MODELING NONLINEARITIES WITH MIXTURES-OF-EXPERTS OF TIME SERIES MODELS

ALEXANDRE X. CARVALHO AND MARTIN A. TANNER

Received 5 February 2006; Revised 21 May 2006; Accepted 28 May 2006

We discuss a class of nonlinear models based on mixtures-of-experts of regressions of exponential family time series models, where the covariates include functions of lags of the dependent variable as well as external covariates. The discussion covers results on model identifiability, stochastic stability, parameter estimation via maximum likelihood estimation, and model selection via standard information criteria. Applications using real and simulated data are presented to illustrate how mixtures-of-experts of time series models can be employed both for data description, where the usual mixture structure based on an unobserved latent variable may be particularly important, as well as for prediction, where only the mixtures-of-experts flexibility matters.

1. Introduction

The last three decades have experienced a great deal of research on nonlinear regression models, as described in [23]. Among the several models proposed in the literature, we can find an important class denoted as mixtures-of-experts (ME), and its extension, denoted as hierarchical mixtures-of-experts (HME). Since the publication of the original papers by Jacobs et al. [26, 33], these two classes of models have been used in many different areas to account for nonlinearities, and other complexities in the data. In these models, the dependent variable $y_t \in \mathcal{S} \subset \Re$ is assumed to have the following conditional density specification:

$$f(y_t \mid x_t, \theta) = \sum_{j=1}^{J} g_j(x_t;\gamma)\pi(y_t;\eta(\alpha_j + x_t'\beta_j), \varphi_j), \tag{1.1}$$

where $x_t \in \mathbb{X} \subset \Re^s$ is a vector of covariates, and $\pi(y_t;\eta(\alpha_j + x_t'\beta_j),\varphi_j)$ is a generalized linear model [38] with mean $\eta(\alpha_j + x_t'\beta_j)$ and dispersion parameter φ_j. The specification

Hindawi Publishing Corporation
International Journal of Mathematics and Mathematical Sciences
Volume 2006, Article ID 19423, Pages 1–22
DOI 10.1155/IJMMS/2006/19423

in (1.1) describes a mixture model, with J components, where the weights $g_j(x_t;\gamma) \in (0,1)$ are also functions of the covariate vector x_t.

Because of its great flexibility, simple construction, and good modeling properties, ME started to be commonly used in models for nonlinear time series data. Let y_t be a univariate stochastic process observed at time epoch t, $t = 1,\dots,T$, and let $\mathbb{I}_{t-1}$ be the available information set at time $t-1$. In the time series ME construction, the conditional density of y_t given $\mathbb{I}_{t-1}$ is assumed to have the form in (1.1), where x_t may include lags of transformations of the observed response y_t, as well as lags of external predictors.

An application of ME to signal processing in a noninvasive glucose monitoring system is presented in [35]. Reference [22] applies ME to gender and ethnic classification of human faces. Reference [37] presents the use of ME to uncover subpopulation structure for both biomarker trajectories and the probability of disease outcome in highly unbalanced longitudinal data. Reference [27] presents an application of ME in modeling hourly measurements of rain rates. Reference [18] studies local mixtures-of-factor models, with mixture probabilities varying in the input space. Reference [48] employs a model based on combinations of local linear principal components projections, with estimation performed via maximum likelihood. In [52], the authors apply ME, what they called "gated experts," to forecast stock returns. Reference [54] studies mixtures of two experts, referred to as "logistic mixture autoregressive models." Finally, [57] treats mixtures of autoregressive experts, what they call "mixtures of local autoregressive models," or MixAR models, where the covariate vector x_t contains only lags of y_t.

The stochastic underlying process represented by (1.1), in a time series context, can be interpreted as follows: imagine there exist J autoregressive processes $\pi(y_{j,t};\eta(\alpha_j + x_t'\beta_j),\varphi_j)$, all belonging to one specific parametric family $\pi(\cdot;\cdot,\cdot)$, and, conditional on the past information $\mathbb{I}_{t-1}$, each component j generates a response $y_{j,t}$, $j = 1,\dots,J$. Additionally, imagine there is a multinomial random variable $I_t \in \{1,2,\dots,J\}$, independent of $y_{j,t}$, where each value j has a probability $g_j(x_t;\gamma) \in (0,1)$, and if $I_t = k$, the value $y_t = y_{k,t}$ is observed. Based on the law of iterated expectations, we conclude that (1.1) is the conditional density for y_t, given $\mathbb{I}_{t-1}$.

In general, the probabilities $g_j(x_t;\gamma)$ are assumed to have a logistic form:

$$g_j(x_t;\gamma) = \frac{\exp(v_j + u_j'x_t)}{\sum_{k=1}^{J}\exp(v_k + u_k'x_t)}, \tag{1.2}$$

where v_j and $u_j \in \Re^s$, $j \in \{1,\dots,J\}$, are unknown parameters. In order to avoid identification problems, we assume that

$$v_J = 0, \qquad u_J = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}'. \tag{1.3}$$

The mixed components $\pi(y_t;\eta(\alpha_j + x_t'\beta_j),\varphi_j)$ are referred to as experts and the probabilities (or weights) $g_j(x_t;\gamma) \in (0,1)$ are called gating functions or simply gates. The grand vector of gating parameters is the list of all the individual gating parameters $\gamma = (v_1,u_1',v_2,u_2',\dots,v_{J-1},u_{J-1}')'$.

Several properties for ME and HME were initially proved by Jiang and Tanner [28–31]. They treated consistency and asymptotic normality of the maximum likelihood estimator, conditions for parameter identifiability, and approximations properties, for exponential family experts. Nonetheless, although these results are quite general, they do not apply directly to time series data. The authors assumed independent observations y_t, $t = 1, \ldots, T$, and a compact covariate space $\mathbb{X}$.

In a series of papers, Carvalho and Tanner [8–11] extended most of the maximum likelihood estimation results proved by Jiang and Tanner to time series applications. Besides, [7, 10] presented parameter conditions to guarantee stochastic stability of the ME construction. In these papers, the authors also treated exponential family distributions, focusing on normal, Poisson, gamma, and binomial autoregressive processes.

By using mixtures of regressions of one of these four distributions, it is possible to treat a great variety of time series problems. Mixtures of binomial experts can be used to model discrete time series with response y_t bounded by some value ν (see, e.g., [51]), whereas mixtures of Poisson experts can be used to model unbounded discrete time series. For continuous responses, we can use mixtures of normal experts for observations assuming values on the whole real line, and mixtures of gamma experts for strictly positive time series. For unbounded count data, mixtures of Poisson experts present an advantage over several models in the literature since the proposed mixture construction allows for both positive and negative autocorrelations, while most of the existing count data models allow only for positive autocorrelation (see, e.g., [3, 32]). Besides, most of count time series models have likelihood functions that are difficult to write explicitly, and computational intensive approaches have to be used. This problem does not happen in the ME context and standard maximization algorithms can be employed for parameter estimation (see, e.g., [34]).

The ME models bear some similarity to other nonlinear models in the literature. We can mention, for example, the threshold autoregressive (TAR) models introduced by [49], where a threshold variable controls the switching between different autoregressive models. Another example is the Bayesian-treed model introduced by [12], where the input space is split in several subregions and a different regression model is fitted in each subregion. In both approaches, after the partition of the covariate space, a different regression curve is fitted in each subregion.

In this paper, we present a survey of the main ideas and results involved in the usage of the ME class of models for time series data. The discussion combines analytical results, simulation illustration, and real applications examples. In Section 2, we provide a more formal definition of ME of time series, with exponential family distributions. Section 3 discusses the probabilistic behavior, focusing on stochastic stability and moment existence, for ME time series models. Section 4 discusses parameter estimation (or model training) using maximum likelihood. In Section 5, Monte Carlo simulations provide evidence to support the BIC in selecting the number J of mixed components. In Section 6, several examples using real and simulated data illustrate how ME can be employed both for data description, where the underlying latent variable I_t may be particularly important, as well as for prediction. Final comments and suggestions for future research are presented in Section 7.

2. Mixtures-of-experts of time series models

In the models discussed here in this paper, the observed stochastic process $y_t \in \mathcal{S} \subset \Re$ has the conditional distribution, given the available information set $\mathbb{I}_{t-1}$, following the conditional density specification in (1.1), where the vector of covariates x_t includes functions of lags of y_t. This formulation follows the specification proposed by [36] for time series based on generalized linear models. The vector x_t at time t has the form $\{\zeta(y_{t-1}),\ldots, \zeta(y_{t-p}), w_{t-1},\ldots, w_{t-q}\}$, where w_t is a vector of external covariates, $\zeta(\cdot)$ is a transformation of the response y_t, and p and q correspond to the maximum lags. Because the covariate vector is known at time $t-1$ ($x_t \in \mathbb{I}_{t-1}$), hereinafter we will use the notation x_{t-1} instead of x_t for the conditioning vector of predictors.

