performance of the MixAR can probably be best appreciated by noting, from Table 1, the significant decrease in prediction error when using the MixAR model, as compared to the other models listed there. What is particularly interesting is that the fit on the training set is not notably better for the highly parametrized MixAR model, compared with the more 'parsimonious' TAR model, although *a priori* one might have expected this. Also, note that the significantly reduced prediction error is also surprising, since the large number of parameters in the MixAR model would suggest overfitting resulting in reduced out-of-sample performance.

Some further evidence supporting the prediction quality of MixAR can be drawn from Figure 3, which depicts the one-step predictions for the *prediction set*. Figure 3(b) shows the state dependency of the model; different patterns are described by the local AR models. It is clear from Figure 3(a) that the resulting predictions track the the out-of-sample values very well.

8.5 Sunspots Time Series

The Wölfer sunspot number data set is perhaps one of the most extensively studied time series in the literature; it makes a natural benchmark. Although yearly averages of this sunspot activity have been

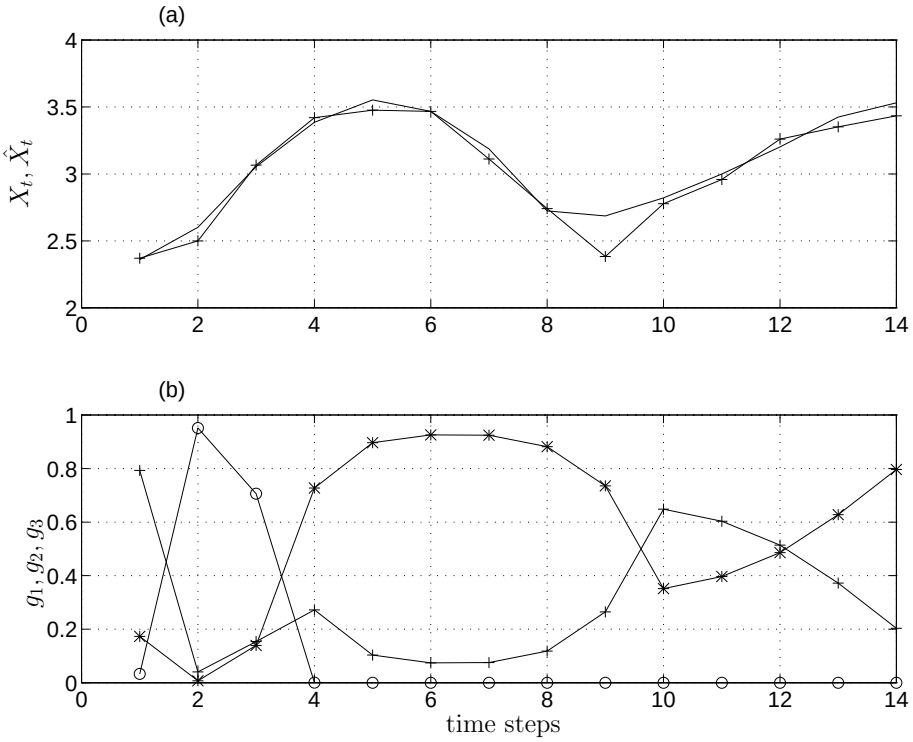

Figure 3: MixAR forecasting characteristics for the 14 point prediction set of the log–transformed Canadian lynx time series. (a) Prediction set and MixAR(3; 6) one step prediction performance (+, line), superimposed on the original series. (b) Probabilities assigned to the three local AR models.

recorded since 1700, the exact underlying mechanism has not yet been accounted for. The rise to peaks tends to be sharper than the subsequent falling pattern, leading to the assumption of a non-linear structure [see, e.g., Tong (1990), §7.3, for a detailed discussion].

The data set was divided into a *training set* (1700-1920) and two prediction sets: *Prediction set* I (1921-1955), and II (1956-1979). In the first stage of the specification of the MixAR model we set the lag size to be $d = 12$, as the series exhibits a clear cycle of about 10-12 years. The smoothed periodogram clearly shows 10^{-1} as the first dominant frequency, and 12^{-1} as the second [cf. Tong (1990, §7.3) and Brockwell and Davis (1991, p. 354) for further details]. Following the model specification procedure outline in Section 8.2, a MixAR(3; 12) model was fitted to the data based on the 221 points in the *training set*; a total of 81 parameters are used in the fitted model.

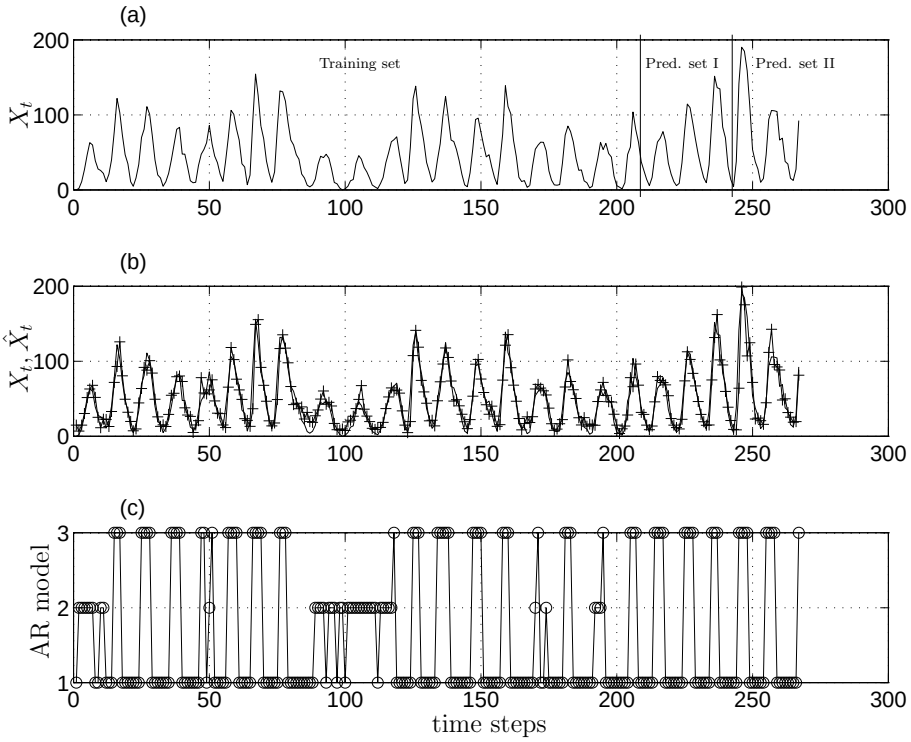

Figure 4: Analysis of the Wölfer number, sunspot time series. (a) Sunspot time series divided into the training set (1-209) and prediction set I (210-243), and II (244-267). (b) MixAR(3; 12) one step prediction (+, line), superimposed on the original data series. (c) Most probable AR model (highest value $\{g_j\}_{j=1}^{3}$).

Figure 4 describes the results obtained for this model; graphs display normalized time scale. The

MixAR$(3, 12)$ one step prediction fits the training set quite well, but accuracy deteriorates as a function of time. This is quite evident in the second prediction set, which is characterized with relatively high variability ($\hat{\sigma}^2 = 2852$) compared to the training set ($\hat{\sigma}^2 = 1174$) and first prediction set ($\hat{\sigma}^2 = 1674$). For the training set, the residual error sequence seems to pass most reasonable tests for randomness (e.g., serial correlation and sign tests). The details are omitted. Figure 5 shows the one-step prediction results and the probabilities assigned to the component AR models for the two prediction sets. Notice in Figure 4(a) the different rate characterizing the rises to, and falls from, local maxima, and the approximate 12 year cyclic behavior. It has been argued by several researchers [see the discussion in Brockwell and Davis (1991), and Tong (1990, §7.3)] that the 'rise' patterns and 'fall' patterns indicate non-linear dynamics. In this context the MixAR model apparently captures this behavior by assigning a linear model to each regime, and using the third for the lower peaks in the series which have a different fall rate. This is evident in Figure 4(c), as well as the results for the *prediction sets* shown in Figure 5. The implication is that the model successfully extrapolates the *training set* patterns to the out-of-sample new observations in the *prediction sets*. This is somewhat surprising given the large number of parameters in the MixAR model relative to the size of the training set (less that 3 observations per fitted parameter). Another obvious result is the decrease in the prediction quality as we compare the error of prediction sets I and II in Figure 5)(a) and (c). Note, that from Figure 5)(b) and (d) it is clear that in the second prediction set only two (out of the three) local AR models are used. This indicates that this set does not exhibit patterns similar to those contained in the *training set* and the first *prediction set*, which were 'assigned' to the third local AR model. Recall also that *prediction set* II is exhibits much more variance compared with the *training set* and *prediction set* I.

