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

Assaf J. Zeevi * Ron Meir† Robert J. Adler‡

October 24, 2000

Abstract

We consider a new class of non-linear time series models based on mixtures of autoregressions, which we call MixAR models. These have the property that, at any given point in time, one of a number of alternative AR models describes the dynamics and the driving AR model is randomly selected from a set of m possible models via a state (lag vector) dependent probability distribution. Thus a MixAR process is a Markov chain whose transition kernel is a mixture distribution with state dependent weights.

We study the flexibility of this model in a prediction setting, and examine issues of stochastic stability. We also discuss parameter estimation algorithms, asymptotics and examples.

KEY WORDS: Mixtures of autoregressions, non-linear models, prediction, mixtures of distributions, EM algorithm, Markov chains, stochastic stability.
AMS 1991 Subject Classifications: Primary 62M10, 62M45; Secondary 60G10.

*Information Systems Lab, Stanford University, Stanford CA. 94305; assaf@isl.stanford.edu Research supported in part by NSF.

†Faculty of Electrical Engineering, Technion, Haifa 32000, Israel. rmeir@dumbo.technion.ac.il Research supported in part by the Israel Science Foundation and Ollendorff Center of the Faculty of Electrical Engineering at the Technion.

‡Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel. robert@ieadler.technion.ac.il Research supported in part by NSF and the Israel Science Foundation.

1 Introduction

We shall be interested in a specific class of models for non-linear time series, which, for reasons that will soon be obvious, we call "MixAR models" and which are defined as follows:

Let $\{X_t\}_{t\in\mathbb{N}}$ be a real valued time series. Fix an integer $d \geq 1$ and write X_{t-d}^{t-1} for the lag vector $(X_{t-1}, \ldots, X_{t-d})$. Fix another integer $m \geq 1$, and let $\{g_j(\cdot;\boldsymbol{\theta}_g)\}_{k=1}^m$ be a collection of non-negative functions from $\mathbb{R}^d \to [0,1]$ so that $\sum_{k=1}^m g_j(\mathbf{x};\boldsymbol{\theta}_g) = 1$ for all $\mathbf{x} \in \mathbb{R}^d$. $\boldsymbol{\theta}_g$ is a parameter to be interpreted later.

Then X_t is said to be a MixAR (mixture of autoregressions) process with lag d and m components (MixAR(m,d)) if it is Markov of order d (i.e. $\mathbb{P}(X_t \in B|X_{t-1}, X_{t-2}, \ldots) = \mathbb{P}(X_t \in B|X_{t-d}^{t-1})$ for all Borel B and all t) and

$$X_t \;=\; \boldsymbol{\theta}_j^T X_{t-d}^{t-1} \;+\; \theta_{k,0} \;+\; \varepsilon_t^{(k)} \qquad \text{with probability } \; g_j(X_{t-d}^{t-1};\boldsymbol{\theta}_g), \tag{1.1}$$

where the $\boldsymbol{\theta}_j$ are d-dimensional parameter vectors, the $\theta_{k,0}$ are real-valued parameters and $\{\{\varepsilon_t^{(k)}\}_{t\in\mathbb{N}}\}_{k=1}^m$ is a collection of mutually independent i.i.d., zero mean, noise processes with (possibly) different variances $\{\sigma_j^2\}_{k=1}^m$.

Mixture models of this kind were first introduced in the neural network literature, and in the context of non-linear regression in [14, 17], and were called *mixtures of experts*. Note that the terminology is somewhat non-standard, since the dependence of the weights g_j on the data means that these are not mixtures in the standard sense of the word (e.g. [32]). Early applications of mixture models of this kind included the prediction of acoustic vectors [38] and speech recognition [26]. Their mathematical properties are beginning to be seriously studied, as in [15, 44].

If $m = 1$ in (1.1), then we have the familiar AR(d) autoregressive process. If the g_j are constant, rather than functions of X_{t-d}^{t-1}, then we recover the mixture transition distribution (MTD) of Raftery [28], which has recently been studied in a time series setting in [19] and [41]. These are models in which the series evolves according to one of m models and the choice of model by which X_{t-d}^{t-1} determines X_t is made at random, the choices at different times being independent of one another and the prior history of the series.

If the functions g_j are allowed to take only the values 0 or 1, so that they become indicator functions for certain regions of $\mathbb{R}^d$, then we recover the threshold autoregressive (TAR) models of Tong (e.g. [36, 37]). These, again, are models in which the series evolves according to one of m models, but here the choice of model at time t is made deterministically, once the value

of X_{t-d}^{t-1} is known.

Although quite different, the MTD and TAR models share a number of interesting properties, including their ability to capture unusual dependencies and apparent local temporal inhomogeneities, as well as often generating multimodal marginal distributions.

In the generality in which (1.1) is written, it offers more flexibility than both the MTD and TAR models. The improvement over MTD modelling is that the mixture probabilities are allowed to depend on the past history of the process, a consideration that seems extremely natural in a time series setting. The improvement over the basic TAR model is that as the series moves into different regions of the state space, the move from model to model is far smoother. Furthermore the MixAR model, despite its highly non-linear nature, is able to exploit the simplicity of the individual AR components to keep the model computationally tractable.

The additional flexibility of the MixAR model involves, of course, increasing the number of parameters that need to be defined and estimated. However, as we shall argue below, we do not believe that this is a problem.

Section 2 presents the MixAR model in detail, specifies the g_j of (1.1) that we shall work with, and gives a result indicating the predictive power of these models. It also briefly discusses the relationships between MixAR and other non-linear models and the issue of parameterisation. Conditions for the stochastic stability of MixAR processes are given in Section 3. Section 4 describes an efficient parameter estimation algorithm and lists some of its properties and Section 5 describes the analysis of one simulated and two standard series. All proofs are relegated to the Appendix.

Acknowledgements: Haya Kaspi was a big help for Section 3. We are extremely grateful to excellent referees, one of whom went above and beyond the call of duty and checked some of the numerics in Section 5, pinpointing a systematic error in some of our computations. We are particularly grateful to him/her.

2 The MixAR model

To fully specify the MixAR model of (1.1) we need only determine the distribtion of the noise processes, which we take to be be standard Normal, and the weighting functions g_j, which we

take to be multinomial logits; viz.

$$g_j(\mathbf{x};\boldsymbol{\theta}_g) \equiv \frac{\exp(\mathbf{a}_j^T\mathbf{x}+b_j)}{\sum_{k=1}^m \exp(\mathbf{a}_j^T\mathbf{x}+b_j)}, \qquad k=1,\ldots,m. \tag{2.1}$$

Thus, in the notation of (1.1), $\boldsymbol{\theta}_g = [\mathbf{a}_1^T, b_1, \ldots, \mathbf{a}_m^T, b_m]^T \in \mathbb{R}^{m(d+1)}$. Writing $\boldsymbol{\theta}_j = [\theta_{j,1}, \ldots, \theta_{j,d}]^T \in \mathbb{R}^d$, it follows that $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1^T, \theta_{1,0}, \ldots, \boldsymbol{\theta}_m^T, \theta_{m,0}, \boldsymbol{\theta}_g^T, \sigma_1^2, \ldots, \sigma_m^2]^T \in \mathbb{R}^{2m(d+1)+m}$ denotes the full parameter vector specifying the model. Normality is an assumption allowing maximum likelihood estimation. The choice of multinomial logits is based on convenience and the accumulated experience of their reliability in similar roles (e.g. [14]). Obviously, other choices, such as radial basis functions, will probably work just as well, as long as some basic technical restrictions (cf. [44]) hold.

In view of (2.1), it follows that the transition density of such a MixAR(m,d) process, considered as a Markov process of order d, is

$$\mathbb{P}\left(X_t \in dx | X_{t-d}^{t-1} = \mathbf{x};\boldsymbol{\theta}\right) = \sum_{j=1}^m g_j(\mathbf{x};\boldsymbol{\theta}_g)\phi(x;\boldsymbol{\theta}_j^T\mathbf{x}+\theta_{j,0},\sigma_j)\,dx, \tag{2.2}$$

where $\phi(x;\mu,\sigma)$ denotes the density of a $N(\mu,\sigma^2)$ random variable. Thus the conditional density of X_t, given the 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.

Note that while we have taken lags in each element of the mixture to be identical, in practice one can always choose the largest lag and pad out the shorter parameter vectors with zeroes.

With the weight functions g_j now specified in (2.1), we can expand our comments in the Introduction relating the MixAR model to others already in the literature.

It is clear that MiaAR models are closely related to both Priestley's locally linear "state dependent models" [27] and Tong's TAR models, already described above. The essential difference between TAR(m,d) and MixAR(m,d) models is that in the former the g_j are indicator functions, while in the latter they are smooth. Thus, while TAR processes can easily be approximated by MixAR processes, the two classes of models are actually disjoint. The same comment holds when looking at the NEAR mixture model of [7], and the other recent extensions of TAR models, including FAR [8], EXPAR [12], STAR and SETAR [36, 37] and ASTAR [21]. While FAR and ASTAR consider more general schemes of partitioning the state space than TAR, the inherent expense of complicating the specification and estimation tasks seems to be far higher than for MixAR and the type of non-linearities somewhat different.

Perhaps somewhat closer in spirit to MixAR models are the "'regime switching models" of Econometrics. These include AR models in which the mean [13] or the AR coefficients [33] depend on the state of an independent (hidden) Markov chain. (See [37, 11, 35] for more details.) These models have a common link with MixAR models, in that the EM algorithm plays a central role in parameter estimation, but are conceptually quite different, in that they require the introduction of a switching mechanism that is exogenous to the time series of interest.

As with most of the above models, and in common with the neural nets that gave birth to them, MixAR models involve a large $(2m(d+1)+m)$ number of parameters and d and m themselves need to be determined. We want to argue, briefly, that this, while somewhat of an anathema to the conventional statistician, need not be a problem.

There is currently vigorous activity in the statistical literature developing modelling techniques, such as "bagging" [5] and "boosting" [30], which involve parameter dimensions so large that they can even exceed the number of data points. 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. What is known at the moment is that such methodologies often work extremely well in practice [10], although there is, as yet, only limited theory [1, 3, 4, 20, 23] to explain why this should be the case. The basic theme of the theory is that the generally accepted measure of parsimony, the number of parameters, is not necessarily appropriate in the non-linear scenario. Rather, it is more relevant to concentrate on meaningful simple estimators - the m AR components in our case - and an efficient way of combing them - via the weights g_j.

In order to demonstrate the effectiveness of the MixAR architecture and estimation process, we look at some classic examples in Section 5. In all the examples the number of parameters used was large, in one case over half the number of data points. The results there, however, show no evidence of over-fitting and (in all but one sub-case) increased predictive and explanatory capabilities over the best linear model.

Before turning to studying MixAR models in depth, we bring one of their basic approximation properties which describes their ability to provide high quality prediction for general stationary processes.

Consider a strictly stationary, real valued, finite variance, discrete time stochastic process X_t. Then the optimal (mean square) predictor of X_t given the past is the conditional mean.

If X is Markov of order d, then this is given by

$$f(\mathbf{x}) \stackrel{\triangle}{=} \mathbb{E}\left[X_t | X_{t-d}^{t-1} = \mathbf{x}\right]. \tag{2.3}$$

If X is MixAR(m,d), then it follows immediately from (2.2) that the conditional mean takes the form

$$f_m(\mathbf{x};\boldsymbol{\theta}) \stackrel{\triangle}{=} \sum_{j=1}^{m} g_j(\mathbf{x};\boldsymbol{\theta}_g)[\boldsymbol{\theta}_j^T\mathbf{x} + \theta_{j,0}]. \tag{2.4}$$

The following result, a proof of which can be found in [44], indicates how well the prediction function of a general Markov process of order d can be approximated by a MixAR(m,d) process.

Theorem 2.1 *Let X_t be a strictly stationary, order d, Markov process and let μ denote the d-dimensional stationary distribution of $(X_t,\ldots,X_{t+d})$, Assume that 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*

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

where f and f_m are defined in (2.3) and (2.4).

Under stronger assumptions on the underlying true process, one can go beyond Theorem 2.1 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 [44] shows that there is a constant c, dependent only on r, d, q and K, such that

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

3 Stationarity and ergodocity

We shall now establish the existence of stationary, ergodic solutions to (1.1), both properties being, of course, essential for consistency once we turn to parameter estimation. A further property, geometric ergodicity, (cf. (3.1) below), implies that the process is exponentially β-mixing, which is important for establishing central limit theorems for estimates.

As a first step, write $\mathbb{X}_t = X_{t-d}^{t-1}$, so that $\mathbb{X}_t$ is a vector-valued process. The fact that X_t is Markov of order d implies that $\mathbb{X}_t$ is a simple Markov chain – i.e. Markov of order 1 – with the

well behaved transition density implicit in (2.2), and so we have no problems with existence. We now need to find a stationary distribution for $\mathbb{X}_t$. Since the first coordinate of $\mathbb{X}_t$ follows the sample path of X_t, existence, stationarity and ergodicity for $\mathbb{X}_t$ will imply equivalent results for X_t.

We start with the case $d = 1$. The transition density of $X_t \equiv \mathbb{X}_t$ is then simply

$$p(x, dy) \quad = \quad \sum_{j=1}^{m} g_j(x; \boldsymbol{\theta}_g)\phi(y; \theta_j x + \theta_{j,0}, \sigma_j)\ dy.$$

Write $p^n(x, y)$ for the n-step transition density and $P^n(x, \cdot)$ for the corresponding measure. A measure, π, satisfying $\pi(B) = \int_B dy \int_{\mathbb{R}} p(x, y)\pi(dx),\ \forall\ B \in \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel subsets of $\mathbb{R}$, is an invariant probability measure of X_t and if π is unique then the process is also ergodic. (For this, and other basic facts about Markov processes, our reference is [24].)

Now assume, without loss of generality, that the AR components of our MixAR$(m, 1)$ have been ordered so that $a_1 \leq a_2 \leq \cdots \leq a_m$, where the a_j parameterize g_j as in (2.1). Then we have the following stability result, whose proof is deferred to the Appendix.

Theorem 3.1 *Let X_t 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; viz. there exists $r > 1$ such that*

$$\sup_{B \in \mathcal{B}(\mathbb{R})} |P^n(x, B) - \pi(B)| \quad = \quad O(r^{-n}). \tag{3.1}$$

Before looking at the case of general d note that it follows from the theorem that it suffices to restrict attention to the two local AR(1) models in the mixture that are dominant in terms of the behaviour 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. Call these the "extreme" models. A priori, it seems plausible that if *all* component AR models are stable (i.e. $|\theta_j| < 1\ \forall j$) then the MixAR$(m, 1)$ would be stable as well. This is in fact the case, but the theorem asserts a more refined result. Roughly, if the two extreme AR models cause an average drift towards 0, then the MixAR$(m, 1)$ is stable. Note that the two extreme models need not be stable simultaneously. We note in passing that the conditions in Theorem 3.1 are similar to the ones obtained for TAR processes by Tong and co-workers (cf. [37] for details).

We now turn to the general MixAR(m, d) process, with $d \geq 1$ where, unfortunately, we have only been able to obtain rather strong sufficient conditions. Set

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

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). Then we have the following result, the proof of which is also in the Appendix.

Theorem 3.2 *Let X_t 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 $\mathbb{X}_t = X_{t-d}^{t-1}$ has a unique stationary probability measure and is geometrically ergodic in the sense of (3.1).

Note that if $d = 1$ then Theorem 3.2 requires that $\alpha_1 < 1$, i.e. $\max_{1 \le i \le m} |\theta_i| < 1$. While this is consistent with Theorem 3.1, it seems that for $d > 1$ our results are probably too stringent in general.

4 Parameter Estimation and Inference

In this Section we shall show how to use the EM algorithm [9] to estimate, via maximum likelihood, the parameters of a MixAR(m, d) process, when both m and d are known. The EM algorithm, with which we assume the reader is familiar, is a natural choice, having been used extensively for both general mixture problems (cf. [32]) and in the mixture of experts scenario (cf. [17, 18]).

Let $\mathcal{D}_N = \{X_t\}_{t=1}^N$ be a sample from a stationary, ergodic, MixAR(m, d) process. Then the corresponding *reduced* likelihood equation is

$$L(\mathcal{D}_N; \boldsymbol{\theta}) = \prod_{t=d+1}^{N} \sum_{j=1}^{m} g_j(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \phi(X_t; \boldsymbol{\theta}_j^T X_{t-d}^{t-1} + \theta_{j,0}, \sigma_j) \tag{4.1}$$

Write $\hat{\boldsymbol{\theta}}_N \equiv \text{argmax}_{\boldsymbol{\theta}} L(\mathcal{D}_N; \boldsymbol{\theta})$ for the maximum likelihood estimator (MLE).

Direct maximization of (4.1) or its logarithm is complicated, since the mixture structure of the MixAR transition kernel leads to logarithms of sums. To overcome this we turn to the EM algorithm.

To set this up, let

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

be "hidden" indicator variables denoting which model is active at any given time. Let $\mathcal{Z} = \{X_t, \{I_t^j\}_{j=1}^m\}_{t=d+1}^N$ denote the observed and missing variables. Then the *complete data* (reduced) log-likelihood is

$$\log P(\mathcal{Z}|\boldsymbol{\theta}) = \sum_{t=d+1}^{N} \sum_{j=1}^{m} I_t^j \log \left[g_j(X_{t-d}^{t-1}; \boldsymbol{\theta}_g) \cdot \phi(X_t; \boldsymbol{\theta}_j^T X_{t-d}^{t-1} + \theta_{j,0}, \sigma_j) \right]. \tag{4.2}$$

This is a random variable in the missing data and now we take an expectation with respect to them to obtain the complete expected log-likelihood

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) \equiv \mathbb{E} \log P(\mathcal{Z}|\boldsymbol{\theta}) &= \sum_{t=d+1}^{N} \sum_{j=1}^{m} \tau_{t,j}^{(k)} \log[g_j(X_{t-d}^{t-1}; \boldsymbol{\theta}_g] \\ &+ \sum_{t=d+1}^{N} \sum_{j=1}^{m} \tau_{t,j}^{(k)} \log \left(\phi(X_t; \boldsymbol{\theta}_j X_{t-d}^{t-1} + \theta_{j,0}, \sigma_j) \right), \end{aligned} \tag{4.3}$$

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

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

$$\tau_{t,j}^{(k)} = \frac{g_j(X_{t-d}^{t-1}; \boldsymbol{\theta}_g^{(k)}) \cdot \phi(X_t; \boldsymbol{\theta}_j^{(k)} X_{t-d}^{t-1} + \theta_{j,0}^{(k)}, \sigma_j^{(k)})}{\sum_{i=1}^m g_i(X_{t-d}^{t-1}; \boldsymbol{\theta}_g^{(k)}) \cdot \phi(X_t; \boldsymbol{\theta}_j^{(k)} X_{t-d}^{t-1} + \theta_{j,0}^{(k)}, \sigma_j^{(k)})}. \tag{4.4}$$

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 (4.3) and (4.4) is the *expectation step* (**E**). The maximization step (**M**) maximizes $Q(\cdot|\boldsymbol{\theta}^{(k)})$, and consequently (cf. [9, 42]) the likelihood (4.1) itself. Differentiating with respect to $\boldsymbol{\theta}$ yields the parameter update equations

$$\begin{aligned} \sigma_j^{(k+1)} &= \left(\frac{1}{\sum_{t=d+1}^N \tau_{t,j}^{(k)}} \sum_{t=d+1}^{N} \tau_{t,j}^{(k)} [X_t - \boldsymbol{\theta}_j^{(k)} X_{t-d}^{t-1} - \theta_{j,0}^{(k)}][X_t - \boldsymbol{\theta}_j^{(k)} X_{t-d}^{t-1} - \theta_{j,0}^{(k)}]^T \right)^{1/2} \\ \left[(\boldsymbol{\theta}_j^{(k+1)})^T, \theta_{j,0} \right]^T &= (R_j^{(k)})^{-1} c_j^{(k)}, \qquad \boldsymbol{\theta}_g^{(k+1)} = \boldsymbol{\theta}_g^{(k)} + \alpha \left. \frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_g} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} \end{aligned} \tag{4.5}$$

where the matrix R_j and the vector c_j are given by

$$R_j^{(k)} = \sum_{t=d+1}^{N} \tau_{t,j}^{(k)} \mathcal{X}_{t-1} \mathcal{X}_{t-1}^T, \qquad c_j^{(k)} = \sum_{t=d+1}^{N} \tau_{t,j}^{(k)} \mathcal{X}_{t-1} x_t, \tag{4.6}$$

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.

We used steepest ascent, as is evident in the update equation, where α is the step size obtained from a line search. With this setup, we are actually in the framework of the GEM (generalized EM) algorithm, although for brevity we continue refering to it simply as EM.

One of the main advantages of the above estimation algorithm is its straightforward implementation and the associated global convergence result given in Proposition 4.1.

As far as our practical experience with the algorithm is concerned, we have not experienced any problems beyond occasional slow convergence in quite a few test cases which we investigated, some of which are reported in Section 5. As in all non-trivial, non-linear, optimization problems, we occasionally needed to experiment with multiple initial conditions.

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 [29]. We also have the following, which summarizes the key properties of the algorithm. A sketch of the proof, which is quite standard, is given in the Appendix.

Proposition 4.1 *Let $\{\boldsymbol{\theta}^{(k)}\}$ be the sequence produced by EM algorithm described above and suppose that it is contained in a compact subset of $\mathbb{R}^{2m(d+1)}$. Write $\mathcal{L}(\boldsymbol{\theta}) = \log P\left(X_{d+1}^n; \boldsymbol{\theta}\right)$. 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 $\boldsymbol{\theta}^*$.

(3) If in addition $\{\boldsymbol{\theta} : \mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^*\} = \{\boldsymbol{\theta}^*\}$, then $\boldsymbol{\theta}^{(k)} \to \boldsymbol{\theta}^*$ as $k \to \infty$.

In order to obtain convergence of the $\boldsymbol{\theta}^{(k)}$ we need the additional condition that the limit set of stationary points is a singleton. In that case $\boldsymbol{\theta}^{(k)} \to \hat{\boldsymbol{\theta}}_N$, the ML estimator.

We complete this section with a brief discussion of the distributional 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 simply a relabeling among classes. Standard techniques, as in [29], now work quite well. 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 parametrization of the weight functions g_j be defined via the

first (in lexicographical order, say) $m-1$ 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 Theorem 1 and Corollary 1 of [16].

Assume that the MixAR process is both stationary and ergodic, sufficient conditions for which are given in Theorems 3.1 and 3.2. Let $\mathbb{X}_t$ be the vector Markov process of Section 3 and set

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

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 $J_{k\ell} \equiv \mathbb{E}[u_k u_\ell]$. Then, under mild side conditions, Theorems 2.1 and 2.2 of [2] apply and establish that

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

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 in detail in [15].

5 Numerical Examples

We now illustrate the MixAR model by applying it to one simulation and two benchmark series. In the latter cases, we shall compare the MixAR results to those obtained via the "best" (in the sense described below) linear (ARMA) and threshold (TAR) models, as well as some other non-linear models.

Both benchmark data sets were partitioned into subsets. The initial portion was used for estimation purposes, and is referred to as the *training* set. The others, the *prediction* subsets, were not revealed during the estimation phase and were subsequently used to evaluate the performance of the fitted model. This is particularly important to keep in mind when reading the results, since the fitted MixAR models involve numbers of parameters that are of the order of 1/3–1/2 of the numbers of observations in the training sets. Thus the fact that our results are rather positive indicate that the issue of parsimony raised in Section 2 is, as intimated there, seemingly not related to a simple parameter count.

Our basic measure of model accuracy was the *normalized mean squared error* (NMSE) defined as follows. For a subset $\mathcal{S}$ of the time series and a predictor $\hat{X}$,

$$\text{NMSE}(\mathcal{S}, \mathcal{T}) \;=\; \frac{\sum_{t\in\mathcal{S}}[X_t - \hat{X}_t(X_{t-d}^{t-1}; \mathcal{T})]^2}{\sum_{t\in\mathcal{S}}[X_t - \bar{X}(\mathcal{S})]^2} \;=\; \frac{1}{|\mathcal{S}|\ \hat{\sigma}_{\mathcal{S}}^2} \sum_{t\in\mathcal{S}} \left[X_t - \hat{X}_t\right]^2, \qquad (5.1)$$

where $|\cdot|$ denotes cardinality, $\mathcal{T}$ denotes the training set, $\bar{X}(\mathcal{S})$ is the empirical mean over $\mathcal{S}$, and $\hat{X}_t(X_{t-d}^{t-1};\mathcal{T})$ is the predictor based on the parameters estimated using the set $\mathcal{T}$. The "best" model was generally considered to be that with the minimal NMSE, when $\mathcal{S}$ was taken to be the predicition set.

Parameter estimation followed the EM algorithm as described in Section 4, using `MATLAB` code. All sequences were linearly normalized to take values in $[0,1]$, so as to avoid standard numerical problems associated with non-linear algorithms. Prediction values were transformed back to the original scale.

5.1 Simulated TAR Data

We start with an example of TAR generated data. 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 TAR(2;1,1) model, 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 5.1 summarizes the results of the analysis, with the data shown in the (X_t, X_{t-1}) plane.

The parameter estimates for the two AR models were $\theta_1 = 0.7501$, $\theta_{1,0} = 0.9244$, $\theta_2 = -0.6004$, $\theta_{2,0} = -0.4655$, in the same ball park as the true values.

It is clear from Figure 5.1(b) that the state-dependent probabilities g_1 and g_2 degenerate in this case, giving a good approximation of the underlying threshold model.

5.2 MixAR Model Specification

Before estimating the parameters of a MixAR(m,d) series, m and d need to be determined somehow. In both of the benchmark series, we adopted a rather heuristic approach, which seemed to work well.

In particular, we choose the maximal lag size d much as one would in the preliminary analysis of a simple ARMA model, which we did by looking at the spectrum. We used the same d for each of the component models, allowing the estimation algorithm to set parameters to zero (or close to zero) if it felt this was extravagent. The number of component models was chosen dynamically, starting with $m = 2$ and increasing in unit steps until the most recently added component was not significantly different from one of the earlier ones, a degeneracy

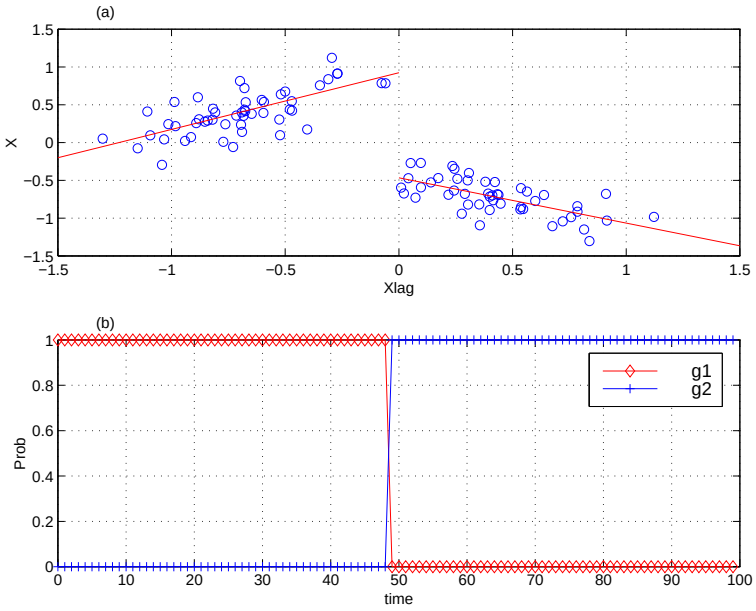

Figure 5.1: Analysis of TAR generated data. (a) Simulated data points and local AR models fitted by the MixAR(2; 1) model (solid line), observed in the (X_t, X_{t-1}) plane. (b) Local AR model probabilities g_1 and g_2

that was easily identified when it occured. We preferred this method to using more automated measures such as AIC or BIC, whose interpretation is not always clear and which have not proven themselves to be highly reliable for non-linear models.

5.3 Canadian Lynx Series

Our analysis of the logarithm (base 10) of the Canadian lynx series is summarized in Figure 5.2. Observations 1–93 (1821–1920) were the training set, and 94–107 (1921–1934) were the prediction set. We ended up with a MixAR$(3, 6)$ with a total of 45 parameters. The precise values of the estimated parameters, along with those of all the other models we quote, can be found in [43].

Figure 5.2 (a) and (b) show good fit and predictive power, but more precise information comes from comparison with other models. The log-lynx data has been heavily analysed, including,

(i) An AR(12) linear model, [6] p550.

(ii) A TAR(2,8,3) threshold model, [36] p190.

(iii) A bilinear model, [31] p204, [6] p552.

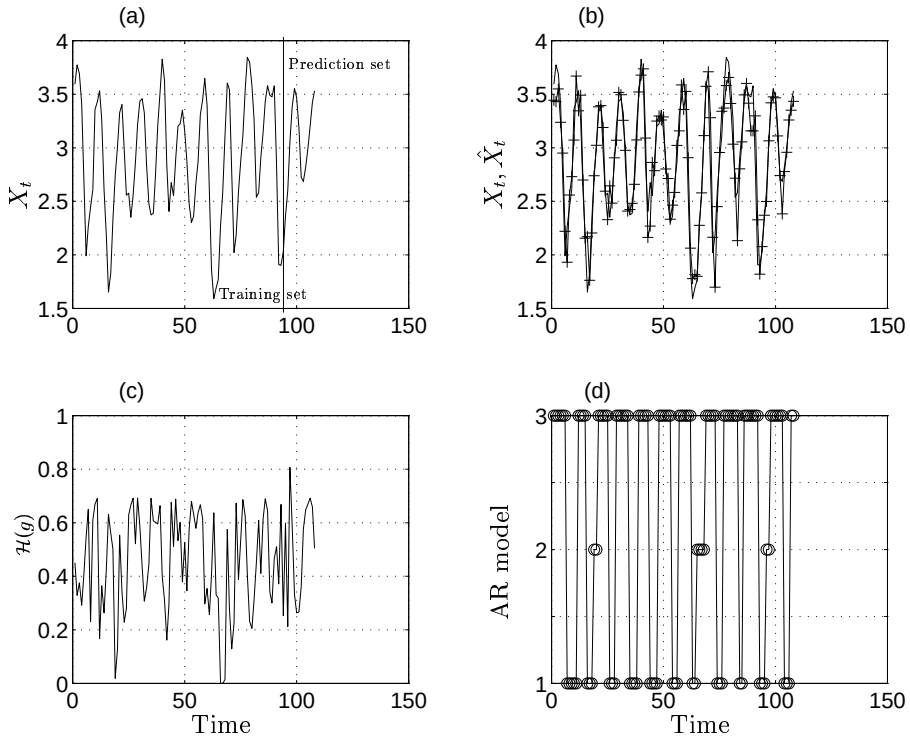

Figure 5.2: Analysis of the log transformed Canadian lynx series (1821-1934). (a) The log lynx series. (b) One step prediction of the MixAR$(3, 6)$ model (+, line) for the training set (1-93) and prediction set (94-107), superimposed on the original data series. (c) Entropy $\mathcal{H}(g)$ of the AR component probabilities $\{g_j\}$. (d) Most probable AR model (largest value of $\{g_j\}_{j=1}^{3}$).

(iv) An AR(2) random coefficient model [25] p144,[6] p552.

Table 1 compares the NMSE for all four models, over both the training and prediction sets. What is rather interesting is that the MixAR does not have the lowest NMSE over the training set, where one would, *a priori*, expect it to outperform other models because of the large number of paramters involved (with the natural, and undesired, side effect of overfitting). However it *does* have the lowest NMSE over the predicition set, where it outperforms the other models by between 25% to 50%.

Returning now to Figure 5.2, note the (non-Gaussian) irreversibility of the series, exhibited by the fact the series rises to, and falls from, its local maximum at different rates.

It is interesting to see how the MixAR model copes with this. Figure 5.2 (d) shows which of the three underlying AR components (1,2 or 3) is dominant at any given time, where dominance is measured by which model exhibits the largest probability g_j at a given time. Comparing

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

this with (a) or (b), one immediately sees that one component models the process during rises, one at falls and one at changes. Note that this structure was not imposed on the model in any way, but rather is a natural outcome of the estimation algorithm.

Additional information is given in Figure 5.2 (c), which gives the entropy

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

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 component ($\mathcal{H} = 0$) and a uniform distribution over all m models ($\mathcal{H} = \log_2 m = \log_2 3 \approx 1.58$). For further details on our analysis of both this and the sunspot series see [43].

5.4 Sunspots Series

The Wölfer sunspot number data set is perhaps one of the most extensively studied time series in the literature, since its exact underlying mechanism is still unclear. What is clear is that it rises to peaks more rapidly than it falls from them and so non-linear models are called for. (cf. [37] §7.3, for a detailed discussion.)

Following convention, we divided the data into a training set of 221 points (1700-1920) and two prediction sets: I (1921-1955) with 35 points and II (1956-1979) with 24. We ended up with a MixAR(3, 12) model, using 81 parameters.

The basic results are evident in Figure 5.3. There is more detailed analysis in [43], which includes a listing of the model parameters. Again it is interesting to note that the final model involved three components, which, on the basis of Figure 5.3 (c), can be seen to correspond to rises, falls and change points of the data. Rather interestingly, only two of these models come into play at the prediction stage.

Prediction was good for the first prediction set (which actually shows a little more variability than the training set ($\hat{\sigma}_I^2 = 1674$ vs. $\hat{\sigma}_T^2 = 1174$) but was poorer for the second prediction set,

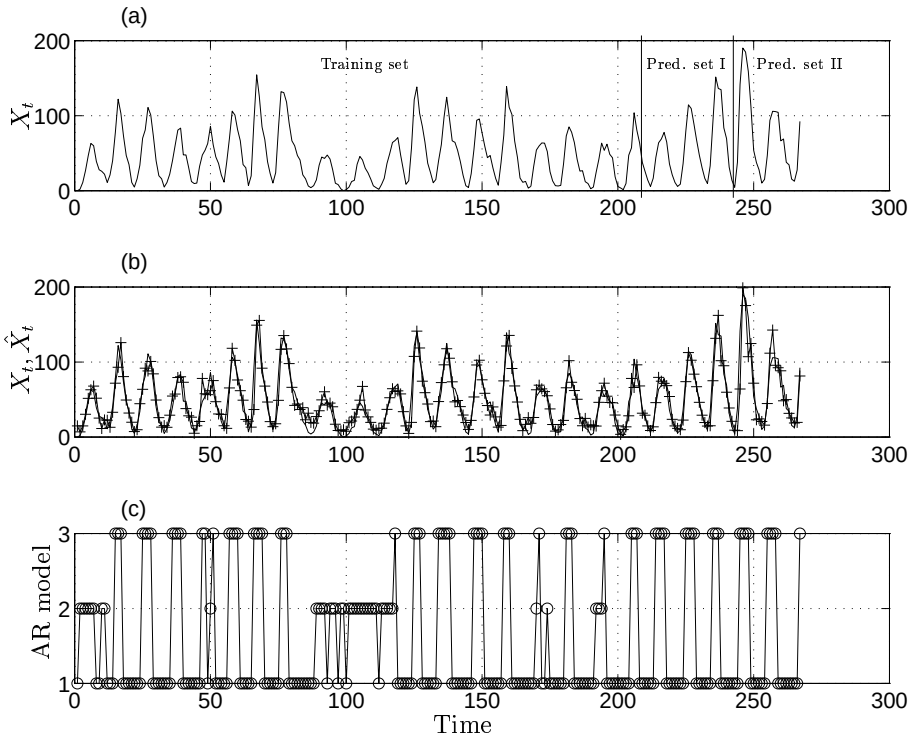

Figure 5.3: Analysis of the Wölfer number, sunspot series. (a) Sunspot series divided into the training set (1-209) and prediction set I (210-243) and II (244-267). (b) MixAR(3, 12) one step prediction (+, line), superimposed on the original data series. (c) Most probable AR model (highest value $\{g_j\}_{j=1}^3$).

which has relatively high variability ($\hat{\sigma}_{II}^2 = 2852$). (This is easier to see in Table 2.) The significant increase in intrinsic variability in the second predicition set makes it clear that prediction here is going to be harder for any stationary model.

For comparison, consider the following models, all of which have been applied to this data:

(i) An AR(9) linear model, [31] p196.

(ii) A TAR(2,4,12) threshold model, [36], p241.

(iii) A NN(3,12) neural network, [40].

(A NN(m, d) neural network is a feed-forward network with an input layer of dimension d, a hidden layer with m sigmoidal units, and one linear output unit, realizing the function $\sum_{j=1}^m \alpha_j \sigma(\mathbf{w}_j^T \mathbf{x} + b_j)$ where $\sigma(x) = (1 + e^{-x})^{-1}$, and $\alpha_j, b_j \in \mathbb{R}$, $\mathbf{w}_j \in \mathbb{R}^d$. For more details see [39, 11].)

Comparisons are summarized in Table 2. The results for models (ii) and (iii) are taken from the original references, while for (i) (not given in the original) they were generated using the model.

Model	Training set	Prediction set I	Prediction set II
AR(9)	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 and prediction sets for the sunspots series.

The results here are interesting. In the first prediction set, which, as noted above, is most like the training set in terms of overall variability, the MixAR model is beaten only by the neural net, whereas in the second prediction set the more traditional AR and TAR models outperform the highly non-linear neural net and MixAR.

While the reason behind the poorer behaviour of MixAR in this case is not clear, it is probably related to the fact that the second prediction set seems to be qualitatively different from both the training set and first prediction set, indicating an intrinsic non-stationarity. (This is the reason why, for this data set, it is "traditional" to work with two prediction sets.) It is not surprising that the highly non-linear neural network and the MixAR models are more sensitive to such a change of behaviour.

A Proofs of Theorems

Throughout this, rather technical, section we shall assume that the reader is familiar with Meyn and Tweedie [24], which we shall cite as MT.

Proof of Theorem 3.1: We start with an outline of the main steps and ideas.

Step 1: We prove that the Markov chain defined by the MixAR$(m, 1)$ process is λ-irreducible (MT, p87), 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: 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 applica-

tion of this 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 p190 is sufficient to prove Harris recurrence (MT, p200). 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, p262) and Theorem 11.0.1 of MT, concludes that the invariant measure is finite, and so π exists and is unique. Consequently the chain is ergodic. It turns out that 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.

1. *Proof of λ-irreducability*: For Borel B, define the hitting time $\tau_B = \inf\{n \geq 1 : X_n \in B\}$. Since the transition probability kernel (2.2) has a positive density, if $\lambda(B) > 0$ then

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

for all $x \in \mathbb{R}$. It follows that the Markov chain $(X_t)_{t\geq 0}$ is λ-irreducible.

2. *Proof of geometric ergodicity*: We deviate slightly from the notation in MT, in order to clarify some definitions. From Theorem 16.0.2 of MT 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

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

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, while slightly different than that used in MT, it is the right definition of drift for our purposes. (MT use petite, rather than compact, sets. However, the structure of (2.2) implies that MixAR processes are weak Feller and consequently, by Proposition 6.2.8 of MT, every compact set is petite.) The issue now is the 'right' choice of Lyapunov function V and corresponding set K.

From the conditions of Theorem 3.1, there must exist two positive reals a and b such that $-a/b < \theta_1 < 1$, $-b/a < \theta_m < 1$. Take

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

and write Φ for the standard 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 $a\int_{\mathbb{R}_+} u\phi(u;\mu,\sigma)du$ and $b\int_{\mathbb{R}_-} u\phi(u;\mu,\sigma)du$ for the m Gaussian components of the mixture and C_1 is a constant depending on $a, b, \max_j |\sigma_j|$ and $\max_j |\theta_{j,0}|$.

Note that $g_1(x) \to 1$ for $x \to \infty$ and $g_j(x) \to 0$ for $j > 1$. In particular $|g_j(x)| \leq \exp\{-(\theta_1 - \theta_j)x\}$ so that $g_j(x)x^k = o(1)$ for all $k \in \mathbb{N}$. Thus, for $\delta \in (1-\rho, 1)$, there exists an $r > 0$ 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

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

and so

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

Thus 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 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

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

so that

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

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 through the same arguments for the case of $x \in \mathbb{R}_-$ and show that for $-b/a < \theta_m < 1$ we can choose a compact set K'' such that the Lyapunov stability condition (A.1) is met. This concludes the proof. ■

Proof of Theorem 3.2: Set $\mathbf{Y}_t = X_{t-d+1}^t = [X_t, \ldots, X_{t-d+1}]^T$. That $\mathbf{Y}_t$ is λ-irreducible follows as in the proof of Theorem 3.1. That it is aperiodic is also easily verified. Motivating the derivations to follow is the choice of the test function $V(\mathbf{y}) = \|\mathbf{y}\|_2$. Define the $d \times d$ matrices

$$\mathbf{A}_j \triangleq \begin{bmatrix} \boldsymbol{\theta}_j(1) & \boldsymbol{\theta}_j(2) & \cdots & \boldsymbol{\theta}_j(d-1) & \boldsymbol{\theta}_j(d) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \qquad \mathbf{D} \triangleq \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{d-1} & \alpha_d \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

and d vectors $\mathbf{b}_j \equiv [\theta_{j,0}, 0, 0, \ldots, 0]^T$. Set

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

let $\{\varepsilon_t^{(j)}\}$ be as in (1.1) and write $\mathbf{Z}_{j,t} \equiv [\varepsilon_t^{(j)}, 0, 0, \ldots, 0]^T$. Let $\{I_t\}_{t \geq 0}$ be a sequence of mutually independent, integer valued random variables such that, given X_{t-d}^{t-1},

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

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

$$\mathbf{Y}_t = \mathbf{A}_{I_t} \mathbf{Y}_{t-1} + \mathbf{b}_{I_t} + \mathbf{Z}_{I_t,t}. \tag{A.2}$$

It is easily verified that $\|\mathbf{A}_j\| > 1$ for all j, (where $\|\mathbf{A}\| \equiv \sup_{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1} \|\mathbf{A}\mathbf{x}\|$ denotes the usual norm of a matrix $\mathbf{A}$) 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 [34] Lemma 3.1, that is suited to our setting.

Lemma A.1 *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

$$\begin{aligned}\mathbf{Y}_{t+h-1} &= \left(\prod_{i=0}^{h-1}\mathbf{A}_{I_{t+i}}\right)\mathbf{Y}_{t-1} + \sum_{i=0}^{h-2}\left(\prod_{j=i+1}^{h-1}\mathbf{A}_{I_{t+j}}\right)\mathbf{b}_{I_{t+i}} \\ &\quad + \sum_{i=0}^{h-2}\left(\prod_{j=i+1}^{h-1}\mathbf{A}_{I_{t+j}}\right)\mathbf{Z}_{I_{t+i},t+i} + \mathbf{b}_{I_{t+h-1}} + \mathbf{Z}_{I_{t+h-1},t+h-1}. \qquad \text{(A.3)}\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}[\|\mathbf{Y}_{t+h-1}\||\mathbf{Y}_{t-1}=\mathbf{y}] &\leq \mathbb{E}\left[\left\|\prod_{i=0}^{h-1}\mathbf{A}_{I_{t+i}}\mathbf{Y}_{t-1}\right\| |\mathbf{Y}_{t-1}=\mathbf{y}\right] \\ &\quad +\mathbb{E}\left[\left\|\sum_{i=0}^{h-2}\left(\prod_{j=i+1}^{h-1}\mathbf{A}_{I_{t+j}}\right)\mathbf{b}_{I_{t+i}}\right\| |\mathbf{Y}_{t-1}=\mathbf{y}\right] \\ &\quad +\mathbb{E}\left[\left\|\sum_{i=0}^{h-2}\left(\prod_{j=i+1}^{h-1}\mathbf{A}_{I_{t+j}}\right)\mathbf{Z}_{I_{t+i},t+i}\right\| |\mathbf{Y}_{t-1}=\mathbf{y}\right] \\ &\quad +\mathbb{E}\left[\left\|\mathbf{b}_{I_{t+h-1}}\right\| |\mathbf{Y}_{t-1}=\mathbf{y}\right] + \mathbb{E}\left[\left\|\mathbf{Z}_{I_{t+h-1},t+h-1}\right\| |\mathbf{Y}_{t-1}=\mathbf{y}\right] \\ &\equiv E_1+E_2+E_3+E_4+E_5 \qquad \text{(A.4)}\end{aligned}$$

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

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

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 is easy to check. (There is a proof in [43].)

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*

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

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

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

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 $\lim_{h\to 0}\|\mathbf{D}^h\|^{1/h}\to\lambda_{\max}(\mathbf{D})$, 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

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

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

$$\mathbb{E}[\|\mathbf{Y}_{t+h-1}\||\mathbf{Y}_{t-1}=\mathbf{y}] \;\leq\; \rho\|\mathbf{y}\|+C \;\leq\; \frac{\rho+1}{2}\|\mathbf{y}\| \tag{A.5}$$

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

Remark A.1 Tjøstheim, [34], Theorem 4.5, gives sufficient conditions for the stability of a multivariate TAR(m,d) model. (cf. also [37].) In spite of the apparent similarity of the MixAR and TAR models, Tjøstheim's result does not carry over to the MixAR setting, since the k-cycle assumption made there is no longer applicable.

Proof of Theorem 2.1: The proof proceeds in three steps, somewhat condensed to remain within the scope of this paper.

1^0: We first 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 which the reader is referred to [44].

3^0. The third step is an application of Theorem A.1 of [44], adapted to the above functional class. A straightforward adaptation of Theorem A.1 there 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 $f_m^*(\mathbf{x}; \boldsymbol{\theta}) = \sum_{j=1}^m \alpha_j g_j(\boldsymbol{\theta}_g; \mathbf{x})$, such that $\|f - f_m^*\|_q < \epsilon$.*

Sketch of Proof for Proposition 4.1: The key here is an application of the so-called global convergence theorem (cf. [22] p125). 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, [22] §7.6, we have, using Corollary 2 of [22] p125, that the composite algorithm is a closed point-to-point mapping. Thus, by Theorem 1 and Theorem 4 of [42], we are done. ■

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*, **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.* **52**, 434–452.

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

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

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

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

[9] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. Roy. Statist. Soc. B*, **39**, 1-38.

[10] 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

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

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

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

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

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

[16] Jiang, W. and Tanner, M. (2000a) On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models, *IEEE Trans. on Inform. Theory*, **46**, 1005-1013.

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

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

[19] Le, N.D., Martin, R.D. and Raftery, A.E., (1996) Modelling flat stretches, bursts and outliers in times series using mixture transition distribution models. *J. Am. Statis. Soc.* **91**, 1504-1514.

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

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

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

[23] Meir, R., El-Yaniv, R. and Ben-David, S. (2000) Localised boosting, *Proc. 13th Annual Conference on Computational Learning Theory*, eds. N. Cesa-Bianchi and S. Goldman, Morgan Kaufmann, San Francisco, 190–199.

[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, **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*, **91**, 953-960.

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

[28] Raftery, A.E., (1985) A model for high-order Markov chains, J. Roy. Statist. Soc., **47**, 528-539.

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

[30] 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.* **26**, 1651–1686.

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

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

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

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

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

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

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

[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. **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*, **1**, 193-209.

[41] Wong, C.S. and Li, W.K., (2000) On a mixture autoregressive model, *J. R. Statist. Soc.* **62**, 95-115.

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

[43] Zeevi, A.J., Meir, R. and Adler, R.J.. (2000) Non-linear models for time series using mixtures of autoregressive models – the long version. `http://iew3.technion.ac.il:8080/~radler/publications.html`

[44] 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*, **44**, 1010–1025.