We assume that the densities $\pi(y_t;\eta(\alpha_j + x'_{t-1}\beta_j),\varphi_j)$, $j = 1,\ldots,J$, belong to the same family, but the parameters $(\alpha_j,\beta'_j,\varphi_j)$ are different for different j's. The gate functions $g_j(x_{t-1};\gamma)$ are assumed to have a logistic form in (1.2). The grand vector of gating parameters is the list of all the individual gating parameters $\gamma = (\upsilon_1,u'_1,\upsilon_2,u'_2,\ldots,\upsilon_{J-1},u'_{J-1})'$.

Examples of ME of exponential family distributions can be based on the experts $\pi(y;\eta,\varphi)$.

Poisson distribution with logarithmic link function. $\eta(\alpha_j + x'_{t-1}\beta_j) = \exp(\alpha_j + x'_{t-1}\beta_j)$, $y_t \in \mathcal{S} = \{0,1,2,3,\ldots\}$, $\varphi_j = 1$ (known dispersion parameters), and

$$\pi(y;\eta,\varphi) = \frac{e^{-\eta}}{y!}\eta^{y}. \tag{2.1}$$

Normal distribution with identity link function. $\eta(\alpha_j + x'_{t-1}\beta_j) = \alpha_j + x'_{t-1}\beta_j$, $\varphi_j = \sigma_j^2 \in (0,+\infty)$, $y_t \in \mathcal{S} = \Re$, and

$$\pi(y;\eta,\varphi) = \left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\frac{(y-\eta)^2}{2\sigma^2}\right\}. \tag{2.2}$$

Gamma distribution with logarithmic link function. $\eta(\alpha_j + x'_{t-1}\beta_j) = \exp(\alpha_j + x'_{t-1}\beta_j)$, $y_t \in \mathcal{S} = (0,+\infty)$, $\varphi_j = \gamma_j \in (0,+\infty)$, and

$$\pi(y;\eta,\varphi) = \frac{(\gamma/\eta)^{\gamma}y^{\gamma-1}}{\Gamma(\gamma)}\exp\left(-\frac{y\gamma}{\eta}\right). \tag{2.3}$$

Binomial distribution (ν trials) with logistic link function. $y_t \in \mathcal{S} = \{0,\ldots,\nu\}$, $\varphi_j = 1$ (known dispersion parameters), $\eta(\alpha_j + x'_{t-1}\beta_j) = \nu\{e^{\alpha_j + x'_{t-1}\beta_j}/(1+e^{\alpha_j + x'_{t-1}\beta_j})\}$, and

$$\pi(y;\eta,\varphi) = \binom{\nu}{y}\left(\frac{\eta}{\nu}\right)^{y}\left(1-\frac{\eta}{\nu}\right)^{\nu-y}. \tag{2.4}$$

For the Poisson case, we can use the transformation $\zeta(y_{t-k}) = \log(y_{t-k}+1)$, while for the gamma case we can use $\zeta(y_{t-k}) = \log y_{t-k}$. These transformations allow for convenient properties in terms of stochastic stability (see [10]) of the mixed process. For the normal and gamma cases, the dispersion parameter is unknown and must be estimated from the data.

The grand vector of parameters for the whole model is $\theta \in \Theta \subset \Re^K$, where θ is the union of all the components $\theta_j = (\alpha_j, \beta'_j, \varphi_j)$, $j = \{1,\dots,J\}$, and γ. The dimension of Θ is $K = J(2+s) + (J-1)(1+s)$. For models with known dispersion parameters, θ has J fewer elements. From the density in (1.1), the conditional expectation for the response y_t is

$$\mu(x_{t-1};\theta) = E(y_t \mid x_{t-1}) = \sum_{j=1}^{J} g_j(x_{t-1};\gamma)\eta(\alpha_j + \beta'_j x_{t-1}), \tag{2.5}$$

and higher moments can be obtained by similar expressions.

Identifiability of the models treated here can be obtained by following the steps in [9, 30]. Because of the mixture structure, we have to impose some order constraints for the experts parameters $\theta_j = (\alpha_j, \beta'_j, \varphi_j)$, $j = 1,\dots,J$, that is, we assume $\theta_1 \prec \theta_2 \prec \cdots \prec \theta_J$ according to some order relation, so there is no invariance caused by the permutation of expert indices. We can impose, for example, an order relation of the following form: if $\alpha_j < \alpha_k$, then $\theta_j \prec \theta_k$; if $\alpha_j = \alpha_k$ and $\beta_{j,1} < \beta_{k,1}$, then $\theta_j \prec \theta_k$; if $\alpha_j = \alpha_k$, $\beta_{j,1} = \beta_{k,1}$, and $\beta_{j,2} < \beta_{k,2}$, then $\theta_j \prec \theta_k, \dots,$ if $\alpha_j = \alpha_k$, $\beta_{j,1} = \beta_{k,1}, \dots, \beta_{j,s} = \beta_{k,s}$, and $\varphi_j < \varphi_k$, then $\theta_j \prec \theta_k$, for all $j,k \in \{1,\dots,J\}$. (As will be discussed in Section 4, parameter estimation can be performed by using maximum likelihood methods. For maximizing the likelihood function, heuristic optimization methods, such as simulated annealing or genetic algorithms, can be employed. In this case, the ordering relation can be imposed directly in the objective function, by using, e.g., the parameterization $\alpha_1 = \alpha$, $\alpha_2 = \alpha + e^{\kappa_2}, \dots, \alpha_J = \alpha + e^{\kappa_J}$, where the new parameters to be estimated are $\alpha, \kappa_2, \dots, \kappa_J$, instead of $\alpha_1, \dots, \alpha_J$. We opted for a simpler approach, where we employ the EM algorithm to the unrestricted maximization problem. One could rearrange the parameter estimates after the EM solution is obtained, so as to impose the ordering relation. However, in practical terms there is no need to do so, and we decided just to use directly the estimated parameters.) Additionally, to guarantee identifiability of the gate parameters, we impose the initialization constraint as presented in (1.3). Finally, given the dependence of the mean function of the exponential family distributions on a vector of covariates x_{t-1}, we need some additional constraints on the marginal distribution of x_{t-1}. Basically, the conditions are imposed so that we do not allow for linear dependence among the elements of vector $(1, x'_{t-1})$.

3. Probabilistic behavior

Stochastic stability properties of the ME of time series models can be studied based on the general results for stochastic processes given in [17, 40]. These properties are specially important, for example, when treating the asymptotic behavior of the maximum likelihood estimators for the model parameters. Especially for ME of autoregressive linear models, which is the case when each expert has a normal distribution, some results are presented initially in [57], and extended in [7]. In a nutshell, these authors show that stationarity of each autoregressive model individually guarantees stationarity and existence of moments for the ME structure. Nonetheless, for mixture models, with constant gate functions (not depending on covariates), reference [54] shows that, even with not all mixed experts being stationary, it is still possible to combine them in a mixture model and obtain a stationary process in the end.

Although stochastic stability for autoregressive linear models can be proved for experts with an arbitrary number p of lags, extending these results to other exponential family distributions is not trivial, since linearity plays a key role in going from one-lag models to multiple-lag models (see [7]). The exception are stochastic processes with bounded sample spaces (e.g., mixtures of Bernoulli or binomial experts). For Poisson and gamma experts, reference [10] shows that, given some simple parameter restrictions on ME models, where each expert has only one lag, stochastic stability holds and the resulting observed process has a moment generating function, and therefore all moments exist.

3.1. Simulated time series. In this section, we present a simulated example to illustrate the capability of mixtures-of-experts models to approximate the behavior of various time series data. Although we do not present a more thorough discussion of the approximation theory for the mixtures-of-experts, the example below, involving normal experts, gives an idea about the flexibility implied by the proposed construction. The reader can refer to [28, 29, 56] for related approximation results. For similar examples on simulated data from ME of Poisson autoregressions, see [11].

We simulate a mixture of two Gaussian autoregressions

(1) $y_t = 3.0 + 0.5y_{t-1} + \epsilon_{1,t}$,

(2) $y_t = -3.0 + 0.5y_{t-1} + \epsilon_{2,t}$,

where $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are normally distributed with mean 0 and unit variance. The gate functions are

(1) $g_1(y_{t-1}) = \exp(0.9y_{t-1})/(1 + \exp(0.9y_{t-1}))$,

(2) $g_2(y_{t-1}) = 1 - g_1(y_{t-1})$.

The upper graph in Figure 3.1 presents the plot of 10 000 observations of the simulated series. (In order to estimate the marginal density of $\{y_t\}$, we simply used a kernel estimator based on the generated time series. Given that the process $\{y_t\}$ is stationary and all moments exist (Carvalho and Skoulakis [7]), we can use the generated series to estimate nonparametrically the density for the marginal process. To have a better precision in these estimates, we used 40 000 time points after the warm-up sample.) To guarantee that the series reaches stationarity, we initially generated 100 000 warm-up data points. The middle graph presents an estimate for the marginal density of $\{y_t\}$. (Depending on the parameter configuration, a warm-up sample of 100,000 observations may be excessive. Nonetheless, given the speed of sample generation, we decided to use a large number to guarantee that the series achieves stationarity.) Note the clear existence of two regimes in the series, which is very similar to the behavior of hidden Markov processes. In fact, when y_t is close to 6.0 (the stationary mean for the first autoregression), the weight for the positive autoregression (first component) is close to one, as can be seen from the lower graph in Figure 3.1, so that the series tends to keep following the first autoregression. Analogously, when y_t is close to -6.0, the weight $g_2(y_{t-1})$ is close to 1, and the series tends to behave according to the second autoregression.