To contrast the performance of the MixAR with other models, we quote results that have been reported in the literature for the following models.

(i) Linear model: AR(9) obtained by Subba and Gabr (1984, p. 196) using the AIC criterion.

(ii) Threshold models: TAR(2;4,12) reported by Tong [see, e.g., Tong (1983, p. 241)]. The results appear in Tong and Lim (1980).

(iii) Neural networks: NN(3;12), i.e., a sigmoidal feedforward neural network with 3 sigmoidal units ('hidden layer') and 12 input units. This network realizes the regression function $f_m(\mathbf{x}) = \alpha_0 +$

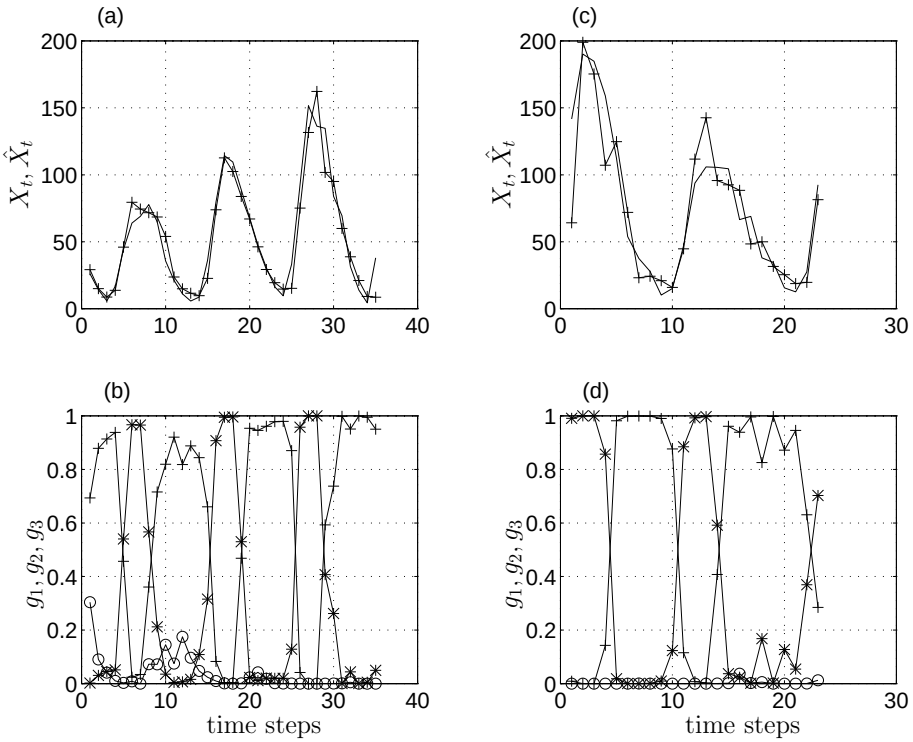

Figure 5: Analysis of the Wölfer number, sunspot prediction set I and II. (a) One-step prediction results for the MixAR(3; 12) model (+, line) superimposed on prediction set I. (b) Probabilities assigned to the component AR models, prediction set I. (c) One-step prediction results for the MixAR(3; 12) model (+, line) superimposed on prediction set II. (d) Probabilities assigned to the component AR models, prediction set II.

$\sum_{k=1}^{3} \alpha_k \sigma(\mathbf{w}_k^T \mathbf{x} + b_k)$, with $\alpha_k, b_k \in \mathbb{R}$ and $\mathbf{w}_k \in \mathbb{R}^{12}$. This architecture was studied in the context of the sunspot series by Weigend *et al.* (1990).

We summarize the results in Table 2, quoting the results for models (ii) and (iii) from the respective references. Detailed results for model (i) where not given in the original reference, therefore we used the model suggested by Subba and Gabr to reproduce the relevant results.

Model	Training set	Prediction set I	Prediction set II
AR$(9, 1)$	0.1705	0.1099	0.1679
TAR$(2; 4, 12)$	0.1268	0.0889	0.1448
NN$(3; 12)$	0.1072	0.0789	0.1837
MixAR$(3; 12)$	0.1190	0.0862	0.1912

Table 2: NMSE for the training set and both prediction sets in the sunspots time series.

The results summarized in Table 2 clearly show that the MixAR performance is competitive with both complex nonlinear models (neural networks), as well as more parsimonious ones (AR and TAR). However, it is also clear that for the second prediction set, the MixAR performance degrades somewhat. It is worth noting, however, that the models we cite in this study were all regularized (except for the MixAR model), so as to improve their out-of-sample performance. The AR was model selected using Akaike's information criterion [cf. Subba and Gabr (1984) for more details]. The TAR model was also judiciously specified and model selected [see Tong (1983, pp. 231-241) for the main ideas, and Tong and Lim (1980) for a more detailed account]. The neural network used by Weigend *et al.* (1990) was trained using cross-validation and the number of non-linear functions in the additive expansion was carefully regularized [cf. §4.1.2 of Weigend *et al.* (1990)]. The authors discuss the issue of over-fitting and ways to avoid it in this class of models, in particular the use of cross validation and a method they term 'weight decay'. In contrast, the MixAR was specified using the procedure outlined in Section 8.2, and no regularization was incorporated in the specification/model selection procedure. Indeed the model that was selected had 81 parameters for a data set of length 221. Despite this, there is no substantial evidence of overfitting in the predictions, in particular in the first prediction set. It seems plausible that more refined model selection procedures will give rise to improved results.

9 Concluding Remarks

The main thrust of this paper is two-fold. First, we set out to describe some basic properties of the MixAR class of models, focusing on stochastic stability, generality, and learning algorithms. The results and proofs expose a basic fact: the analysis of this class of models is greatly simplified, essentially made tractable, by the special structure of MixAR models. Secondly, we investigate, via a preliminary numerical study, whether or not this model is indeed reasonable to apply in practice. The results of the numerical study emphasize again the same key idea; the combination of local linear models via the proposed mixture structure gives rise to an intuitive model that is easy to interpret, and performs well. The results show that, despite the large number of parameters (relative to the short lengths of data sets we consider), MixAR models do not suffer from substantial degradation in out-of-sample performance. The interpretability of MixAR models is a key in the heuristic specification procedure we propose.

The MixAR modeling framework does raise a number of interesting questions and challenges. For a start, we do not currently have a theoretically sound model selection procedure to offer. Rather, we adopted a rather heuristic approach which exploited the special structure of the MixAR model. This method was certainly successful in practice. Nevertheless, one would certainly want to develop more refined, and certainly more automated methods. Moreover, the large number of parameters (relative to the number of data points) will typically preclude asymptotic analysis, e.g., large sample hypothesis testing. At this point we do not have any concrete alternative methodology to offer, and consider this a very important and challenging area for future research. Of course in 'large enough' data sets this issue will not pose a difficulty.

Extensions of the basic model that we have presented here are also clearly of interest. Among these must be MixAR processes with heavy-tailed noise, and MixARMA processes. Other forms of the weighting functions g_j should also be considered, although we feel confident that the model is reasonably robust as far as this is concerned.