To have an idea about how different parameter values change the observed time series, we simulated a model similar to the mixture of two experts above, using an autoregressive coefficient of 0.6 instead of 0.5. The results are shown in Figure 3.2. Observe that, for a higher autoregressive coefficient, the frequency for regime changes decreases. This is

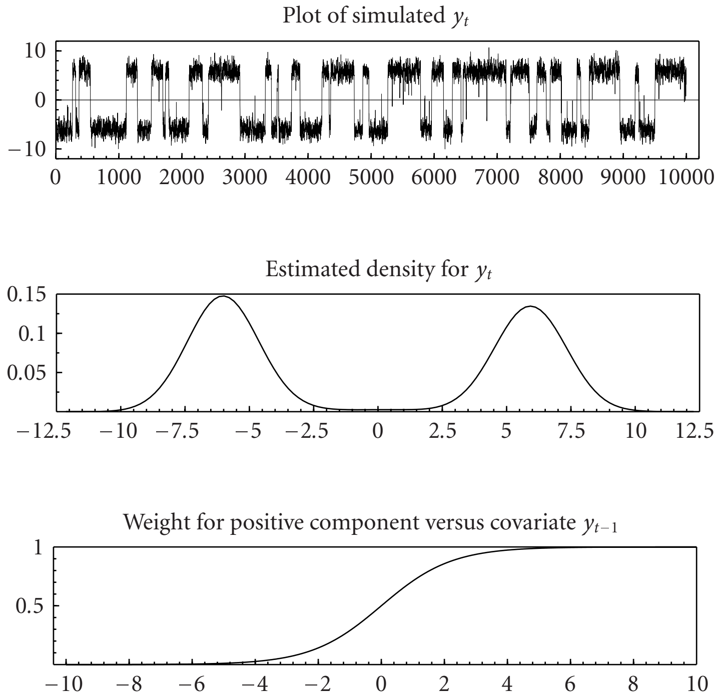


Figure 3.1. Generated time series (a), density estimate for the observed series (b), and weight $g_1(y_{t-1})$ for the positive mean autoregression (c) in the first example. The autoregressive coefficient for each expert is assumed to be 0.5.

because when the autoregressive coefficient changes from 0.5 to 0.6, the stationary mean for the first expert becomes 7.5 and the stationary mean for the second expert becomes −7.5. Therefore, regime changes become less likely, because it becomes more difficult for an observed y_t to jump to regions in $\Re$ where the weight for the other expert is sufficiently high. Some additional experiments show that, for autoregressive coefficients closer to 1.0, the probabilities of regime change are even lower.

4. Maximum likelihood estimation

Estimation of the parameter vector θ for the ME of time series models studied in this paper can be performed by maximizing the partial likelihood function [53]. (We are using the partial likelihood function because we are modeling only the conditional process of y_t given x_{t-1}. We are not using the overall likelihood function, where we model the stochastic process for both y_t and x_{t-1} jointly.) From the density in expression (1.1), we can write down the conditional likelihood function based on a sample $\{y_t, x'_{t-1}\}_{t=1}^{T}$. If the vector x_{t-1} contains functions of lags of the response variable y_t, such that the maximum lag order is p, we will require $T + p$ observations so that our sample has an effective size

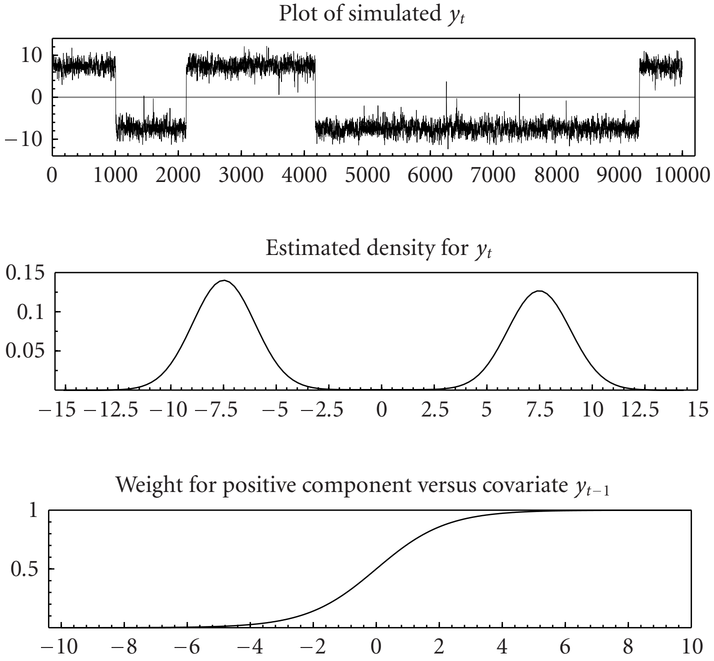


Figure 3.2. Generated time series (a), density estimate for the observed series (b), and weight $g_1(y_{t-1})$ for the positive mean autoregression (c) in the first example. The autoregressive coefficient for each expert is assumed to be 0.6.

T. The likelihood function $\log L_T(\theta)$ to be maximized over Θ is given by

$$\sum_{t=1}^{T} \log\left[\sum_{j=1}^{J} g_j(x_{t-1};\gamma)\pi\left(y_t;\left(\alpha_j + x'_{t-1}\beta_j\right),\varphi_j\right)\right]. \tag{4.1}$$

Numerical optimization can be performed by applying the EM algorithm (see [25, 33]), described in Section 4.1. (In this paper, we focus on the frequentist approach, employing the maximum likelihood for parameter estimation. However, one can also use Bayesian methods, which present various nice properties as discussed in Section 7.) In Section 4.2, we discuss formal results for the asymptotic properties of the maximum likelihood estimator.

4.1. The EM algorithm. For simple problems, where the parameter space is low-dimensional, maximization of log-likelihood function in (4.1) can be performed directly by using some standard optimization algorithm, such as Newton-Raphson. However, in most of the practical problems, the dimension of Θ is high enough so that the usual optimization methods become very unstable. The alternative, commonly used in mixture of distribution models, is the EM algorithm, proposed by [15]. The use of the EM algorithm for mixtures-of-experts models is thoroughly described in [25, 33], and that is the procedure used here for estimation.

To initialize the EM algorithm, choose a starting value θ^0 for the parameter vector $\theta = (\theta_1', \dots, \theta_J', \lambda')'$. Then, obtain the sequence $\{\theta^i\}$ iterating between Step 1 (expectation step) and Step 2 (maximization step), for $i = 0, 1, 2, \dots$.

Step 1. Construct

$$Q^i(\theta) = \sum_{t=1}^{T} \sum_{j=1}^{J} h_{j,t}(\theta^i) \log \pi(y_t \mid x_{t-1}; \theta_j) + \sum_{t=1}^{T} \sum_{j=1}^{J} h_{j,t}(\theta^i) \log g_j(x_{t-1}; \lambda), \tag{4.2}$$

where

$$h_{j,t}(\theta) = \frac{g_j(x_{t-1}; \lambda) \pi(y_t \mid x_{t-1}; \theta_j)}{\sum_{l=1}^{J} [g_l(x_{t-1}; \lambda) \pi(y_t \mid x_{t-1}; \theta_l)]}. \tag{4.3}$$

Step 2. Find $\theta^{i+1} = \arg\max_{\theta \in \Theta} Q^i(\theta)$.

Note that, at each iteration i, maximization of $Q^i(\theta)$ in (4.2) can be obtained by maximizing separately the J terms Q_j^i, corresponding to parameters for each expert distribution individually,

$$Q_j^i(\theta_j) = \sum_{t=1}^{T} h_{j,t}(\theta^i) \log \pi(y_t \mid x_{t-1}; \theta_j), \tag{4.4}$$

and the term Q_{gates}^i, corresponding to the parameter vector λ for the gating functions,

$$Q_{\text{gates}}^i(\lambda) = \sum_{t=1}^{T} \sum_{j=1}^{J} h_{j,t}(\theta^i) \log g_j(x_{t-1}; \lambda) = \sum_{t=1}^{T} \sum_{j=1}^{J-1} h_{j,t}(\theta^i) \log \left[\frac{e^{z_{t-1}' \omega_j}}{\sum_{k=1}^{J} 1 + e^{z_{t-1}' \omega_k}} \right], \tag{4.5}$$

where $\omega_j = (v_j, \mathbf{u}_j')'$, $\lambda = (\omega_1', \omega_2', \dots, \omega_{J-1}')'$, and $z_{t-1} = (1, x_{t-1}')'$. We used the notation $\pi(y_t \mid x_{t-1}; \theta_j) = \pi(y_t; \alpha_j + x_{t-1}' \beta_j, \varphi)$ so as to make explicit the dependence on the target parameter θ_j.