A Proofs of Theorems

Proof of Theorem 5.1: The proof follows a standard line of attack for analyzing the stability of

Markov chains over a general state space. An excellent reference, whose results we use throughout the proof, is Meyn and Tweedie (1993). For brevity we will refer to this as MT. Since the proof involves several technical details, we will first outline the main steps and ideas.

Step 1^0: We prove that the Markov chain defined by the MixAR$(m;1)$ process is λ-irreducible (MT, p. 87), where λ denotes Lebesque measure. This assures that the chain enters every set $A \in \mathcal{B}(\mathbb{R})$, with $\lambda(A) > 0$, almost surely in a finite number of steps. It also ensures the existence of a *maximal irreducability measure* ψ, which appears in all the main theorems we quote from MT.

Step 2^0: The proof of geometric ergodicity (and stability in general) relies on finding an appropriate Lyapunov test function, which satisfies the so-called Foster-Lyapunov conditions. Theorem 16.0.2 of MT, involves the right conditions for geometric ergodicity. For the application of the last theorem, we also require that the chain $(X_t)_{t\geq 0}$ is aperiodic, but this is readily seen to hold for the MixAR Markov chain. Note, that condition V1 of MT (p. 190), is sufficient to prove Harris recurrent (MT, p. 200). In that case, an invariant measure exists, and is unique up to constant multiples. However, it may not be finite. An application of the stronger criterion V2 (MT, p. 262) and Theorem 11.0.1 (MT, p. 256), concludes that the invariant measure is finite, and so π exists and is unique. Consequently the chain is ergodic. As it turns out, the conditions needed to ensure the existence and uniqueness of the (finite) invariant measure, are tantamount (in this example) to the conditions ensuring geometric ergodicity.

We now turn to the details of the proof. For completeness, recall the transition kernel of the MixAR$(m;1)$ process:

$$p(X_t \in B | X_{t-1} = x; \boldsymbol{\theta}) = \sum_{j=1}^{m} g_j(\mathbf{x}; \boldsymbol{\theta}_g) \nu(B; \theta_j x + \theta_{j,0}, \sigma_j),$$

with $\boldsymbol{\theta}$ a fixed parameterization, and $B \in \mathcal{B}(\mathbb{R})$. Here, as before, $\nu(B; \mu, \sigma)$ is the probability measure corresponding to the Gaussian distribution with mean μ and variance σ^2. The associated density is denoted $n(x; \mu, \sigma^2)$. For Borel B, define the hitting time

$$\tau_B = \inf\{n \geq 1 : X_n \in B\} \quad .$$

1^0. *Proof of λ-irreducability*: It is clear that

$$P(\tau_B < \infty | X_0 = x) \;\geq\; P(X_1 \in B | X_0 = x) \;>\; 0$$

for Borel B such that $\lambda(B) > 0$, and $x \in \mathbb{R}$. The last inequality is due to the fact that the transition probability kernel possesses an almost everywhere positive density. It follows that the Markov chain $(X_t)_{t\geq 0}$ is λ-irreducible.

2^0. *Proof of geometric ergodicity*: We deviate slightly from the notation in MT, in order to clarify some definitions. From Theorem 16.0.2 (MT, p. 384) it suffices to prove that if for some $\rho \in (0,1)$ there exists a test function $V : \mathbb{R} \to \mathbb{R}^+$, a compact set K, and a finite number M such that

$$(\Delta V)(x) \equiv \mathbb{E}[V(X_1)|X_0 = x] - \rho V(x) \leq \begin{cases} M, & x \in K, \\ -1, & x \in K^c, \end{cases} \tag{17}$$

then π exists, is unique, and in addition $\sup_B |p^n(x,B) - \pi(B)| \to 0$ at a geometric rate. Here Δ is defined formally as $\Delta \equiv P - \rho I$, and it is the right definition of drift for our purposes (note that the definition is slightly different than that used in MT). Note that MT use the concept of *petite sets* to generalize the notion of compact sets. However, the transition kernel of the MixAR implies that it is weak Feller (MT, p. 128), and consequently Proposition 6.2.8 (MT, p. 136) concludes that every compact set is petite. The issue here is the 'right' choice of Lyapunov function V, and corresponding set K.

From the conditions of Theorem 5.1, there must exist two positive real numbers a and b such that

$$-a/b < \theta_1 < 1$$

$$-b/a < \theta_m < 1 \qquad .$$

Let us take

$$V(x) = \begin{cases} ax, & x > 0, \\ -bx, & x \leq 0 \end{cases} .$$

In what follows, let

$$\Phi(z) = \int_{-\infty}^{z} n(u;0,1)du \quad ,$$

the Gaussian CDF. We will evaluate $(\Delta V)(x)$ for $x \in \mathbb{R}^+$ and subsequently for $x \in \mathbb{R}^-$. We have

$$\begin{aligned}
(\Delta V)(x) &= \mathbb{E}[V(X_1)|X_0 = x] - \rho a x \\
&\overset{(a)}{\leq} \frac{a+b}{\sqrt{2\pi}} \sum_j g_j(x;\boldsymbol{\theta}_g)\sigma_j - b\sum_j g_j(x;\boldsymbol{\theta}_g)[\theta_j x + \theta_{j,0}]\Phi\left(-\frac{\theta_j x + \theta_{j,0}}{\sigma_j}\right) \\
&\quad + a\sum_j g_j(x;\boldsymbol{\theta}_g)[\theta_j x + \theta_{j,0}]\left[1 - 2\Phi\left(-\frac{\theta_j x + \theta_{j,0}}{\sigma_j}\right)\right] - a\rho x \\
&\leq C_1 - b\sum_j g_j(x;\boldsymbol{\theta}_g)\theta_j x\Phi\left(-\frac{\theta_j x + \theta_{j,0}}{\sigma_j}\right) \\
&\quad + a\sum_j g_j(x;\boldsymbol{\theta}_g)\theta_j x\left[1 - 2\Phi\left(-\frac{\theta_j x + \theta_{j,0}}{\sigma_j}\right)\right] - a\rho x
\end{aligned}$$

where (a) follows from evaluating $a\int_{\mathbb{R}^+} un(u;\mu,\sigma)du$ and $b\int_{\mathbb{R}^-} un(u;\mu,\sigma)du$ for the m Gaussian components of the mixture, and C_1 is a constant depending on $a, b, \max_j |\sigma_j|$ and $\max_j |\theta_{j,0}|$. Observe that: (1) $g_1(x) \to 1$ for $x \to \infty$, and $g_j(x) \to 0$ for $j > 1$, in particular $|g_j(x)| \leq \exp\{-(\theta_1 - \theta_j)x\}$ thus, $g_j(x)x^k = o(1)$ for all $k \in \mathbb{N}$; (2) $\Phi : \mathbb{R} \cup \{-\infty, \infty\} \to [0,1]$. Fix $\delta \in (1-\rho, 1)$. Then, by (1) and (2) there exists $r > 0$ and an interval $K_1 =: [-r, r] \subset \mathbb{R}$ such that on $K_1^c = [-r,r]^c$ we have $\max_{j>1} |g_j(x)\theta_j x| \leq \delta/m$ and $g_1(x) > 1 - \delta$. There are two cases to consider.

case(i) $\theta_1 > 0$ **:** On K^c we have

$$-b\sum_j g_j(x;\boldsymbol{\theta}_g)\theta_j x\Phi\left(-\frac{\theta_j x + \theta_{j,0}}{\sigma_j}\right) \leq -b\delta - b(1-\delta)\theta_1 x \leq 0$$

thus

$$(\Delta V)(x) \leq C_1 + a\delta + ax\big((1-\delta)\theta_1 - \rho\big)$$

and it is clear that if $\theta_1 < 1$ we can choose r_1 sufficiently large (and $r_1 \geq r$) such that on $K_1^c = [-r_1, r_1]^c$ we have $(\Delta V)(x) \leq -1$ while $(\Delta V)(x)$ is clearly uniformly bounded on K_1.

case(ii) $\theta_1 < 0$ **:** On K^c we have

$$-b\sum_j g_j(x;\boldsymbol{\theta}_g)\theta_j x\Phi\left(-\frac{\theta_j x + \theta_{j,0}}{\sigma_j}\right) \leq -b\theta_1 x$$

while

$$a\sum_j g_j(x;\boldsymbol{\theta}_g)\theta_j x\left[1 - 2\Phi\left(-\frac{\theta_j x + \theta_{j,0}}{\sigma_j}\right)\right] \leq a\delta + (1-\delta)a\theta_1 x \leq a\delta$$

thus

$$(\Delta V)(x) \leq C_1 + a\delta - x\big(b(1-\delta)\theta_1 + a\rho\big)$$

and it is clear that if $\theta_1 > -a/b$ we can choose r_2 sufficiently large (and $r_2 \geq r$) such that on K_2^c we have $(\Delta V)(x) \leq -1$ while $(\Delta V)(x)$ is uniformly bounded on K_2.