Therefore, the EM algorithm in our case consists of calculating, at each iteration i, the weights $h_{j,t} \in (0,1)$, $j = 1, \dots, J$, $t = 1, \dots, T$, and then maximizing the functions $Q_1^i(\theta_1), \dots, Q_J^i(\theta_J), Q_{\text{gates}}^i(\lambda)$, to find the new value θ^{i+1}. Maximizing $Q_j^i(\theta_j)$ can be seen as a weighted maximum likelihood estimation, where each observation in the sample is weighted by its corresponding gating function value. Maximizing $Q_{\text{gates}}^i(\lambda)$ corresponds to estimating a multinomial logistic regression. The limit of the sequence $\{\theta^i\}$, denoted by $\widehat{\theta}(\theta^0)$, is a root of the first-order condition $\partial_\theta \log L_T(\theta) = 0$ (see [47]).

When the log-likelihood function is multimodal, the limits $\widehat{\theta}(\theta^0)$ may not correspond to the global maximum of the log-likelihood function, so we used multiple starting points to initialize the algorithm. In this case, the point with maximum likelihood from multiple points is an approximation to the global maximum, and the maximum likelihood estimator $\widehat{\theta}$ is approximately the root corresponding to the largest likelihood value $L_T(\widehat{\theta}(\theta^0))$. Alternatively, one can resort to heuristic algorithms such as genetic algorithms [20, 41] and simulated annealing [50]. Besides, several methods which take advantage of the specificities of the mixtures structures in ME models are also available [46].

4.2. Asymptotic properties of the MLE. Given the simple structure of the likelihood function for ME models, the main method for parameter estimation is via maximum likelihood. By using the EM algorithm or any other global search heuristic method, maximizing the log-likelihood function is a rather simple task and does not involve maximizing simulated likelihoods. Therefore, it is expected that the MLE will present all the nice asymptotic properties of regular parametric models, for example. In fact, that is exactly what happens.

Carvalho and Tanner [9–11] present a series of very general results guaranteeing consistency and asymptotic normality of the MLE for several different situations. In fact, given stationarity and ergodicity of the conditioning series (predicting variables) $\{x_{t-1}\}_{t=0}^{\infty}$ and some hypotheses about moment existence of $\{x_{t-1}\}_{t=0}^{\infty}$, both consistency and asymptotic normality hold:

$$\begin{gathered}\widehat{\theta}_{\text{MLE}} \xrightarrow{P} \theta_0, \\ \sqrt{T}\left[\widehat{\theta}_{\text{MLE}} - \theta_0\right] \xrightarrow{L} N\left(0, \mathbf{I}^{-1}\right),\end{gathered} \tag{4.6}$$

where $\mathbf{I} \equiv -E\{\partial_\theta \partial_{\theta'} \log f(y_t | x_{t-1}; \theta_0)\}$ is the Fisher information matrix and θ_0 is the true parameter value. By imposing the existence of a little higher moments, the same results hold for nonstationary time series.

If we assume that there is a single parameter θ^* that minimizes the Kullback-Leibler pseudodistance, [9, 10] show that, under some regularity conditions on the true data generating processes, consistency and asymptotic normality of the MLE still hold. In this case, if one is interested in statistical inference, such as hypothesis testing or confidence interval construction, the asymptotic covariance matrix of $\widehat{\theta}_{\text{MLE}}$ is no longer the Fisher information matrix, and some correction has to be done (see [8]).

More generally, one can show that if the Kullback-Leibler pseudodistance [6] achieves a global minimum at all the elements of a nonempty set Ψ_0, the maximum likelihood estimator is consistent to Ψ_0, in the sense that $\mathrm{P}\{\min_{\theta \in \Psi_0} | \widehat{\theta} - \theta| < \epsilon\} \to 1$ as $T \to \infty$, for any $\epsilon > 0$. (see, e.g., [28]). The importance of this fact is that even if there is more than one parameter θ^* resulting in the best approximation for the true data generating process, the maximum likelihood will provide a parameter estimate close to one of these best approximation parameters, which is important for prediction purposes.

5. Selecting the number of experts

The selection of the correct number of experts has no easy answer. Basically, log-likelihood ratio tests are not applicable in this case, as long as, under the null hypothesis of fewer experts, the alternative hypothesis implies a nonidentified problem (see, e.g., [43]). We will examine the use of information criteria such as BIC [45] or AIC [1, 2] in selecting the right number of experts.

In [55], BIC is used to select the number of experts for spatially adaptive nonparametric regression models. For well-behaved models, we know that the BIC is consistent for model selection, since, with probability tending to one as the sample size goes to infinity, the true model will be chosen because it has the largest BIC. However, when the model is overidentified, the usual regularity conditions to support this result fail. Fortunately, [55] presents some evidence that, even when we have overidentified models, the BIC may still be consistent for model selection.

In this section, we present some results about the Monte Carlo simulations to evaluate the performance of the Bayesian information criteria (BIC) and the Akaike information criteria (AIC) in selecting the right number of mixed experts. We performed simulations under true models composed by three experts and, for each generated data set, we estimated mixtures-of-experts models with various number of mixed distributions. For each simulated data set, we stored the BIC and the AIC values. We expect that one of the two criteria (or both) will present the smallest value for the estimated model with the same number of experts as the simulated true model. We performed simulations for normal and binomial distributions. We report that simulations for other distributions presented similar conclusions. For each true model, we generated 400 data sets, with $T = 100$ and $T = 200$ observations. Each model includes an external covariate x_t, which was generated as an autoregressive process of order 1, with autoregressive coefficient equal to 0.5.

For the normal distribution, the expressions for the experts $(y_{j,t})$ and for the gates $(g_j(x_{t-1};\gamma))$, assuming three experts in the true model, are

$$\begin{gathered} y_{1,t} = 3.0 + 0.4y_{t-1} + 1.2x_{t-1} + \epsilon_{1t}, \qquad y_{2,t} = -2.0 + 0.7y_{t-1} - 1.1x_{t-1} + \epsilon_{2t}, \\ y_{3,t} = 1.0 - 0.6y_{t-1} + 0.5x_{t-1} + \epsilon_{3t}, \end{gathered} \tag{5.1}$$

with gate functions

$$\begin{gathered} \xi_{1,t} = -2.0 - 0.3y_{t-1} + 0.4x_{t-1}, \qquad \xi_{2,t} = -1.1 + 0.1y_{t-1} + 0.1x_{t-1}, \\ g_1(x_{t-1};\gamma) = \frac{\exp(\xi_{1,t})}{1+\exp(\xi_{1,t})+\exp(\xi_{2,t})}, \qquad g_2(x_{t-1};\gamma) = \frac{\exp(\xi_{2,t})}{1+\exp(\xi_{1,t})+\exp(\xi_{2,t})}, \\ g_3(x_{t-1};\gamma) = 1 - g_{1t} - g_{2t}, \end{gathered} \tag{5.2}$$

where $\epsilon_{1t} \sim N(0,2.0)$, $\epsilon_{2t} \sim N(0,1.5)$, and $\epsilon_{3t} \sim N(0,1.5)$. The results for the normal distribution with three experts are presented in Table 5.1.

For the binomial case with three experts, the expressions for the experts $y_{j,t}$'s and for the gating functions are presented below. In all models, we considered 50 trials for the

Table 5.1. Mixtures of 3 normal experts.

Selected number of experts	$T = 100$			
	BIC		AIC	
	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency
2	0	0%	0	0%
3	365	91.25%	21	5.25%
4	35	8.75%	379	94.75%
Total	400	100.00%	400	100.00%
Selected number of experts	$T = 200$			
	BIC		AIC	
	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency
2	0	0%	0	0%
3	391	97.75%	29	7.25%
4	9	2.25%	371	92.75%
Total	400	100.00%	400	100.00%

binomial random variables:

$$\begin{aligned} E\{y_{1,t} \mid y_{t-1}, x_{t-1}\} &= \left[\frac{e^{4.0-0.5y_{t-1}-0.2x_{t-1}}}{1+e^{4.0-0.5y_{t-1}-0.2x_{t-1}}}\right], \\ E\{y_{2,t} \mid y_{t-1}, x_{t-1}\} &= \left[\frac{e^{1.5-0.6y_{t-1}-3.0x_{t-1}}}{1+e^{1.5-0.6y_{t-1}-3.0x_{t-1}}}\right], \\ E\{y_{3,t} \mid y_{t-1}, x_{t-1}\} &= \left[\frac{e^{1.0-0.5y_{t-1}-0.1x_{t-1}}}{1+e^{1.0-0.5y_{t-1}-0.1x_{t-1}}}\right], \end{aligned} \tag{5.3}$$

with gate functions

$$\begin{gathered} \xi_{1,t} = -1.5 + 0.2y_{t-1} + 0.4x_{t-1}, \qquad \xi_{2,t} = 1.5 - 0.2y_{t-1} - 1.2x_{t-1}, \\ g_1(x_{t-1};\gamma) = \frac{\exp(\xi_{1,t})}{1+\exp(\xi_{1,t})+\exp(\xi_{2,t})}, \qquad g_2(x_{t-1};\gamma) = \frac{\exp(\xi_{2,t})}{1+\exp(\xi_{1,t})+\exp(\xi_{2,t})}, \\ g_3(x_{t-1};\gamma) = 1 - g_{1t} - g_{2t}. \end{gathered} \tag{5.4}$$

The results for binomial experts are summarized in Table 5.2. As can be seen from the tables, the BIC performed very well in selecting the correct number of mixed experts, for the two distributions studied in the simulations. The AIC tends to pick more experts than needed, especially in the normal case. Therefore, the use of the BIC seems to be very appropriate for model selection in this case, and its performance tends to improve as the sample size T increases. We replicated similar experiments with true models containing

Table 5.2. Mixtures of 3 binomial experts.