Combining the two cases, we conclude that for $-a/b < \theta_1 < 1$ we can choose $r' = \max\{r_1, r_2\}$ such that $K' = [-r', r']$ is the desired compact set, associated with our choice of Lyapunov function V. By symmetry, we can go thru the same arguments for the case of $x \in \mathbb{R}^-$ and get that for $-b/a < \theta_m < 1$ we can choose a compact set K'' such that the Lyapunov stability condition (17) is met. Summarizing, we have established that if $\theta_1 < 1$, $\theta_m < 1$ and $\theta_1\theta_m < 1$, then for some (all) $0 < \rho < 1$ there exists V and $K = K' \cup K''$, such that the Foster–Lyapunov condition (17) is satisfied. This concludes the proof. ■

Proof of Theorem 5.2: As usual, set $\mathbf{Y}_t = [X_t, \ldots, X_{t-d+1}]^T$. That the vectorized Markov chain $\mathbf{Y}_t$ is λ-irreducible follows straightforwardly using the same arguments as in the proof of Theorem 5.1. That it is aperiodic is also easily verified. Motivating the derivations to follow is the choice of the test function $V(\mathbf{y}) = \|\mathbf{y}\|_2$ (the purpose of $V(\cdot)$ is detailed in the proof of Theorem 5.1). Define the following matrices and vectors

$$\mathbf{A}_j \triangleq \begin{bmatrix} \boldsymbol{\theta}_j(1) & \boldsymbol{\theta}_j(2) & \cdots & \boldsymbol{\theta}_j(d-1) & \boldsymbol{\theta}_j(d) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{d\times d}$$

$$\mathbf{D} \triangleq \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{d-1} & \alpha_d \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{d\times d}$$

and $\mathbf{b}_j \equiv [\theta_{j,0}, 0, 0, \ldots, 0]^T \in \mathbb{R}^d$. Here, and in what follows, set

$$\alpha_k \equiv \max_{j=1,2,\ldots,m} |\boldsymbol{\theta}_j(k)|, \qquad k = 1, 2, \ldots, m,$$

where $\boldsymbol{\theta}_j(k)$ is the k-th component of the parameter vector $\boldsymbol{\theta}_j$ (defining the j-th autoregression in the MixAR model). Let $\{\varepsilon_{j,t}\}$ be a family of mutually independent, zero mean, Gaussian variables such that for each $j = 1, 2, \ldots, m$, $\{\varepsilon_{j,t}\}_{t\geq 0}$ are independent identically distributed zero mean Gaussian with variance σ_j^2. Set $\mathbf{Z}_{j,t} \equiv [\varepsilon_{j,t}, 0, 0, \ldots, 0]^T \in \mathbb{R}^d$. Let $\{I_t\}_{t\geq 0}$ be a sequence of mutually independent, integer valued random variables such that, given X_{t-d}^{t-1},

$$I_t = j \;\text{ with probability }\; g_j(X_{t-d}^{t-1}; \boldsymbol{\theta}_g),$$

$j = 1, \ldots, m$. Using these definitions, we can write the MixAR$(m; d)$ process as

$$\mathbf{Y}_t = \mathbf{A}_{I_t} \mathbf{Y}_{t-1} + \mathbf{b}_{I_t} + \mathbf{Z}_{I_t,t}. \tag{18}$$

It is easily verified that $\|A_j\| > 1$ for all j (where $\|\cdot\|$ denotes the spectral matrix norm) and consequently direct application of the stability criterion, as in the proof of Theorem 4.1, is not immediate. However, we can apply a variation of the stability criterion, as in Tjøstheim (1989, Lemma 3.1), that is suited to our setting.

Lemma A.1 (Tjøstheim, 1989) *Let $\mathbf{Y}_t$ be an aperiodic Markov chain, and h a fixed integer. If $(\mathbf{Y}_{th})$ is geometrically ergodic then so is $(\mathbf{Y}_t)$.*

The same implication holds also for the properties of recurrence, transience, and positive Harris recurrence.

Straightforward algebra now yields

$$\begin{aligned}
\mathbf{Y}_{t+h-1} &= \left(\prod_{i=0}^{h-1} \mathbf{A}_{I_{t+i}} \right) \mathbf{Y}_{t-1} + \sum_{i=0}^{h-2} \left(\prod_{j=i+1}^{h-1} \mathbf{A}_{I_{t+j}} \right) \mathbf{b}_{I_{t+i}} \\
&\quad + \sum_{i=0}^{h-2} \left(\prod_{j=i+1}^{h-1} \mathbf{A}_{I_{t+j}} \right) \mathbf{Z}_{I_{t+i},t+i} + \mathbf{b}_{I_{t+h-1}} + \mathbf{Z}_{I_{t+h-1},t+h-1}.
\end{aligned} \tag{19}$$

Taking conditional expectations of $\|\mathbf{Y}_t\|$ and the I_t, and applying the triangle inequality yields

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{Y}_{t+h-1}\| | \mathbf{Y}_{t-1} = \mathbf{y}\right] &\leq \mathbb{E}\left[\left\| \prod_{i=0}^{h-1} \mathbf{A}_{I_{t+i}} \mathbf{Y}_{t-1} \right\| | \mathbf{Y}_{t-1} = \mathbf{y} \right] \\
&\quad + \mathbb{E}\left[\left\| \sum_{i=0}^{h-2} \left(\prod_{j=i+1}^{h-1} \mathbf{A}_{I_{t+j}} \right) \mathbf{b}_{I_{t+i}} \right\| | \mathbf{Y}_{t-1} = \mathbf{y} \right] \\
&\quad + \mathbb{E}\left[\left\| \sum_{i=0}^{h-2} \left(\prod_{j=i+1}^{h-1} \mathbf{A}_{I_{t+j}} \right) \mathbf{Z}_{I_{t+i},t+i} \right\| | \mathbf{Y}_{t-1} = \mathbf{y} \right] \\
&\quad + \mathbb{E}\left[\left\| \mathbf{b}_{I_{t+h-1}} \right\| | \mathbf{Y}_{t-1} = \mathbf{y} \right] + \mathbb{E}\left[\left\| \mathbf{Z}_{I_{t+h-1},t+h-1} \right\| | \mathbf{Y}_{t-1} = \mathbf{y} \right] \\
&\equiv E_1 + E_2 + E_3 + E_4 + E_5
\end{aligned} \tag{20}$$

To evaluate the terms in the above expression we proceed as follows. First, note that

$$E_5 \leq \max_{j=1,2,\ldots,m} \mathbb{E}\left[\|\mathbf{Z}_{j,t}\| \, | \mathbf{Y}_{t-1} = \mathbf{y}\right] \leq \max_{j=1,2,\ldots,m} \sigma_j \equiv \sigma_{\max},$$

and, similarly,

$$E_4 \;\leq\; \max_{j=1,2,\ldots,m} \|\mathbf{b}_j\| \;=\; \max_{j=1,2,\ldots,m} |\theta_{j,0}| \;=\; \alpha_0$$

The following property of matrix norms, the proof of which is temporarily deferred, will be useful.