Selected number of experts	$T = 100$			
	BIC		AIC	
	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency
2	0	0%	0	0%
3	399	99.75%	291	72.75%
4	1	0.25%	109	27.25%
Total	400	100.00%	400	100.00%

Selected number of experts	$T = 200$			
	BIC		AIC	
	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency
2	0	0%	0	0%
3	400	100.00%	288	72.00%
4	0	0%	112	28.00%
Total	400	100.00%	400	100.00%

one and two experts, and with other distributions (Poisson and gamma), and the conclusions are basically the same. For some of these distributions, we also simulated samples with 1 000 observations and noticed that the BIC still selected the true number of experts for 100% of the samples, while the AIC continued to present a bias towards selecting a higher number of experts. These results seem to suggest the consistency of the BIC for selecting the number of components. This conclusion agrees, for example, with the results presented in [14], where the authors show that the BIC is an almost surely consistent for estimating the order of a Markov chain.

6. Applications

In this section, we present examples where ME are used to model different time series. In the first example, we present an application of mixtures of binomial experts (for applications using ME of Poisson experts, see [10, 11]), where we are interested not only in predicting the response variable, but also in presenting some plausible description of the data-generating process, based on the stochastic underlying mixture structure behind the mixtures-of-expert models. In these cases, the latent variable I_t, which determines which regime (or expert) is observed, has a meaning and helps us interpret the results.

In example two, we are not interested in explaining the data anymore, but only in using a flexible modeling structure, such as ME, so as to approximate and predict the conditional density function of the observed process. In this case, the underlying latent variable I_t has no meaning, but only the functional form for the density in (1.1). For the simulated time series, clearly the data-generating process does not follow a ME model. However, as we will discuss in these examples, we are still able to reasonably approximate the conditional process.

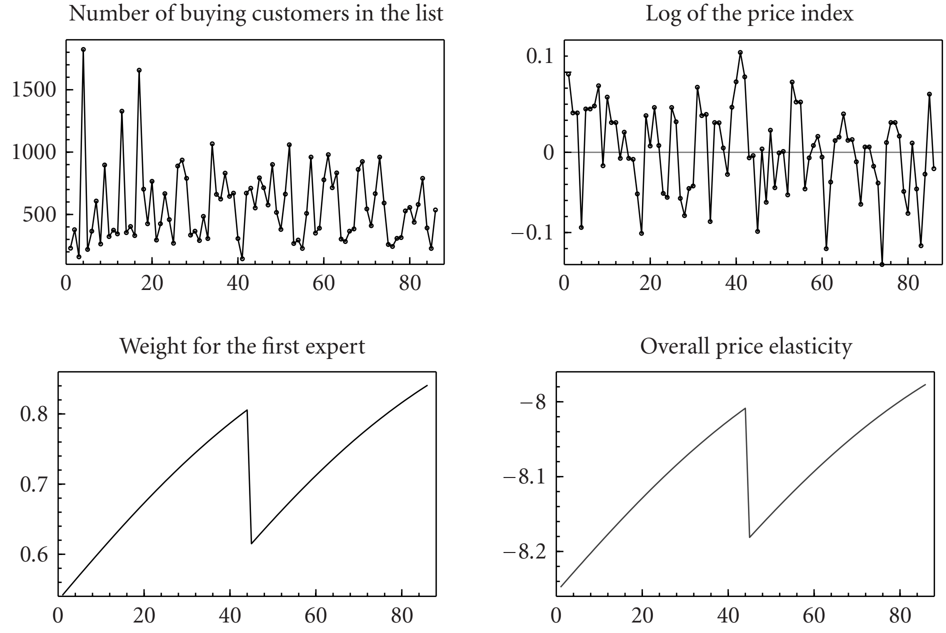


Figure 6.1. Mixture of 2 binomial experts.

6.1. Number of buying customers. In this example, we consider the problem of modeling the buying behavior of a list of regular customers in a supermarket. Basically, we have 20 months of scanner data, and we selected a list of 3 497 regular customers who bought one of 8 brands of canned tuna at least once. By regular customers, we mean the customers that appeared in the store during the 20 months. Therefore, we have a binomial time series, where the response variable y_t (see upper-left graph in Figure 6.1) is the number of customers buying one of the 8 brands on week t. The number of trials $\nu = 3\,497$.

One natural covariate in this case is some price index. For these 8 brands, we have 8 different prices in each week (promotions are launched in a weekly basis). The price index was composed by finding a weighted average of the 8 individual prices, with weights given by the overall market shares during the 20 months. After calculating the weighted average, we applied a logarithm transformation, obtaining the covariate p_t (some preliminary estimations have shown that using the logarithm of prices provides better values for the BIC and AIC than using the untransformed prices). The logarithm of the price index is presented in the upper-right graph in Figure 6.1.

To model some nonstationarities in the data, we also included, in the vector of predictors, a linear time trend t, $1 \le t \le 86$, where t is the number of the focused week. Finally, in the middle of the overall period, one competitor entered the neighborhood, which may have caused some impact in the buying behavior of the list of customers, and to model the competitor effect, we used a dummy variable d_t, with $d_t = 0$, if $t \leq 42$, and $d_t = 1$ otherwise.

After trying different numbers of experts and different numbers of lags for each predicting variable, the resulting model is a mixture of two experts (all parameters are significant with level 1%). (In selecting the final model, we employed the BIC for choosing the number of experts and the t-statistics for selecting the number lags, starting from an initial model with high number of lags. This procedure was based on the general-to-specific approach, commonly used for model building in econometrics (see [13, 24]). Nonetheless, there is still need for further research on model selection in ME models, as discussed in Section 7.) The two binomial regressions for each expert are given by

$$y_{j,t} \sim \operatorname{Bin}\left(\nu = 3{,}497;\ \frac{e^{h_{j,t}}}{1+e^{h_{j,t}}}\right), \quad j = 1,2, \tag{6.1}$$

where

$$\begin{aligned} h_{1,t} &= -1.9971 - 7.8328\log(p_t) + 2.5268\log(p_{t-1}) + 1.0988\log(p_{t-2}), \\ h_{2,t} &= -1.0908 - 8.7378\log(p_t) + 2.4033\log(p_{t-1}) + 1.0322\log(p_{t-2}). \end{aligned} \tag{6.2}$$

As expected, the contemporary price elasticities are negative, which implies the effectiveness of price reductions in increasing the number of buying customers. Observe that the second expert presents a higher-price sensitivity. Both regressions present significant positive coefficients for the first and second lags of the logarithm of the price index, what implies the existence of a dynamic effect. Basically, the inclusion of the lags of the price index suggests that if there is a promotion in the current week, some of the customers buy and stock up canned tuna so that, even if there is another price reduction the next week, the price effect will not be so pronounced.

The gating function, corresponding to the weight of the first expert, is given by

$$g_{1,t} = \frac{\exp(0.1375 + 0.0292t - 0.9828d_t)}{1+\exp(0.1375 + 0.0292t - 0.9828d_t)}. \tag{6.3}$$

Intuitively, we can regard the overall price elasticity as a linear combination of the price elasticities in both experts, weighted by the corresponding gate functions. Therefore, when we increase the weight for the first expert, we decrease the overall price sensitivity. As we can note from the expression for $g_{1,t}$ (see lower graphs in Figure 6.1), the price sensitivity decreases with the time trend and increases with the entrance of the competitor.

The above conclusion about the positive effect that the competitor caused in the overall price sensitivity is quite surprising, if we take into account the fact that the competitor has the tradition of being a more inexpensive store. Basically, we expect that the competitor will attract the more price sensitive customers, so the remaining tuna buyers will be less-price sensitive. One plausible explanation for this contraction can be found by looking at the plot for the logarithm of the price index in the upper-right graph in Figure 6.1. Apparently, the studied store changed its price strategy, increasing the number of promotions after the appearance of the competitor. Actually, the averages of the logarithm of the

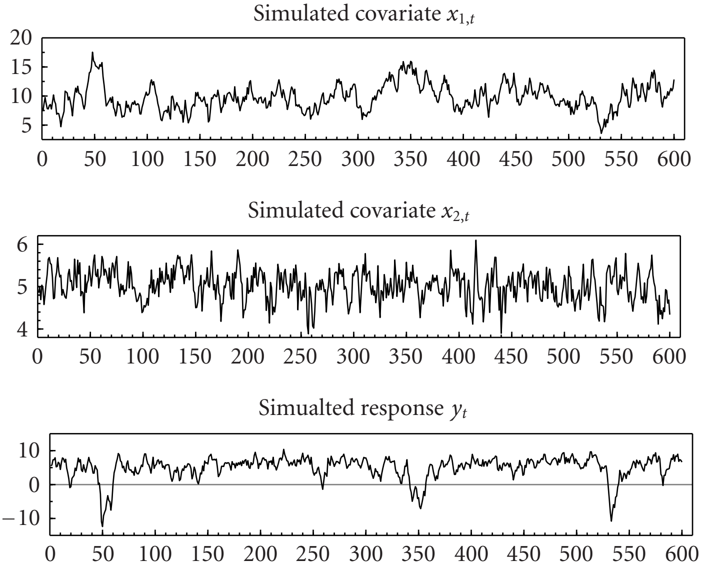


Figure 6.2. Simulated time series for both covariates the $x_{1,t}$ and $x_{2,t}$ and the response y_t.

price index before and after the competitor are 0.0169 and −0.0114. In this way, it seems that the studied store regained its price sensitive customers.

6.2. Simulated data. The following example illustrates the application of ME of Gaussian autoregressions to simulated data, so as to illustrate the ability of ME of time series in modeling the conditional density in time series processes. The artificial time series present nonlinearities not only in the conditional mean function but also in other conditional moments. To evaluate the performance of the estimated models, we present several graphical criteria as suggested in [16].

The simulated time series corresponds to a variance covariate dependent process. The response y_t and the covariates $x_{1,t}$ and $x_{2,t}$ obey the autoregressions

$$
\begin{gathered}
y_t = 3.0 + 0.6y_{t-1} - 0.2\left(x_{t-1} - 10.0\right)^2 + \sigma_t\epsilon_t, \quad \text{where } \sigma_t^2 = 1.0 + 0.9\left(x_{t-2} - 5\right)^4, \\
x_{1,t} = 1.0 + 0.9x_{1,t-1} + \epsilon_t, \qquad (6.4) \\
x_{2,t} = 2.0 + 0.6x_{2,t-1} + 0.3\epsilon_t, \quad \epsilon_t \sim N(0, 1.0).
\end{gathered}
$$

Note that there is an explicit nonlinearity in how the conditional variance of y_t depends on the second lag of $x_{2,t}$. Besides, the conditional mean function of y_t is a nonlinear function of the lagged covariate $x_{1,t}$. The simulated time series are presented in Figure 6.2.

We estimated a mixture of normal experts model to 600 observations. These observations were obtained from the data generating process in (6.4), after 100000 warm-up data points. By selecting the number of experts via BIC, the resulting model is a mixture of three experts, with lags up to order 2 of both the covariates $x_{1,t}$ and $x_{2,t}$ and the response y_t in the experts and in the gates.

In order to assess the goodness-of-fit of the estimated ME of normal autoregressions in modeling the simulated series studied in this paper, we use the methodology based on the probability integral transform, initially defined by [44]. This approach has been employed by a number of recent papers such as [4, 16]. The analysis is based on the relationship between the data generating process $f_t(y_t|x_{t-1})$, for the response variable y_t, and the sequence of estimated conditional densities $p_t(y_t|x_{t-1})$, obtained by using the mixture model. The probability integral transform u_t is the conditional cumulative distribution function corresponding to the density $p_t(y_t \mid x_{t-1})$ evaluated at the actual observed value y_t,

$$u_t = \int_{-\infty}^{y_t} p_t(v \mid x_{t-1})\,dv \equiv P_t(y_t \mid x_{t-1}). \tag{6.5}$$

We then have the following fact, a proof of which can be found in [16], which is the backbone for the model-checking analysis in this paper: if a sequence of density estimates $\{p_t(y_t|x_{t-1})\}_{t=1}^{T}$ coincides with the true data-generating process $\{f_t(y_t \mid x_{t-1})\}_{t=1}^{T}$, then under the usual conditions of nonzero Jacobian with continuous partial derivatives, the sequence of probability integral transforms $\{u_t\}_{t=1}^{T}$ of $\{y_t\}_{t=1}^{T}$ with respect to $\{p_t(y_t \mid x_{t-1})\}_{t=1}^{T}$ is i.i.d. $U(0,1)$.

In this paper, instead of working directly with the sequence $\{u_t\}_{t=1}^{T}$, we followed the suggestion in [4] and worked with the transformation $\{\Phi^{-1}(u_t)\}_{t=1}^{T}$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function. The aforementioned fact implies that $\{\Phi^{-1}(u_t)\}_{t=1}^{T}$ is a standard normal i.i.d. sequence. Therefore, after estimating the mixtures of autoregressive Gaussian experts, we evaluated the model fitting by checking for the hypothesis of independence and standard normality for the constructed series $\{z_t\}_{t=1}^{T}$, where $z_t = \Phi^{-1}(u_t)$, $t = 1,\dots,T$. Following [16], we employed a number of graphical methods for assessing goodness-of-fit. The analysis can be done by plotting the density estimate for the series z_t and comparing it to the standard normal density function. Our density estimation employs the Gaussian kernel and uses the optimal bandwidth for i.i.d. Gaussian data. Additionally, we also plotted the normal quantile plot for the series $\{u_t\}_{t=1}^{T}$ and compared it to the normal quantile plot for a standard normal variable. The two upper graphs in Figure 6.3 present the normal quantile plots (left-upper graph) and the density estimates (right-upper graph) for the simulated example.

To check for the independence hypothesis in the series $\{z_t\}_{t=1}^{T}$, we can plot the autocorrelation function for the series $(z_t - \bar{z})$, $(z_t - \bar{z})^2$, $(z_t - \bar{z})^3$, and $(z_t - \bar{z})^4$, as suggested by [16], where $\bar{z}$ is the sample mean for $\{z_t\}_{t=1}^{T}$. The four lower graphs in Figure 6.3 contain the plots of autocorrelation functions for the four trasformed series for the ME model applied to the two artificial time series, along with the corresponding 5% significance limits. For these limits, we used the approximation $\pm 1.96T^{-1/2}$ (see [5], for details).

According to Figure 6.3, the ME model seems to be a good approximation for the true data-generating process in the simulated example. Note that the normal quantile plots and the density plots seem to support the standard normality of $\{z_t\}_{t=1}^{T}$. Besides, the ACF

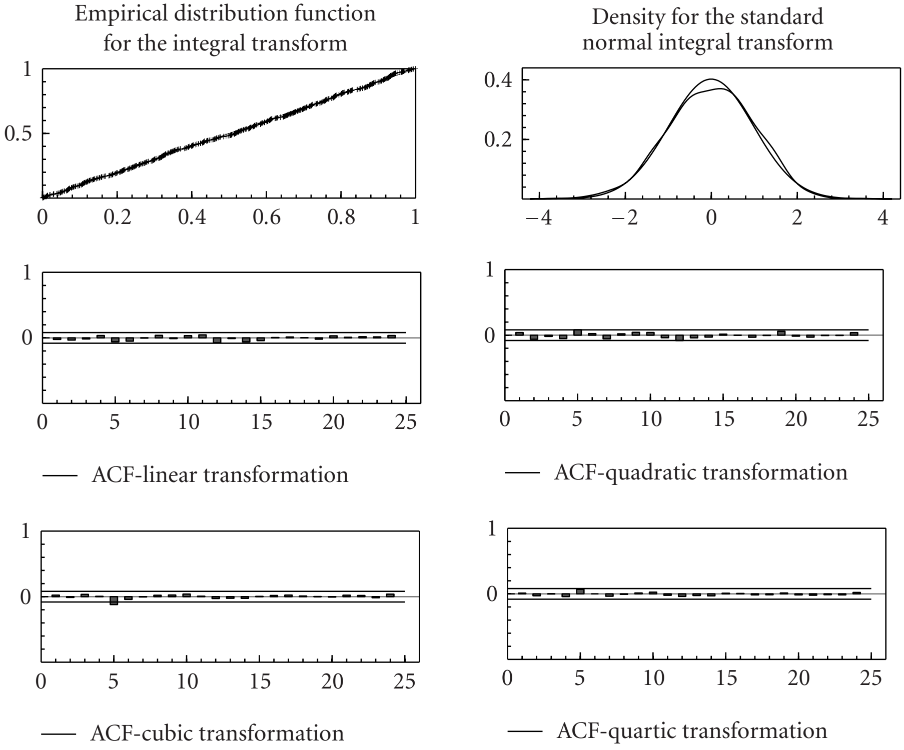


Figure 6.3. Density evalution and autocorrelation analysis for the simulated data.

plots seem to provide support for the independence of the contructed series $\{z_t\}_{t=1}^{T}$. For more examples on simulated and real data regarding ME of Gaussian autoregressions, see [9].