Proposition A.1 *Let $\{\mathbf{A}_k\}_k \in \mathbb{R}^{d\times d}$ be any finite collection of matrices and let $\mathbf{D} \in \mathbb{R}^{d\times d}$ be such that $|(\mathbf{A}_k)_{ij}| \leq (D)_{ij}$ for $i, j, = 1, 2, \ldots, d$ and for all k. Then*

$$\left\| \prod_{k=1}^{m} \mathbf{A}_k \right\| \leq \|D^m\|,$$

where $\|\cdot\|$ is the spectral matrix norm.

We now evaluate the remaining three terms E_1, E_2, E_3. A straightforward application of Proposition A.1 gives

$$\begin{aligned} E_1 &\leq \mathbb{E}\left[\left\| \prod_{i=0}^{h-1} \mathbf{A}_{I_{t+i}} \right\| \|\mathbf{Y}_{t-1}\| | \mathbf{Y}_{t-1} = \mathbf{y} \right] \\ &\leq \left\|\mathbf{D}^h\right\| \|\mathbf{y}\|, \end{aligned}$$

and

$$\begin{aligned} E_3 &\overset{(a)}{\leq} \mathbb{E}\left[\sum_{i=0}^{h-2} \left\| \prod_{j=i+1}^{h-1} \mathbf{A}_{I_{t+j}} \right\| \|\mathbf{Z}_{I_{t+i},t+i}\| | \mathbf{Y}_{t-1} = \mathbf{y} \right] \\ &\leq \sum_{i=0}^{h-2} \left\|D^{h-1-i}\right\| \mathbb{E}\left[\|\mathbf{Z}_{I_{t+i},t+i}\| | \mathbf{Y}_{t-1} = \mathbf{y} \right] \\ &\overset{(b)}{\leq} \sum_{i=0}^{h-2} \left\|D^{h-1-i}\right\| \sigma_{\max} \\ &\equiv C_3(\boldsymbol{\theta}, h) \end{aligned}$$

where (a) follows from Minkowski's inequality, and (b) follows from the the bound on E_5. Here $C(\boldsymbol{\theta}, h)$ is a constant depending only on the parameters of the MixAR model $\boldsymbol{\theta}$ and on h. The same steps applied to E_2 yield

$$E_2 \;\leq\; \sum_{i=0}^{h-2} \left\|D^{h-1-i}\right\| \alpha_0 \;\equiv\; C_2(\boldsymbol{\theta}, h).$$

To finish the proof observe that

$$\det\left(z\mathbf{I} - \mathbf{D}\right) = \mathcal{P}(z),$$

and by assumption all roots of $\mathcal{P}(z)$ are less than 1 in modulus. Thus, $|\lambda_{\max}(\mathbf{D})| < 1$, where $\lambda_{\max}(\mathbf{D})$ is the largest eigenvalue of $\mathbf{D}$ in modulus. Consequently, since $\lim_h \|\mathbf{D}^h\|^{1/h} \to \lambda_{\max}(\mathbf{D})$, there exists a $\rho < 1$ and $h(\rho)$ such that $\|\mathbf{D}^{h(\rho)}\| \leq \rho$. This implies that for the above choice of ρ and $h(\rho)$ we have

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{Y}_{t+h-1}\| | \mathbf{Y}_{t-1} = \mathbf{y}\right] &\leq \left\|\mathbf{D}^h\right\| \|\mathbf{y}\| + C_2 + C_3 + \alpha_0 + \sigma_{\max} \\ &\leq \rho\|\mathbf{y}\| + C(\boldsymbol{\theta}, h, \rho) \quad . \end{aligned}$$

To apply the stability criterion for geometric ergodicity in Theorem 4.1, let $V(\mathbf{y}) = \|\mathbf{y}\|$, and $K = \{\mathbf{y} : \|\mathbf{y}\| \leq 2C/(1-\rho)\}$. Consequently, on K^c, we have

$$\mathbb{E}\left[\|\mathbf{Y}_{t+h-1}\| | \mathbf{Y}_{t-1} = \mathbf{y}\right] \;\leq\; \rho\|\mathbf{y}\| + C \;\leq\; \frac{\rho+1}{2}\|\mathbf{y}\| \tag{21}$$

by choice of K^c. Since $(\rho+1)/2 < 1$, the proof is complete. ■

Remark A.1 Tjøstheim (1989, Theorem 4.5) gives sufficient conditions for the stability of a multivariate TAR($m; d$) model [for further discussion see Tong (1990)]. In spite of the apparent similarity in the formulation of the MixAR($m; d$) in (18) and the TAR($m; d$) model, we cannot apply Tjøstheim's result, since the k-cycle assumption made there is not plausible in the MixAR setting.

We now prove Proposition A.1. Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{D}$ be such that $|A_{ij}| \leq D_{ij}$ (in all cases we refer to the i, j elements in the matrices in question). Denote $|\mathbf{y}| = [|y_1|, |y_2|, \ldots, |y_d|]^T$. Then

$$\|\mathbf{A}\| \;\equiv\; \sup_{\mathbf{y}\neq 0} \frac{\|\mathbf{A}\mathbf{y}\|}{\|\mathbf{y}\|} \overset{(a)}{\leq} \sup_{\mathbf{y}\neq 0} \frac{\||D|\mathbf{y}|\|}{\|\mathbf{y}\|} \overset{(b)}{=} \|D\|$$

where (a) follows since $|\mathbf{A}\mathbf{y}| \leq \mathbf{D}|\mathbf{y}|$ by the triangle inequality and (b) follows from the fact that the matrix $\mathbf{D}$ has all elements non-negative, thus allowing us to restrict attention to vectors $\mathbf{y}$ with non-negative elements in taking the supremum.

Now, let $A, B \in \mathbb{R}^{d\times d}$ and D be such that $|A_{ij}|, |B_{ij}| \leq D_{ij}$. Then

$$|(AB)_{ij}| \;=\; \sum_{k=1}^{d} (A)_{ik}(B)_{kj} \;\leq\; \sum_{k=1}^{d} |A_{ik}||B_{kj}| \;\leq\; \sum_{k=1}^{d} (D)_{ik}(D)_{kj} \;=\; (D^2)_{ij}$$

and using the previous result we have that $\|AB\| \leq \|D^2\|$. Repeated application of the above concludes the proof. ■

Proof of Theorem 6.1: The proof proceeds in three steps, somewhat condensed to avoid introducing machinery and notation that are beyond the scope and purpose of this paper.

1^0: The first step establishes that $f \in L^q(\mathbb{R}^d, \mu)$, where μ is the stationary distribution of X_{t-d}^{t-1}.

$$\begin{aligned}\int |f(\mathbf{x})|^q \mu(d\mathbf{x}) &= \int |\mathbb{E}[X_t | X_{t-d}^{t-1} = \mathbf{x}]|^q \mu(d\mathbf{x}) \\ &\leq \int E\left[|X_t|^q \; | X_{t-d}^{t-1} = \mathbf{x}\right] \mu(d\mathbf{x}) \\ &= \mathbb{E}|X_t|^q,\end{aligned}$$

which is finite by assumption.

2^0. The second step is to show that the model prediction function $f_m(\mathbf{x}; \boldsymbol{\theta}) = \sum_j g_j(\mathbf{x}; \boldsymbol{\theta}_g)[\boldsymbol{\theta}_j^T \mathbf{x} + \theta_{j,0}]$ is in the vector space $L^q(\mathbb{R}^d, \mu)$.

$$\begin{aligned}\int |f_n(\mathbf{x}; \boldsymbol{\theta})|^q \mu(d\mathbf{x}) &= \int \left| \sum_{j=1}^m g_j(\mathbf{x}; \boldsymbol{\theta}_g)[\boldsymbol{\theta}_j^T \mathbf{x} + \theta_{j,0}] \right|^q \mu(d\mathbf{x}) \\ &\leq \int \sum_{j=1}^m g_j(\mathbf{x}; \boldsymbol{\theta}_g) |\boldsymbol{\theta}_j^T \mathbf{x} + \theta_{j,0}|^q \mu(d\mathbf{x}) \\ &\leq 2^{q-1} \int \sum_{j=1}^m g_j(\mathbf{x}; \boldsymbol{\theta}_g) |\boldsymbol{\theta}_j^T \mathbf{x}|^q \mu(d\mathbf{x}) + 2^{q-1} \int \sum_{j=1}^m g_j(\mathbf{x}; \boldsymbol{\theta}_g) |\theta_{j,0}|^q \mu(d\mathbf{x}) \\ &\leq 2^{q-1} \int \sum_{j=1}^m g_j(\mathbf{x}; \boldsymbol{\theta}_g) \|\boldsymbol{\theta}_j\|^q \|\mathbf{x}\|^q \mu(d\mathbf{x}) + 2^{q-1} \max_j |\theta_{j,0}|^q \\ &\leq 2^{q-1} \max_j \|\boldsymbol{\theta}_j\|^q \int \|\mathbf{x}\|^q \mu(d\mathbf{x}) + c_1 \\ &\leq 2^{q-1} c_2 d^q E|X_t|^q + c_1,\end{aligned}$$

and the last line is finite by the moment assumption on X_t and since the parameters take values in compact sets. For an explicit identification of these compact sets the reader is referred to Zeevi *et al.* (1998).

3^0. The third step is an application of Theorem A.1 of Zeevi *et al.* (1998), adapted to the above functional class. A straightforward adaptation of Theorem A.1 in this paper gives us the following lemma.

Lemma A.2 *Let $K \subset \mathbb{R}^d$ be compact. For any $f \in L^q(K, \lambda)$, and any $\epsilon > 0$, there exists an m sufficiently large (depending on f and ϵ), and a corresponding function $f_m^*(\mathbf{x}; \boldsymbol{\theta}) = \sum_{j=1}^m \alpha_j g_j(\boldsymbol{\theta}_g; \mathbf{x})$, so that*

$$\|f - f_m^*\|_{L^q([-1,1]^d, \lambda)} < \epsilon,$$

where λ denotes the Lebesgue measure on $\mathbb{R}^d$.

Note that a consequence of the proof is that it actually suffices to work with a MixAR with locally constant components; i.e. AR models that have only a constant term. Although this is hardly very interesting statistically, a corollary of the lemma is that the coefficients $\boldsymbol{\theta}_j$ may be chosen with arbitrary restrictions without affecting the inherent approximation capacity of the model prediction function f_m. Now, since $f, f_m \in L^q(\mathbb{R}^d, \mu)$ there exists a $T \in \mathbb{R}$ such that

$$\int_{\mathbb{R}^d \setminus [-T,T]^d} |f - f_m|^q d\mu < \frac{\epsilon}{2},$$

and, since μ has a bounded continuous density function,

$$\int_{[-T,T]^d} |f - f_m|^q d\mu \leq C \int_{[-T,T]^d} |f - f_m|^q d\lambda,$$

which subsequently can be made smaller than $\epsilon/2$ (for m sufficiently large) following Lemma A.2. ■

Sketch of Proof for Proposition 7.1: The key here is an application of the so-called global convergence theorem (cf. Luenberger, 1973, p. 125). Since both $Q(\cdot; \boldsymbol{\theta})$ and $Q(\boldsymbol{\theta}; \cdot)$ are continuous for all $\boldsymbol{\theta} \in \Theta$, and by the basic properties of the steepest ascent algorithm [cf. Luenberger (1973), ch. 7.6] we have, using Corollary 2 of Luenberger (1973, p. 125), that the composite algorithm given in 7.1 is a closed point-to-point mapping. Thus, by Theorem 1 and Theorem 4 of (Wu, 1983) we have the result. ■

B Specification of Models in Section 8

In this section we present the full specification of the models used in Sections 8.4 and 8.5.

B.1 Models Used for Canadian Lynx Time Series

For each model used in Section 8.4 we give the values of all estimated parameters. In all cases these are quoted from the respective sources, excluding of course the MixAR model.

1. AR(12): This model was fitted using the AICC (corrected Akaike's Information Criterion) by

Brockwell and Davis (1990, p. 550). They obtain

$$X_t = 1.123 + 1.084X_{t-1} - 0.477X_{t-2} + 0.265X_{t-3} - 0.218X_{t-4} + 0.180X_{t-9} - 0.224X_{t-12} + \varepsilon_t,$$

with $\mathbb{E}\varepsilon_t^2 = 0.0396$.

2. TAR(2;8,3): This model was fitted by Tong and Lim (1980). We quote the estimated model from Tong (1983, p. 190).

$$X_t = \begin{cases} 0.5240 + 1.0360X_{t-1} - 0.1760X_{t-2} + 0.1750X_{t-3} - 0.4340X_{t-4} + 0.3460X_{t-5} \\ \quad - 0.3030X_{t-6} + 0.2170X_{t-7} + 0.0040X_{t-8} + \varepsilon_t^{(1)} & \text{if } X_t \leq 3.1160 \\ 2.6560 + 1.4250X_{t-1} - 1.1620X_{t-2} - 0.1090X_{t-3} + \varepsilon_t^{(2)} & \text{if } X_t > 3.1160 \end{cases}$$

where the $\mathrm{Var}[\varepsilon_t^{(1)}] = 0.0255$ and $\mathrm{Var}[\varepsilon_t^{(2)}] = 0.0516$.

3. A bilinear model studied by Subba and Gabr [see, e.g., Subba and Gabr (1984, p. 204)]. This model takes the general form

$$\begin{aligned} &X_t + a_1X_{t-1} + a_2X_{t-2} + a_3X_{t-3} + a_4X_{t-4} + a_9X_{t-5} + a_{12}X_{t-12} \\ &\quad = a_0 + b_{3,9}X_{t-3}\varepsilon_{t-9} + b_{9,9}X_{t-9}\varepsilon_{t-9} + b_{6,2}X_{t-6}\varepsilon_{t-2} + b_{1,1}X_{t-1}\varepsilon_{t-1} \\ &\qquad + b_{2,7}X_{t-2}\varepsilon_{t-7} + b_{4,2}X_{t-4}\varepsilon_{t-2} + \varepsilon_t \end{aligned}$$

With fitted parameters (quoted from Subba and Gabr's study):

$a_1 = -0.7728$	$a_2 = 0.0916$	$a_3 = -0.0831$	$a_4 = 0.2615$
$a_4 = 0.2615$	$a_9 = -0.2256$	$a_{12} = 0.2458$	$a_0 = -1.4863$
$b_{3,9} = -0.7893$	$b_{9,9} = 0.4798$	$b_{6,2} = 0.3902$	$b_{1,1} = 0.1326$
$b_{2,7} = 0.0794$	$b_{4,2} = -0.3212$		

4. A random coefficient AR(2) studied by Nicholls and Quinn [see, e.g., Nicholls and Quinn (1982), p. 143]. We quote the parameters from their study.

$$X_t = 2.8802 + (1.4132 + Z_t^{(1)})(X_{t-1} - 2.8802) + (-0.7942 + Z_t^{(2)})(X_{t-1} - 2.8802) + \varepsilon_t$$

where $\{Z_t^{(1)}, Z_t^{(1)}\}$ are i.i.d. multivariate zero mean Gaussian r.v.'s independent of all other r.v.'s, with

$$\mathbb{E}[Z_t^{(1)}, Z_t^{(1)}][Z_t^{(1)}, Z_t^{(1)}]^T = \begin{bmatrix} 0.0701 & -0.0406 \\ -0.0406 & 0.0492 \end{bmatrix}$$

and $\mathbb{E}\varepsilon_t^2 = 0.0391$.