7. Final comments and suggestions for future research

In this paper, we discussed some of the recent results on a nonlinear class of models for time series data. This class is based on the idea of combining several simple models, in a mixture structure, where the weight for each model is a function of the covariates. Each combined simple model is called expert, whereas the weights are denoted as gates. The combined resulting model is denoted as mixtures-of-experts of time series. To incorporate time series dynamics, the covariates in the experts and in the gates may include lags or transformed lags of the dependent variable. Therefore, we can regard these models as nonlinear autoregressive structures, and they include several archtectures suggested in the literature [25, 51, 52, 54, 57].

Some simulated examples showed that, even with a relatively simple and intuitive structure, mixtures-of-experts can reproduce a great variety of time series behaviors, even with a small number of components. When the number of mixed components go to infinity, ME of time series models constitute a universal approximator for the conditional function of y_t given x_{t-1}, in the same way as artificial neural networks, stochastic neural networks, and other sieves-type models [28, 29]. However, because of the mixture construction, ME of time series models are also able to capture more than only approximations in the mean function. In fact, it may also capture multiple modes [54], asymmetries (skewed conditional distribution), heavy tails, and nonhomogeneity in higher conditional moments (e.g., conditional heterocedasticity). Moreover, one can easily extend the ideas presented in this paper and combine densities from different families, such as normal and gamma, or Poisson and binomial. Therefore, ME of time series models may be able to provide not only good approximations for the conditional-mean function, but also to provide good approximations to the entire conditional distribution of the response variable y_t. This fact was illustrated in this paper using a simulated example. More examples, with simulated and real data, can be found in [9].

We discussed several important results regarding model identification and stochastic stability for the ME of time series models. The main two assumptions for model identification are no two experts have the same parameter vector ($\theta_i \neq \theta_j$, for all $i \neq j$, $1 \leq i, j \leq J$); and the design matrix obtained from stacking the covariate vectors x_{t-1}'s is full rank with probability 1 [9, 11, 30]. For stochastic stability, a sufficient condition is that all autoregressive experts are stationary individually. Given that condition, no additional assumptions on the gates are necessary [7, 10]. Nonetheless, as [54] has pointed out, even when some of the experts are nonstationary, the whole system may still be stationary. Therefore, providing more general conditions for ME stability still remains an open question.

Parameter estimation of the ME model can be performed by maximum likelihood, employing the EM algorithm, which exploits the mixture construction. Alternatively, one can use heuristic methods for likelihood maximization (genetic algorithms, simulated annealing, etc.), instead of using the EM method. Several analytical results have been shown that, when the true data generating process follows a ME construction, maximum likelihood parameter estimates are consistent for the true parameters, and asymptotic normality holds. Additionally, even when the model is misspecified and the true data-generating process does not belong to a ME of time series family, the parameter estimates are still consistent and asymptotic normal. In this case, some easily computable corrections have to be done for the estimated covariance matrix. For more details, refer to [9, 10, 28, 31]. Finally, simulated examples show that BIC seems to be consistent for selecting the number of experts.

Several important questions still remain regarding ME of time series models. The analytical results for approximation capability and stochastic stability can be extended to more general conditions. Moreover, there is still work to be done on alternative estimation algorithms. Besides, model selection still deserves further investigation. Even though the BIC seems to be consistent in selecting the number of experts, there is still need for research on the selection of covariates, and on the selection of the number of lags.

In terms of estimation algorithms, one possibility is to use Bayesian techniques, which have been successfully employed for nonlinear time series models and for mixture models (see [19, 25, 39, 42], e.g.). One of the advantages of using Bayesian methods is that, in terms of forecasting k-steps ahead, the Markov chain Monte Carlo (MCMC) approach will automatically provide samples from the predictive distribution. Besides, one can employ reversible jump MCMC to obtain the posterior distribution for the number of experts in ME construction (see [21]). Some of these topics are under current investigation by the authors.

References

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, Second International Symposium on Information Theory (Tsahkadsor, 1971), Akadémiai Kiadó, Budapest, 1973, pp. 267–281.

[2] ______, *Likelihood of a model and information criteria*, Journal of Econometrics **16** (1981), 3–14.

[3] M. A. Al-Osh and A. A. Alzaid, *First-order integer-valued autoregressive (INAR(1)) process*, Journal of Time Series Analysis **8** (1987), no. 3, 261–275.

[4] J. Berkowitz, *Testing density forecasts, with applications to risk management*, Journal of Business & Economic Statistics **19** (2001), no. 4, 465–474.

[5] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, Springer Texts in Statistics, Springer, New York, 1996.

[6] K. P. Burnham and D. R. Anderson, *Model Selection and Inference. A Practical Information-Theoretic Approach*, Springer, New York, 1998.

[7] A. X. Carvalho and G. Skoulakis, *Ergodicity and existence of moments for local mixtures of linear autoregressions*, Statistics & Probability Letters **71** (2005), no. 4, 313–322.

[8] A. X. Carvalho and M. A. Tanner, *Hypothesis testing in mixtures-of-experts of generalized linear time series*, Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering, Hong Kong, 2003.

[9] ______, *Mixtures-of-experts of autoregressive time series: asymptotic normality and model specification*, IEEE Transactions on Neural Networks **16** (2005), no. 1, 39–56.

[10] ______, *Modeling nonlinear time series with local mixtures of generalized linear models*, The Canadian Journal of Statistics **33** (2005), no. 1, 97–113.

[11] ______, *Modelling nonlinear count time series with local mixtures of poisson autoregressions*, 2005, Technical Report, Department of Statistics, Northwestern University.

[12] H. Chipman, E. George, and R. McCulloch, *Bayesian treed models*, Machine Learning **48** (2002), no. 1–3, 299–320.

[13] M. P. Clements and D. F. Hendry, *Forecasting Economic Time Series*, Cambridge University Press, Cambridge, 1998.

[14] I. Csiszár and P. C. Shields, *The consistency of the BIC Markov order estimator*, The Annals of Statistics **28** (2000), no. 6, 1601–1619.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society. Series B. Methodological **39** (1977), no. 1, 1–38.

[16] F. Diebold, T. Gunther, and A. Tay, *Evaluating density forecasts with applications to financial risk management*, International Economic Review **39** (1998), 863–882.

[17] M. Duflo, *Random Iterative Models*, Applications of Mathematics (New York), vol. 34, Springer, Berlin, 1997.

[18] Z. Ghahramani and G. Hinton, *The EM algorithm for mixtures of factor analyzers*, Tech. Rep., Department of Computer Science, University of Toronto, Toronto, 1996.

[19] E. Ghysels, R. McCulloch, and R. Tsay, *Bayesian inference for periodic regime-switching model*, Journal of Applied Econometrics **13** (1998), no. 2, 129–143.

[20] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley Professional, Massachusetts, 1989.

[21] P. J. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika **82** (1995), no. 4, 711–732.

[22] S. Gutta, J. Huang, P. Jonathon, and H. Wechsler, *Mixture of experts for classication of gender, ethnic origin, and pose of human faces*, IEEE Transactions on Neural Networks **11** (2000), no. 4, 948–960.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Series in Statistics, Springer, New York, 2001.

[24] D. F. Hendry, *Dynamic Econometrics*, Advanced Texts in Econometrics, The Clarendon Press Oxford University Press, New York, 1995.

[25] G. Huerta, W. Jiang, and M. A. Tanner, *Mixtures of time series models*, Journal of Computational and Graphical Statistics **10** (2001), no. 1, 82–89.

[26] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, *Adaptative mixtures of local experts*, Neural Computation **3** (1991), 79–87.

[27] N. Jeffries and R. Pfeiffer, *A mixture model for the probability distribution of rain rate*, Environmetrics **12** (2001), no. 1, 1–10.

[28] W. Jiang and M. A. Tanner, *Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation*, The Annals of Statistics **27** (1999), no. 3, 987–1011.

[29] ———, *On the approximation rate of hierarchical mixtures-of-experts for generalized linear models*, Neural Computation **11** (1999), no. 5, 1183–1198.

[30] ———, *On the identiability of mixtures-of-experts*, Neural Networks **12** (1999), no. 9, 1253–1258.

[31] ———, *On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models*, IEEE Transactions on Information Theory **46** (2000), no. 3, 1005–1013.

[32] H. Joe, *Time series models with univariate margins in the convolution-closed infinitely divisible class*, Journal of Applied Probability **33** (1996), no. 3, 664–677.

[33] M. Jordan and R. Jacobs, *Hierarchical mixtures-of-experts and the EM algorithm*, Neural Computation **6** (1994), 181–214.