5. MixAR(3;6), fitted via the procedure discussed in Section 8.2, and using Algorithm 7.1. Recall that we had normalized the log-lynx time series to the unit interval using a translation and re-scaling. The fitted model was

$$X_t = \begin{cases} 1.1491X_{t-1} - 0.3656X_{t-2} + 0.2932X_{t-3} - 0.4552X_{t-4} + 0.4189X_{t-5} \\ \quad - 0.1683X_{t-6} + 0.1441 + \varepsilon_t^{(1)} & \text{w.p. } g_1(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \\ 0.5283X_{t-1} + 0.9234X_{t-2} - 1.1189X_{t-3} + 0.5174X_{t-4} - 0.6379X_{t-5} \\ \quad + 0.3086X_{t-6} + 0.1929 + \varepsilon_t^{(2)} & \text{w.p. } g_2(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \\ 1.3208X_{t-1} - 0.5921X_{t-2} + 0.0661X_{t-3} - 0.5637X_{t-4} + 0.1120X_{t-5} \\ \quad + 0.1578X_{t-6} + 0.2841 + \varepsilon_t^{(3)} & \text{w.p. } g_3(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \end{cases}$$

with the estimated parameters for the multinomial logit function g_j being

		$t-1$	$t-2$	$t-3$	$t-4$	$t-5$	$t-6$
$j=1$:	0.8784	1.7511	4.6790	10.3615	0.3490	3.9757	-5.5212
$j=2$:	1.1866	-4.5887	-10.3318	-17.8458	-9.8066	-5.7954	12.1835
$j=3$:	-1.5152	1.8276	5.3746	6.9576	8.6806	2.3046	-5.8366

where the displayed values are for $b_j, a_{j,1}, \ldots, a_{j,6}$ for $j = 1, 2, 3$. Recall

$$g_j(\mathbf{x}, \boldsymbol{\theta}_g) \equiv \frac{\exp(\mathbf{a}_j^T \mathbf{x} + b_j)}{\sum_{k=1}^m \exp(\mathbf{a}_k^T \mathbf{x} + b_k)}, \qquad j = 1, \ldots, m \quad .$$

The variances associated with the local AR models were respectively: $0.0051, 0.0015$, and 0.0340 (these numbers have already been scaled up to correct for the transformation to $[0, 1]$).

B.2 Models Used for Sunspots Time Series

For each model used in Section 8.4 we give the values of all estimated parameters. In all cases these are quoted from the respective sources, excluding the case of the MixAR model.

1. AR(9): This model was fitted using the AIC (Akaike's Information Criterion) by Subba and Gabr (1984, p. 196). They obtain

$$X_t = 8.5086 + 1.2163X_{t-1} - 0.4670X_{t-2} - 0.1416X_{t-3} + 0.1691X_{t-4}$$
$$-0.1473X_{t-5} + 0.0543X_{t-6} - 0.0534X_{t-7} + 0.0667X_{t-8} + 0.1129X_{t-9} + \varepsilon_t,$$

with $\mathbb{E}\varepsilon_t^2 = 199.27$.

2. TAR(2;4,12): This model was fitted by Tong and Lim (1980)

$$X_t = \begin{cases} 10.544 + 1.6920X_{t-1} - 1.1592X_{t-2} + 0.2367X_{t-3} + 0.1503X_{t-4} + \varepsilon_t^{(1)} & \text{if } X_t \le 36.6 \\ 7.8041 + 0.7432X_{t-1} - 0.0409X_{t-2} - 0.2020X_{t-3} + 0.1730X_{t-4} - 0.2266X_{t-5} \\ \quad + 0.0189X_{t-6} + 0.1612X_{t-7} - 0.2560X_{t-8} + 0.3190X_{t-9} - 0.3891X_{t-10} \\ \quad + 0.4306X_{t-11} - 0.3970X_{t-12} + \varepsilon_t^{(2)} & \text{if } X_t > 36.6 \end{cases}$$

where the $\text{Var}\varepsilon_t^{(1)} = 254.64$ and $\text{Var}\varepsilon_t^{(2)} = 66.8$.

3. NN(3;12) fitted by Weigend *et al.* (1990). Recall that the neural network realizes the additive expansion

$$f_m(\mathbf{x}) = \alpha_0 + \sum_{k=1}^{3} \alpha_k \sigma(\mathbf{w}_k^T \mathbf{x} + b_k) \quad ,$$

with $\alpha_k, b_k \in \mathbb{R}$ and $\mathbf{w}_k \in \mathbb{R}^{12}$. In their analysis Weigend *et al.* (1990) normalize the sunspots series to the unit interval by dividing all values by 191.2. Their estimated parameters are: $\alpha_0 = 0.798$, $\alpha_1 = -1.565$, $\alpha_2 = 2.247$ and $\alpha_3 = -1.599$. The corresponding values of b_j's were $b_1 = -0.858$, $b_2 = -1.960$ and $b_3 = -0.512$. The values for the vectors $\mathbf{w}_j$ are as follows (for $j = 1, 2, 3$)

$t-1$	$t-2$	$t-3$	$t-4$	$t-5$	$t-6$	$t-7$	$t-8$	$t-9$	$t-10$	$t-11$	$t-12$
0.114	-0.200	0.317	0.030	0.500	-0.129	-0.255	-0.060	-1.101	-0.328	-0.646	0.153
0.814	-3.103	-0.995	-0.168	0.533	0.000	0.414	1.078	-0.048	-0.039	-0.094	-0.205
-4.010	-0.362	-0.293	-0.160	-0.208	0.215	0.198	0.869	0.703	0.130	0.080	0.000

The estimated (scaled up) variance was: $\sigma^2 = 125.87$.

4. A MixaR(3;12): Recall that we had normalized the time series (by dividing the values by 191.2) so that values lie in the unit interval. The parameters were fitted using Algorithm 7.1.

$$X_t = \begin{cases} 0.7755X_{t-1} + 0.0021X_{t-2} - 0.1744X_{t-3} + 0.1158X_{t-4} \\ \quad - 0.1329X_{t-5} - 0.0697X_{t-6} + 0.1973X_{t-7} - 0.1634X_{t-8} + 0.1423X_{t-9} \\ \quad\quad - 0.1330X_{t-10} + 0.0979X_{t-11} + 0.0584X_{t-12} + 0.0319 + \varepsilon_t^{(1)} & \text{w.p.} \quad g_1(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \\ 0.9410X_{t-1} + 0.5574X_{t-2} - 1.4491X_{t-3} + 0.8714X_{t-4} \\ \quad - 0.6319X_{t-5} - 0.0716X_{t-6} + 0.5728X_{t-7} - 0.3843X_{t-8} - 0.1595X_{t-9} \\ \quad\quad + 0.1628X_{t-10} + 0.0264X_{t-11} + 0.0423X_{t-12} + 0.0931 + \varepsilon_t^{(2)} & \text{w.p.} \quad g_2(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \\ 1.0874X_{t-1} - 0.3137X_{t-2} - 0.7351X_{t-3} + 0.6386X_{t-4} \\ \quad - 0.1347X_{t-5} + 0.4621X_{t-6} - 0.6249X_{t-7} + 0.7624X_{t-8} - 0.3772X_{t-9} \\ \quad\quad - 0.1825X_{t-10} + 0.5235X_{t-11} - 0.2524X_{t-12} + 0.1276 + \varepsilon_t^{(3)} & \text{w.p.} \quad g_3(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \end{cases}$$

The estimated parameters for the multinomial logit function g_j were

		$t-1$	$t-2$	$t-3$	$t-4$	$t-5$	$t-6$
$j=1$:	-0.3244	-8.4759	12.3953	-0.0876	1.0168	-0.1975	0.0185
$j=2$:	4.5782	-6.4894	-6.7294	1.0699	4.1966	-2.6170	-2.7179
$j=3$:	-2.2408	15.8723	-3.3894	-1.0070	-4.7607	3.7900	1.8148
	$t-7$	$t-8$	$t-9$	$t-10$	$t-11$	$t-12$	
$j=1$:	5.6104	1.3044	-3.5242	3.8971	-2.7666	3.5431	
$j=2$:	-5.6577	-3.6030	2.3371	-10.2604	9.1528	0.1161	
$j=3$:	1.0747	1.3794	0.7436	6.4400	-4.6413	-1.6831	

where the displayed values are for $b_j, a_{j,1}, \ldots, a_{j,6}$ for $j = 1, 2, 3$. Recall

$$g_j(\mathbf{x}, \boldsymbol{\theta}_g) \equiv \frac{\exp(\mathbf{a}_j^T \mathbf{x} + b_j)}{\sum_{k=1}^{m} \exp(\mathbf{a}_k^T \mathbf{x} + b_k)}, \qquad j = 1, \ldots, m \quad .$$

The variances associated with the local AR models were respectively: $80.6112, 7.3315$, and 315.6292 (these numbers have already been scaled up accordingly to correct for the transformation to $[0, 1]$).

References

[1] Anthony, M. and Bartlett, P.L. (1999) *Learning in Neural Networks: Theoretical Foundations*, C.U.P, Cambridge.

[2] Billingsley, P. (1961) *Statistical Inference for Markov Processes*, Holt, New York.

[3] Bartlett, P.L. (1998) The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is more important than the Size of the Network. *IEEE Trans. Inform. Theory*, vol. 44, 525–536.

[4] Bartlett, P.L., Long, Philip M. and Williamson, R.C. (1996) Fat-shattering and the Learnability of Real-valued Functions . *J. Comput. System Sci.* vol 52, 434–452.

[5] Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

[6] Breiman, L. (1998) Arcing Classifiers. With Discussion and a Rejoinder by the Author. *Ann. Statist.* vol. 26, 801–849.

[7] Brockwell, P.J. and Davis, R.A. (1991) *Time Series: Theory and Methods*, Second Edition, Springer Verlag, New York.

[8] Chan, K. (1988) On the Existence of the Stationary and Ergodic NEAR(p) Model , *J. of Time Ser. Anal.*, vol. 9, 319-328.

[9] Chen, R. and Tsay, R. (1993) Functional Coefficient Autoregressive Models, *JASA*, vol. 88, 298-308.

[10] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. Roy. Statist. Soc.*, vol. B39, 1-38.

[11] Freund, Y. and Schapire, R. (1996) Experiments with a new boosting algorithm, *Proc. of the Thirteeth Int. Conf. on Machine Learning*, 148-156.

[12] Friedman, J.H., Hastie, T. and Tibshirani, R. "Additive Logistic Regression: a Statistical View of Boosting". (1998) Preprint. http://www-stat.stanford.edu/ jhf/#reports

[13] Granger, C.W.J. and Teräsvirta, T. (1993) *Modeling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

[14] Haggan, V. and Ozaki, T. (1981) Modeling Nonlinear Vibrations Using an Amplitude Dependent Autoregressive Time Series, *Biometrica*, vol. 68, 189-196.

[15] Hamilton, J.D. (1990) Analysis of Time Series Subject to Changes in Regime, *Journal of Econometrics*, vol. 45, 39-70.

[16] Jacobs, R.A., Jordan, M.I., Nowlan, S.J. and Hinton, G.E. (1991) Adaptive Mixtures of Local Experts, *Neural Computation*, vol. 3, 79-87.

[17] Jiang, W. and Tanner, M. A. (1999). On the Identifiability of Mixtures-of-Experts. *Neural Networks*, vol. 12, 1253–1258.

[18] Jiang, W. and Tanner, M. (2000) On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models, *IEEE Trans. on Inform. Theory*, to appear.

[19] Jordan, M. and Jacobs, R. (1994) Hierarchical Mixtures of Experts and the EM Algorithm, *Neural Computation*, vol. 6, 181-214.

[20] Jordan, M.I. and Xu, L. (1995) Convergence Results for the EM Approach to Mixtures of Experts Architectures, *Neural Networks*, vol. 8, 1409-1431.

[21] Lee, W.S., Bartlett, P.L. and Williamson, R.C. (1998) The Importance of Convexity in Learning with Squared Loss. *IEEE Trans. Inform. Theory*, vol 44, 1974–1980.

[22] Lewis, P.A.W. and Stevens, J.G. (1991) Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS), *JASA*, vol. 86, 864-877.

[23] Luenberger, D.G. (1973) *Introduction to Linear and Nonlinear Programming*, Addison-Wesley Publishers, Massachusetts.

[24] Meyn, S.P. and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

[25] Nicholls, D. F. and Quinn, B. G. (1982). *Random coefficient autoregressive models: an introduction.* Lecture Notes in Statistics, vol. 11, Springer-Verlag, New York.

[26] Peng, F., Jacobs, R.A. and Tanner, M.A. (1996) Bayesian Inference in Mixtures of Experts Models With an Application to Speech Recognition, *JASA*, vol. 91, 953-960.

[27] Priestley M.B. (1988) *Non-linear and Non-stationary Time Series Analysis*, Academic Press, New York.

[28] Redner, R.A. and Walker, H.F. (1984) Mixture Densities, Maximum Likelihood and the EM Algorithm , *SIAM Review*, vol. 26, 195-239.

[29] Schapire, R.E., Freund, Y., Bartlett, P. and Lee, W.S. (1998) Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. *Ann. Statist.* vol. 26, 1651–1686.

[30] Subba Rao, T. and Gabr, M. M. *An Introduction to Bispectral Analysis and Bilinear Time Series models*. Lecture Notes in Statistics, vol. 24, Springer-Verlag, New York.

[31] Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions*. Wiley.

[32] Tjøstheim, D. (1986) Some Doubly Stochastic Time Series models, *Jour. of Time Ser. Analysis*, vol. 7, 51-72.

[33] Tjøstheim, D. (1989) Non-Linear Time Series and Markov Chains, *Adv. Appl. Prob.*, 22, 587-611.

[34] Tjøstheim, D. (1994) Nonlinear Time Series: A Selective Review , *Scand. J. of Stat.*, vol. 21, 97-130.

[35] Tong, H. (1983) *Threshold Models in Non-linear Time Series Analysis*, Springer Verlag, New York.

[36] Tong, H. (1990) *Nonlinear Time Series: A Dynamical Systems Approach*, Oxford University Press, Oxford.

[37] Tukey, J.W. (1961) Discussion emphasizing the connection between analysis of variance and spectrum analysis, *Technometrics*, vol. 3, 191–219.

[38] Waterhouse, S.R. and Robinson, A.J. (1994) Non-linear Prediction of Acoustic Vectors Using Hierarchical Mixtures of Experts, in *Neural Information Processing Systems* 7, Morgan Kaufmann.

[39] Weigend, A.S. and Gershenfeld, N.A. (1994) *Time Series Prediction: Forecasting the Future and Understanding the Past*, Santa Fe Institute Studies in the Science of Complexity, Proc. Vol. XV. Reading, MA: Addison-Wesley.

[40] Weigend, A.S Huberman, B.A. and Rumelhart, D.E. (1990) Predicting the Future: A Connectionist Approach, *Intl. Jour. of Neural Systems*, vol. 1, 193-209.

[41] Wu, C.F.J., (1983) On the Convergence Properties of the EM Algorithm, *Ann. Statist.*, vol. 11, 95-103.

[42] Zeevi, A.J., Meir, R. and Maiorov, V. (1998) Error Bounds for Functional Approximation and Estimation Using Mixtures of Experts, *IEEE Trans. on Inform. Theory*, vol. 44, 1010–1025.