[34] R. Jung, M. Kukuk, and R. Liesenfeld, *Time series of count data: modelling and estimation*, Economics Working Paper, Department of Economics, Christian-Albrechts-Universitat, Kiel, 2005.

[35] R. Kurnik, J. Oliver, S. Waterhouse, T. Dunn, Y. Jayalakshmi, M Lesho, M Lopatina, J Tamada, C Wei, and R. O. Potts, *Application of mixture of experts algorithm for signal processing in a noninvasive glucose monitoring system*, Sensors and Actuators B: Chemical **60** (1999), no. 1, 19–26.

[36] W. Li, *Time series models based on generalized linear models: some futher results*, Biometrics **50** (1994), 506–511.

[37] H. Lin, C. McCulloch, B. Turnbull, E. Slate, and L. Clark, *A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations*, Statistics in Medicine **19** (2000), no. 10, 1303–1318.

[38] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Monographs on Statistics and Applied Probability, vol. 37, Chapman & Hall/CRC, London, 1998.

[39] R. McCulloch and R. Tsay, *Bayesian analysis of threshold autoregressive processes with a random number of regimes*, Computing Science and Statistics **25** (1994), 253–262.

[40] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Communications and Control Engineering Series, Springer, London, 1993.

[41] M. Mitchell, *An Introduction to Genetic Algorithms (Complex Adaptive Systems)*, MIT Press, Massachusetts, 1998.

[42] F. Peng, R. Jacobs, and M. A. Tanner, *Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition*, Journal of the American Statistical Association **91** (1996), no. 435, 953–960.

[43] B. G. Quinn, G. J. McLachlan, and N. L. Hjort, *A note on the Aitkin-Rubin approach to hypothesis testing in mixture models*, Journal of the Royal Statistical Society. Series B. Methodological **49** (1987), no. 3, 311–314.

[44] M. Rosenblatt, *Remarks on a multivariate transformation*, Annals of Mathematical Statistics **23** (1952), 470–472.

[45] G. Schwarz, *Estimating the dimension of a model*, The Annals of Statistics **6** (1978), no. 2, 461–464.

[46] M. B. L. A. Siek and D. P. Solomatine, *Optimizing mixtures of local experts in tree-like regression models*, Proceedings of IASTED Conference on Artificial Intelligence and Applications, Innsbruck, 2005.

[47] M. A. Tanner, *Tools for Statistical Inference*, Springer Series in Statistics, Springer, New York, 1996.

[48] M. Tipping and C. Bishop, *Mixtures of probabilistic principal component analyzers*, Neural Computation **11** (1999), no. 2, 443–482.

[49] H. Tong, *Threshold Models in Nonlinear Time Series Analysis*, Lecture Notes in Statistics, vol. 21, Springer, New York, 1983.

[50] P. J. van Laarhoven and E. H. Aarts, *Simulated Annealing: Theory and Applications*, Mathematics and Its Applications, D. Reidel, Dordrecht, 1987.

[51] P. Wang and M. L. Puterman, *Mixed logistic regression models*, Journal of Agricultural, Biological, and Environmental Statistics **3** (1998), no. 2, 175–200.

[52] A. Weigend, M. Mangeas, and A. Srivastava, *Nonlinear gated experts for time series: discovering regimes and avoid overfitting*, International Journal of Neural Systems **6** (1995), no. 4, 373–399.

[53] W. H. Wong, *Theory of partial likelihood*, The Annals of Statistics **14** (1986), no. 1, 88–123.

[54] C. S. Wong and W. K. Li, *On a logistic mixture autoregressive model*, Biometrika **88** (2001), no. 3, 833–846.

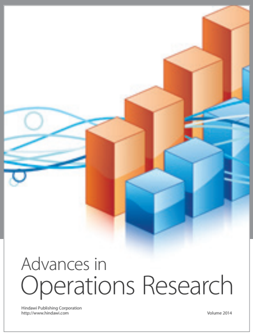
[55] S. A. Wood, W. Jiang, and M. A. Tanner, *Bayesian mixture of splines for spatially adaptive nonparametric regression*, Biometrika **89** (2002), no. 3, 513–528.

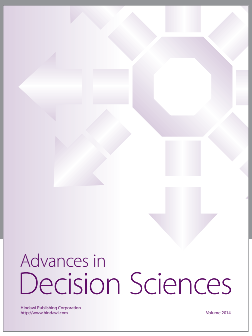
[56] A. Zeevi and R. Meir, *Density estimation through convex combinations of densities: approximation and estimation bounds*, Neural Networks **10** (1996), no. 1, 99–109.

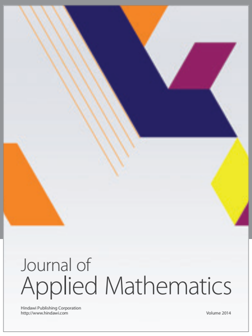
[57] M. Zeevi, R. Meir, and R. Adler, *Nonlinear models for time series using mixtures of autorregressive models*, 1999, Unpublished Technical Report, http://www.isl.stanford.edu/ azeevi/.

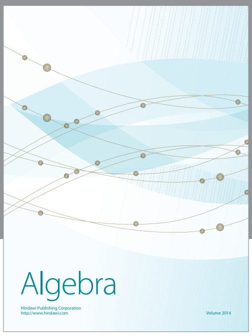
Alexandre X.Carvalho: SBS, Quadra 1, Bloco J, Edifício BNDES, Sala 718, Brasília,
DF, CEP 70076-900, Brazil
E-mail address: alexandre.ywata@ipea.gov.br

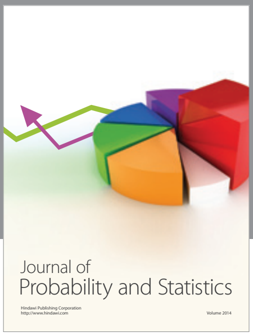
Martin A.Tanner: Department of Statistics, Weinberg College of Arts and Sciences,
Northwestern University, Evanston, IL 60208, USA
E-mail address: mat132@northwestern.edu


Advances in
Operations Research

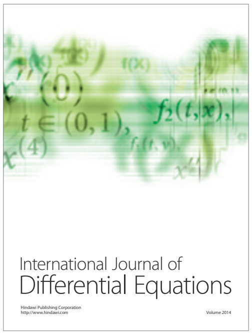

Advances in
Decision Sciences

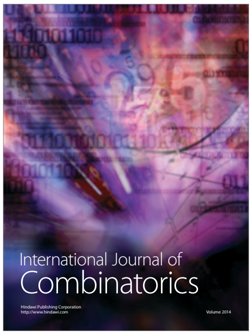

Journal of
Applied Mathematics

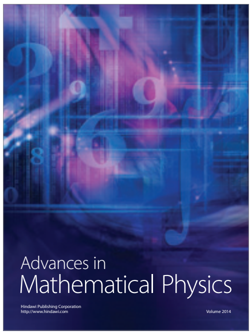

Algebra

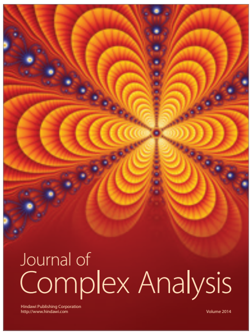

Journal of
Probability and Statistics

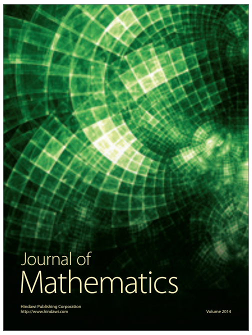
The Scientific
World Journal

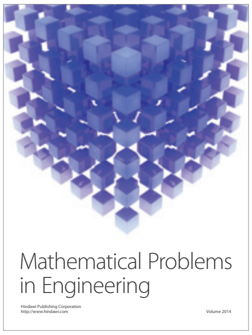

International Journal of
Differential Equations

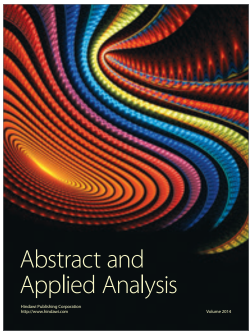

International Journal of
Combinatorics

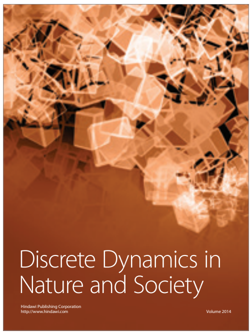

Advances in
Mathematical Physics

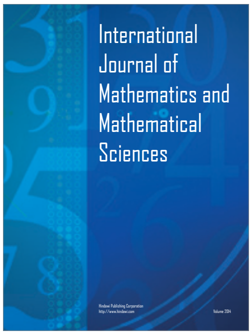

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

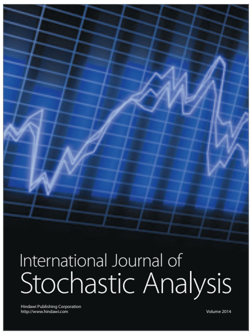

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization