Statistica Sinica **13**(2003), 1097-1118

TIME SERIES MODELING VIA HIERARCHICAL MIXTURES

Gabriel Huerta[1], Wenxin Jiang[2] and Martin A. Tanner[2]

[1] *University of New Mexico and* [2] *Northwestern University*

Abstract: We address the problem of model comparison and model mixing in time series using the approach known as *Hierarchical Mixtures-of-Experts*. Our methodology allows for comparisons of arbitrary models, not restricted to a particular class or parametric form. Additionally, the approach is flexible enough to incorporate exogenous information that can be summarized in terms of covariables or simply time, through weighting functions that define the hierarchical mixture. Huerta, Jiang and Tanner (2001) showed how to estimate the parameters of such models using the EM-algorithm. Here we present some theoretical properties of the method in the context of time series modeling. In addition, we consider model estimation using a full Bayesian approach based on Markov Chain Monte Carlo simulation. Methods for model checking and diagnostics for this class of models are presented. Finally, we explore our methodology by analyzing an economic-financial series: the monthly *US industrial production index* from 1947 to 1993.

Key words and phrases: Covariables, EM-algorithm, hierarchical mixture, Markov Chain Monte Carlo, time series.

1. Introduction

Recent advances in computational statistics offer a wide variety of tools for parametric and non-parametric modeling, particularly in mixture modeling and model comparison. Time series methods are no exception, and mixing is crucial to improve forecasting or to detect changes in structure across time. Also, mixture models in time series offer the possibility of approximating non-linearities with the advantage that mixtures of *simple*, perhaps linear components, are usually more tractable than more *parsimonious* non-linear processes.

The problem of model mixing and model comparison in time series has a long tradition. From a Bayesian perspective, the weights of the mixture are defined through the marginal posterior probability of each individual model. Based on these posterior probabilities, comparisons and marginal inference are feasible. Some examples in this area relate to the work by McCulloch and Tsay (1994) for Difference Stationary-Trend Stationary modeling (see Section 3). Assuming a class of linear autoregressive models, Troughton and Godsill (1997) propose a MCMC reversible jump algorithm to deal with uncertainty about model order or *autoregressive lag*, which consequently allows comparisons of different autoregressions up to an arbitrary order. In this direction, but using a stochastic variable search approach, Huerta and West (1999) incorporate model order uncertainty but put emphasis on inference for latent component structure. Comparisons of multiple autoregressive models with very high orders are feasible with this method. In terms of Bayesian Dynamic Linear Models (DLM), Harrison and Stevens (1976) introduced an approach known as Multi-Process models. Since DLMs are sequential in nature, Multi-Processes allow for model mixtures based on posterior probabilities and comparison of different dynamic models at each time given the past information. The class of DLMs is quite flexible since most of the known time series models can be expressed as an element of this class. On the other hand, particular cases like GARCH or EGARCH models lead to non-normal DLMs which may involve challenging computational issues. Further explanations of these issues may be found in West and Harrison (1997, Chap. 12 and Chap. 15).

The time series modeling approach that we adopt for this paper is based on the idea of mixing models through the neural network architecture known as Hierarchical Mixtures-of-Experts (HME). HME was first introduced in Jordan and Jacobs (1994). The HME approach easily allows for model comparison and permits one to represent the mixture weights as a function of time or other covariables. With the additional hierarchy, it is possible to localize the comparisons to specific *regions* or *regimes*. Furthermore, the defining elements of the mixture do not have to be restricted to a particular class of models, permitting very general comparisons. First, we review how to estimate the model parameters via maximum likelihood using the EM-algorithm. We then consider a full Bayesian approach to explore the posterior distribution of the parameters based on a Markov Chain Monte Carlo (MCMC) scheme that follows the lines of Peng, Jacobs and Tanner (1996).

In this paper, we introduce the general framework of hierarchical mixtures for time series, review inference via the EM-algorithm and consider some theoretical properties of this approach in the context of time series modeling. We present a MCMC method to implement a full Bayesian solution of the time series hierarchical mixture and discuss other approaches to averaging time series models. We apply the methodology to a financial-economic time series. We consider approximately 45 years of the US industrial production index and discriminate between stochastic-trend models and deterministic-trend models.

2. Hierarchical Mixtures of Time Series: Models and Theory

2.1. A general framework for time series modeling via HME

Let $\{y_t\}_0^n$ be a time series of endogenous or response variables, and $\{\mathbf{x}_t\}_0^n$ be a time series of exogenous variables or covariates. Suppose the main interest is to draw inference on $\{y_t\}_0^n$ conditional on $\{\mathbf{x}_t\}_0^n$. Let the conditional probability density function (pdf) of y_t be $f_t(y_t|\mathcal{F}_{t-1},\mathcal{X};\theta)$, where θ is a parameter vector; $\mathcal{X}$ is the σ-field generated by $\{\mathbf{x}_t\}_0^n$, representing the external information; and for each t, $\mathcal{F}_{t-1}$ is the σ-field generated by $\{y_s\}_0^{t-1}$ representing "the previous history" at time $t-1$. Typically, the conditional pdf f_t is assumed to depend on $\mathcal{X}$ through $\mathbf{x}_t$ only. In Mixtures-of-Experts (ME) methodology (Jacobs, Nolan and Hinton (1991)) and Hierarchical Mixtures-of-Experts (HME) (Jordan and Jacobs (1994)), the pdf f_t of the response variable is assumed to be a conditional mixture of the pdfs from simpler, well established models. In a time series context, this mixture can be represented by the finite sum

$$f_t(y_t|\mathcal{F}_{t-1},\mathcal{X};\theta)=\sum_J \mathrm{g}_t(J|\mathcal{F}_{t-1},\mathcal{X};\gamma)\pi_t(y_t|\mathcal{F}_{t-1},\mathcal{X},J;\eta), \tag{1}$$

where the functions $\mathrm{g}_t(\cdot|\cdot,\cdot;\gamma)$ are the mixture weights; $\pi_t(\cdot|\cdot,\cdot,J;\eta)$ are the conditional pdfs from simpler models each defined by a label J; and γ and η are vectors of sub-parameters from θ.

The simpler models in HME are often referred to as the "experts". In a time series context, one "expert" could be an AR(1) model, another "expert" could be a GARCH(1,1) model or an EGARCH(1,1) model. For example, in a situation where it is not clear whether to use a stochastic or a deterministic trend, one expert could be a *trend-stationary process*, another a *difference-stationary process*. A somewhat simpler situation occurs when all the experts propose a model of the same type, e.g., linear autoregressive, but perhaps with different values for the coefficients or for the model order.

Furthermore, the HME models we use in this paper have an additional layer designed with the purpose of local time series modeling. The HME partitions the covariate space, which could include time, into O overlapping regions called "overlays". In each overlay, M models are to compete with each other, in the hope that the model most suitable to the specific region is favored by a high weight. By having multiple overlays, the hierarchical mixture model allows for modeling multiple switching across regions.

Therefore, the expert index J can be expressed as $J=(o,m)$, where the overlay index o takes a value from $\{1,\ldots,O\}$ and the model-type index m from $\{1,\ldots,M\}$, so that the mixture can now be represented by

$$f_t(y_t|\mathcal{F}_{t-1},\mathcal{X};\theta)=\sum_{o=1}^{O}\sum_{m=1}^{M}\mathrm{g}_t(o,m|\mathcal{F}_{t-1},\mathcal{X};\gamma)\pi_t(y_t|\mathcal{F}_{t-1},\mathcal{X},o,m;\eta). \tag{2}$$

We allow the same type of model m to assume different versions or more specifically different parameter values, at each possible overlay. It is worth noting that this framework defines the two-layer HME architecture of Jordan and Jacobs

(1994), where the first layer of gating functions hypothesizes O overlays on the entire time axis, and the second layer of gating functions defines weights for each of the M model types within each overlay. When the input space for the gating functions is time, the hierarchical mixture model can identify the region over which a model or a set of models is (are) dominant in a data-adaptive manner. Thus, the present approach allows for modeling multiple regime switching. The mixing weights are often referred to as “gating functions”. They can depend on the previous history (Zeevi, Meir and Adler (1988)), exogenous information, or can exclusively depend on t. Typically the gating functions have the form

$$\mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma) = \left\{ \frac{e^{v_o + \mathbf{u}_o^T \mathbf{w}_t}}{\sum_{s=1}^{O} e^{v_s + \mathbf{u}_s^T \mathbf{w}_t}} \right\} \left\{ \frac{e^{v_{m|o} + \mathbf{u}_{m|o}^T \mathbf{w}_t}}{\sum_{l=1}^{M} e^{v_{l|o} + \mathbf{u}_{l|o}^T \mathbf{w}_t}} \right\}, \tag{3}$$

where the v’s and $\mathbf{u}$’s are parameter components of γ, and $\mathbf{w}_t$ is an “input” at time t which is measurable with respect to the σ-field induced by $\mathcal{F}_{t-1} \cup \mathcal{X}$. For example, the input $\mathbf{w}_t$ could be the covariate $\mathbf{x}_t$, the “two-lag” history $(y_{t-1}, y_{t-2})^T$, or exclusively depend on time t.

In the context where one is interested in how the weighting for individual models is assigned across different time periods, $\mathbf{w}_t$ can be taken as (t/n). The use of (t/n) instead of simply t is suggested to avoid computational overflow errors when implementating the EM algorithm or a MCMC method with these models. Therefore, one can adopt the following parametric form for the gating functions:

$$\mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma) = g_{om}(t; \gamma) \equiv \left\{ \frac{e^{v_o + u_o(t/n)}}{\sum_{s=1}^{O} e^{v_s + u_s(t/n)}} \right\} \left\{ \frac{e^{v_{m|o} + u_{m|o}(t/n)}}{\sum_{l=1}^{M} e^{v_{l|o} + u_{l|o}(t/n)}} \right\}. \tag{4}$$

Here γ includes the following components: $v_1, u_1, \ldots, v_{O-1}, u_{O-1}, v_{1|1}, u_{1|1}, \ldots, v_{M-1|1}, u_{M-1|1}, \ldots, v_{M-1|O}, u_{M-1|O}$. For identifiability, we set $v_O = u_O = v_{M|o} = u_{M|o} = 0$ for all $o = 1, \ldots, O$. This restriction ensures that the gating functions are uniquely identified by the γ parameter.

More generally, write the right hand side of (3) as $g_{om}(\mathbf{w}_t; \gamma)$, which is a function defined on $\mathbf{W}$, the support of $\mathbf{w}_t$. Suppose we impose the constraint $v_O = \mathbf{u}_O = v_{M|o} = \mathbf{u}_{M|o} = \mathbf{0}$ for all $o = 1, \ldots, O$, and denote by γ the collection of all other parameters v_o’s and $\mathbf{u}_{m|o}$’s. Denote g as the vector of all g_{om} components. Then we have the following.

Proposition 1. *Suppose the support* $\mathbf{W}$ *of the covariate* $\mathbf{w}_t$ *contains* $p+1$ *‘design points’* $\mathbf{w}^0, \ldots \mathbf{w}^p$ *such that the design matrix* $D = [(1, \mathbf{w}^0)^T, \ldots, (1, \mathbf{w}^p)^T]$ *is nonsingular. Then,* $g(\mathbf{w}; \gamma) = g(\mathbf{w}; \gamma')$ *for all* $\mathbf{w} \in \mathbf{W}$*, if and only if* $\gamma = \gamma'$*.*

Proof. The 'if' is obvious. To prove the 'only if', note that $g(\mathbf{w};\gamma) = g(\mathbf{w};\gamma')$ implies that

$$\frac{\sum_{m=1}^{M} g_{om}(\mathbf{w},\gamma)}{\sum_{m=1}^{M} g_{Om}(\mathbf{w},\gamma)} = \frac{\sum_{m=1}^{M} g_{om}(\mathbf{w},\gamma')}{\sum_{m=1}^{M} g_{Om}(\mathbf{w},\gamma')},$$

which implies that $v_o + \mathbf{u}_o^T\mathbf{w} = v_o' + \mathbf{u}_o'^T\mathbf{w}$ for all $\mathbf{w}$ in $\mathbf{W}$. Choose the $p+1$ support points making D nonsingular and solve the equation, we obtain $(v_o, \mathbf{u}_o^T) = (v_o', \mathbf{u}_o'^T)$, which holds for all o. Similarly, from $g(\mathbf{w};\gamma) = g(\mathbf{w};\gamma')$ we are able to cancel the first factors from both sides, and conclude that $v_{m|o} + \mathbf{u}_{m|o}^T\mathbf{w} = v_{m|o}' + \mathbf{u}_{m|o}'^T\mathbf{w}$ for all $\mathbf{w}$ in $\mathbf{W}$. Based on the non-singularity of D, we obtain $(v_{m|o}, \mathbf{u}_{m|o}^T) = (v_{m|o}', \mathbf{u}_{m|o}'^T)$ for all o and m. Therefore $\gamma = \gamma'$.

Remark 1. This argument can be repeated to prove identifiability of mixing weights for HME with more than two layers.

Remark 2. Provided extra conditions on the "expert" densities which basically say that the experts are different, it is possible to establish the identifiability of the probability density functions up to some permutations of the expert labels, which can become completely identified with some order restriction. A detailed study in the context of non-hierarchical mixtures of generalized linear models is included in Jiang and Tanner (1999).

Remark 3. The condition on $\mathbf{W}$, in the case of a single covariate t, reduces to requiring that there exist two design points t_1 and t_2 in the support of t such that $t_1 \neq t_2$. This condition typically holds for any time series of length $n \geq 2$.

As an example, suppose that one believes that the entire time period under observation may be modeled with three overlays decomposed into three regions $O1$, $O2$ and $O3$ with unknown locations, and in each of these regions one of the two types of candidate models may be most appropriate: (M1) a random walk with a drift, $y_t = y_{t-1} + \alpha + \epsilon_t$, and (M2) a linear trend model with no intercept, $y_t = \beta(t/n) + e_t$. Here, $M = 2$ and $O = 3$. In the gating function (4), the first factor softly splits the entire time period into the three regions $O1$, $O2$ and $O3$ indexed by $o = 1, 2, 3$, and the second factor weights the model types $m = 1$ and 2 (named M1 and M2) in each region o (see Figure 1). Assuming independent standard normal innovation errors e_t and ϵ_t, the pdfs defined by the experts are $\pi_t|_{o,m=1} = \phi(y_t - y_{t-1} - \alpha_o)$ and $\pi_t|_{o,m=2} = \phi(y_t - \beta_o t)$ for $o = 1, 2, 3$, where $\phi(\cdot)$ denotes the pdf of a standard normal. Notice that the parameters or versions of each model-type can differ for distinct regions.

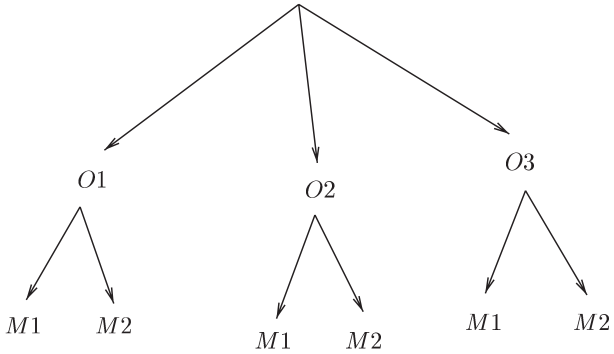


Figure 1. A graphical representation of a two-layer HME.

For describing such a situation, a single-layer HME (i.e., the Mixtures-of-Experts or ME (Jacobs et al. (1991))) can also be used in principle. Multiple versions of M1 and M2 can be combined by single-layer gating functions (here a single-layer gating function is a multinomial logit-linear function of time which may look similar to a single factor in (4)). The disadvantage of this ME approach compared to a two layer HME is that the competition between types of models within each region becomes confused.

Inference on the parameter θ can be based on the log-likelihood function, conditional on y_0, $\mathcal{X}$ and "averaged" in time, which is

$$\mathcal{L}_n(\cdot) = n^{-1} \sum_{t=1}^{n} \log f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \cdot). \tag{5}$$

We denote the maximum likelihood estimate (MLE) of θ as $\hat{\theta} = \arg\max \mathcal{L}_n(\cdot)$. For a Bayesian analysis, inferences on θ are based on the elicitation of a prior distribution $p(\theta)$ that leads, through Bayes theorem, to the posterior distribution $p(\theta|\mathcal{F}_n, \mathcal{X})$.

The free vector of parameters γ in the gating functions automatically determines the location and the "softness" of the splitting of the regions. The number of distinct model-types M usually is specified by the practitioner, depending on the number of models that are of interest to the specific problem. The number of regions O can sometimes be selected based on subjective considerations, especially if there is historical information which may be thought to influence the time series.

Given θ, there are two ways in evaluating the relative weighting of each of the M model types at time t. One is the *conditional* probability of each model m (with the current response y_t being conditioned on) defined by:

$$P_t(m|y_t, \mathcal{F}_{t-1}, \mathcal{X}, \theta) \equiv h_m(t) \equiv \sum_{o=1}^{O} h_{om}(t; \theta). \tag{6}$$

Another approach is to consider the *unconditional* probability / weight of model m at time t (unconditional on the current response y_t):

$$P_t(m|\mathcal{F}_{t-1}, \mathcal{X}, \theta) \equiv g_m(t) \equiv \sum_{o=1}^{O} g_{om}(t;\theta). \tag{7}$$

Point estimates of both probabilities can be obtained with the MLE $\hat{\theta}$ or by taking the expectation with respect to the posterior distribution $p(\theta|\mathcal{F}_n, \mathcal{X})$. As we show in Section 3, the point estimates of (6) can vary point-wise over time due to the conditioning on y_t. The point estimates of the second weighting scheme (7) are more smoother when describing a regional change of preference for model m, when the gating functions depend on a single covariate t. We compare each of these weighting approaches in the data analysis.

2.2. The EM algorithm

As noted in Huerta, Jiang and Tanner (2001), the EM algorithm provides a method for frequentist inference in this context. The EM algorithm starts with an initial estimate of the parameters θ^0. Then a sequence $\{\theta^i\}$ is obtained by iterating between the following two steps: For $i = 0, 1, 2, \ldots,$

E-step: Construct

$$Q^i(\theta) = \sum_{t=1}^{n}\sum_{o,m} h_{om}(t;\theta^i)\log\{\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m;\eta)\mathrm{g}_t(o,m|\mathcal{F}_{t-1}, \mathcal{X};\gamma)\}, \tag{8}$$

where $\theta = (\gamma, \eta)$, $\theta^i = (\gamma^i, \eta^i)$, $h_{om}(t;\theta^i) = h_{om}(t;\theta)|_{\theta=\theta^i}$, and

$$h_{om}(t;\theta) = \frac{g_{om}(t;\gamma)\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m;\eta)}{\sum_{s=1}^{O}\sum_{l=1}^{M} g_{sl}(t;\gamma)\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, s, l;\eta)} \tag{9}$$

is the "conditional probability" of choosing the expert (o, m) at time t and $g_{om}(t;\gamma)$ is the corresponding "unconditional probability".

M-step: Find $\theta^{i+1} = \arg\max_\theta Q^i(\theta)$.

In fact, $Q^i(\theta)$ is the posterior expectation with respect to $z_{om}(t)$ and conditional on θ^i of the *augmented log-likelihood*,

$$\mathcal{L}_A(\theta) = \sum_{t=1}^{n}\sum_{o,m} z_{om}(t)\log\{\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m;\eta)\mathrm{g}_t(o,m|\mathcal{F}_{t-1}, \mathcal{X};\gamma)\}, \tag{10}$$

where $z_{om}(t)$ are indicator variables that take the value 1 if expert (o, m) is chosen at time t and zero otherwise.

It is well-known (Tanner (1996)) that the limit of the sequence $\{\theta^i\}$, denoted by $\hat{\theta}(\theta^0)$, is a root of the likelihood equation $\nabla_\theta \mathcal{L}_n = 0$ corresponding to a stationary point. When the likelihood is multimodal, the limit depends on the different modes so we use multiple starting points and find the corresponding limits via the EM algorithm. We adopt as the point estimate the limit which results in the largest likelihood $\mathcal{L}_n$ over the multiple starting points.

A nice feature of the EM algorithm is that the objective function Q^i, in each step, has the form of a double sum of logarithms, instead of a "sum log sum" typical for the log likelihood function $\mathcal{L}_n$. For this reason, the maximization of the objective function can be decomposed into a number of smaller maximization problems which involve fewer parameters and usually define "known" maximizations of widely used models. For example, suppose the expert pdf has the form

$$\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta) = p_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, m; \eta_{om}), \tag{11}$$

where η is decomposed into a collection of sub-parameter η_{om}, each of which only appears in the pdf of one expert (see the example in Section 3). The parameter η_{om} carries an index o in addition to m to allow one type of model to take different versions (parameters) in different overlays. In such a situation, in the M step, the maximization over the η_{om}'s and γ can be performed separately. For example, for each o, m, i,

$$\eta_{om}^{i+1} = \arg\max_{\eta_{om}} \sum_{t=1}^{n} h_{om}(t; \theta^i) \log p_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, m; \eta_{om}), \tag{12}$$

which becomes the "standard" (albeit weighted by the h's) maximum likelihood problem for model type-m.

2.3. Some analytical considerations

Standard results on the properties of the MLEs, such as the ones described in Lehmann (1991, Chap.6), are established for the situation when data are independent and identically distributed. The observed data $\{y_t\}_0^n$ in this paper, however, are dependent and non-stationary. It is a very difficult problem to study the properties of MLEs for such nonstandard cases. However there is some relevant work, mostly based on various approaches of relaxing the independence assumption, under the frameworks of martingale dependence (Crowder (1976)), mixingale dependence, near epoch dependence (Gallant and White (1988), Chap.3−Chap.5), etc. Consistency and asymptotic normality of the MLEs can be proved under certain regularity conditions (see, e.g., Crowder (1976), Sarma (1986) and Weiss (1973)). However, these regularity conditions can be nonintuitive and difficult to check.

In the following proposition, we show that the ("unaveraged") score functions form a martingale and, under very mild conditions, the true parameter θ_0 is an approximate solution to the likelihood equation in the large sample size limit with a certain rate of precision. This, in a very general context, provides some justification of the maximum likelihood procedure which searches for an exact solution of the likelihood equation. Stronger results can be obtained if further regularity conditions are imposed, an exact solution of the likelihood equation can be shown consistent and asymptotically normal (see, e.g., Crowder (1976), Sarma (1986) and Weiss (1973)). It is noted, however, that these regularity conditions typically involve the properties of certain expectations which may be difficult to check, due to the nonlinearity of the time series.

Proposition 2. *Suppose $f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta)$ is differentiable with respect to θ, and ∇_θ and $\int dy_t$ commute when acting on f_t at $\theta = \theta_0$. Let $u_t(\theta) = \nabla_\theta \log f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta)$, the "score increment" at time t so the "averaged" score function is $\nabla_\theta \mathcal{L}_n(\theta) = n^{-1}\sum_{t=1}^n u_t(\theta)$. Let $||\cdot||$ be the Euclidean norm. We have:*

(i) *$\{u_t(\theta_0)\}$ is a sequence of martingale differences (so the unaveraged score functions form a martingale);*

(ii) *$P\{||\nabla_{\theta_0}\mathcal{L}_n(\theta_0)|| > \epsilon\} \le (n\epsilon)^{-2}\sum_{t=1}^n E||u_t(\theta_0)||^2$, for each positive ϵ;*

(iii) *$\nabla_{\theta_0}\mathcal{L}_n(\theta_0) = O_p(n^{-\delta/2})$ if $\sum_{t=1}^n E||\nabla_{\theta_0}\log f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta_0)||^2 = O(n^{2-\delta})$.*

Proof. For each $t = 1, 2, \ldots$,

$$\begin{aligned} & E\{u_t(\theta_0)|\mathcal{F}_{t-1}, \mathcal{X}\} \\ &= E\{\nabla_{\theta_0}\log f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta_0)|\mathcal{F}_{t-1}, \mathcal{X}\} \\ &= \int dy_t\{f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta_0) f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta_0)^{-1}\nabla_{\theta_0} f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta_0)\} = 0, \end{aligned}$$

by exchangeability of ∇_θ and $\int$, and $\int f_t = 1$. Hence (i).

(ii) is straightforward from Chebyshev's inequality, noting that $E\{u_s(\theta_0)^T u_t(\theta_0)\} = E||u_t(\theta_0)||^2\delta_{st}$ by using (i). Here δ_{st} is the Kronecker's delta.

(iii) follows from (ii) by taking $\epsilon = Mn^{-\delta/2}$ for a positive constant M which can be arbitrarily large.

Remark 4. This proposition shows that the true parameter satisfies the likelihood equation approximately in the large-n limit, if the exponent δ in (iii) is positive. This is a very mild requirement – it is satisfied as long as the growth of $E||\nabla_{\theta_0}\log f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta_0)||^2$ in time is slower than t, which basically disallows the variances of the "score increments" to blow up as fast as t.

2.4. A Bayesian approach

A standard way to deal with parameter uncertainties in a mixture model context is through a Bayesian approach with MCMC methods. To explore the posterior distributions of an HME time series model, we propose an MCMC scheme with a format similar to the HME approach used by Peng, Jacobs and Tanner (1996) for speech recognition tasks. A key benefit of MCMC methods is the ability to obtain samples from the posterior distribution of any functional form of the parameters of the model. In this way, more information is obtained than can be provided by just a point estimate as given by the EM-algorithm.

First, we assume that the prior distribution for $\theta = (\eta, \gamma)$ has the form

$$p(\theta) = p(\eta)p(\gamma). \tag{13}$$

So the "expert" parameters η and the "gating" parameters γ are a-priori independent.

We define $\mathbf{Z} = \{\mathbf{Z}(t); t = 1, \ldots, n\}$ and, for each t, $\mathbf{Z}(t) = \{z_{om}(t); o = 1, 2, \ldots, O, m = 1, 2, \ldots, M\}$ is the set of indicator variables at time t. The sets $\mathbf{Z}(t), t = 1, \ldots, n$ are mutually independent and, given θ, $p(\mathbf{Z}(t)|\theta, \mathcal{X})$ has a multinomial distribution with total count 1 and cell probabilities $g_{om}(t; \gamma)$.

Our MCMC method is based on the fact that is simpler to obtain samples of $(\theta, \mathbf{Z})$ from the augmented posterior distribution $p(\theta, \mathbf{Z}|\mathcal{F}_n, \mathcal{X})$ than samples from the posterior $p(\theta|\mathcal{F}_n, \mathcal{X})$. This is in accordance with the principles for calculation of posterior distributions by data augmentation introduced by Tanner and Wong (1987). The general structure of the conditional posterior distributions required for our MCMC-data augmentation algorithm are briefly outlined here. Specifically,

- The conditional posterior distribution for $p(\mathbf{Z}|\theta, \mathcal{F}_n, \mathcal{X})$ is sampled via the marginal conditional posterior distributions $p(\mathbf{Z}(t)|\theta, \mathcal{F}_n, \mathcal{X})$ for each value of t. Given θ, $\mathcal{F}_n$ and $\mathcal{X}$, the vector $\mathbf{Z}(t)$ has a multinomial distribution with total count 1 and for which

$$Pr[z_{om}(t) = 1|\theta, \mathcal{F}_n, \mathcal{X}] = h_{om}(t; \theta), \tag{14}$$

  where $h_{om}(t; \theta)$ is defined at (9). Certainly, $Pr[z_{om}(t) = 1|\theta, \mathcal{F}_n, \mathcal{X}] = Pr[z_{om}(t) = 1|\theta, \mathcal{F}_t, \mathcal{X}]$ since all the information of $\{y_t\}_0^n$ for $z_{om}(t)$ is summarized in $\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta_{om})$.

- The parameter $\theta = (\eta, \gamma)$ is sampled in two stages. η is sampled from the conditional posterior distribution $p(\eta|\gamma, \mathbf{Z}, \mathcal{F}_n, \mathcal{X})$ and γ is sampled from the conditional posterior distribution $p(\gamma|\eta, \mathbf{Z}, \mathcal{F}_n, \mathcal{X})$. By Bayes theorem,

$$p(\eta|\gamma, \mathbf{Z}, \mathcal{F}_n, \mathcal{X}) \propto \prod_{t=1}^{n} \prod_{o=1}^{O} \prod_{m=1}^{M} \pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta)^{z_{om}(t)} p(\eta). \tag{15}$$

Similarly,

$$p(\gamma|\eta, \mathbf{Z}, \mathcal{F}_n, \mathcal{X}) \propto \prod_{t=1}^{n} \prod_{o=1}^{O} \prod_{m=1}^{M} \mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma)^{z_{om}(t)} p(\gamma). \qquad (16)$$

If η is decomposed into a collection of sub-parameters η_{om} and these are assumed a-priori independent, each η_{om} may be sampled individually. If η_{om} has a prior that is conjugate with respect to the pdf of model "m" and expert "o", its conditional posterior has a closed-form and is usually easy to sample. For γ, we use Metropolis-Hastings steps to obtain samples from its full conditional distribution. Specific issues of model implementation and prior specifications depend on the particular application, as illustrated in Section 3.

In contrast to the EM algorithm, the full Bayesian approach allows complete assessment of uncertainties in estimating the parameters and the conditional and unconditional model probabilities. However, the EM-based approach is computationally less intensive and could be used as a first step to the MCMC implementation.

2.5. Comparison with other approaches

We now compare our HME approach with other methods for model selection/averaging, among a candidate set of models labeled by $m = 1, \ldots, M$.

Our method provides model selection probabilities which can change *gradually* over time, over the multiple overlapping sub-regions. The smooth parametric g's provide the "unconditional" weights for the models of interest, in addition to the less smooth h's. The gating function g's may be regarded as a "prior" probability of model m that depends on some parameter γ to be estimated from the data. The g's are also dependent on time t and are parameterized in the multinomial logit form, which can be suitable for describing regional changes for the model selection probability. In some other model selection methods, however, *only* a point-wise (rather than a regional) description of the model selection probability is provided. One such example is hidden Markov modeling (see the papers by Hamilton (1989), and a Bayesian approach by McCulloch and Tsay (1994)), where the choice over $M = 2$ models is formulated as a two-state Markov chain and the resulting model probabilities fluctuate pointwise in time. Filardo (1994) extends this Markov switching modeling to allow time-varying transition probabilities that depend on exogenous variables. He suggests a logistic functional form for the transition probabilities and uses direct maximum likelihood to estimate model parameters. Still, the methodology seems only capable of hosting problems in which the means, variances and coefficients of an autoregressive process evolve in time in a dichotomous manner.

McCulloch and Tsay (1993, Section 4) consider random mean shift and random variance shift models. They suggest a probit structure (see their equation 13) to relate the probability of a shift to possible exogenous variables. In their approach they consider a baseline model and at any given time point, there can be a shift from that model. In our context, we consider M specific competing models, where $M \geq 2$.

It is noted that ME, HME and the related hidden Markov decision trees have had wide applications in time series modeling − see for example Jordan, Ghahramani and Saul (1996) and Zeevi, Meir and Adler (1998). The focus there, however, is mainly on the flexibility of the methodology (e.g., improvement in fit), rather than on model selection. Also, in that literature, experts typically have the same model type, albeit in different "versions" (i.e., with different parameters). The paper by Weigand, Mangeas and Srivastava (1995) uses gated experts (GE) to discover regime switching. In GE, the experts combined are standard neural networks with a linear output unit and tanh hidden units. This provides a very flexible modeling scheme since it switches among unknown submodels, due to the approximation property of the neural network experts. However, in the situation of selecting among M candidate models, *types of which are specified by the researcher*, GE is less efficient. Our HME approach uses the M specific models, and does not involve approximation, while the GE uses M neuralnets, each with P tanh nodes to approximate the exact models. The GE will have to employ many nodes and can only achieve an approximation to the exact model. Thus GE requires more parameters, and has a less clear interpretation for the purpose of model comparison.

2.6. Model checking

For issues on model checking and diagnostics, we propose the use of one-step-ahead predictive distribution functions as in Kim, Shephard and Chib (1998) and Elerian, Chib and Shephard (2001). For this, let

$$F_t(y_t|\mathcal{F}_{t-1},\mathcal{X},\theta) = Pr(Y_t \leq y_t|\mathcal{F}_{t-1},\mathcal{X},\theta). \tag{17}$$

Given the definition of a ME or HME,

$$F_t(y_t|\mathcal{F}_{t-1},\mathcal{X},\theta) = \sum_J \mathrm{g}_t(J|\mathcal{F}_{t-1},\mathcal{X};\gamma)F_{\pi_t}(y_t|\mathcal{F}_{t-1},\mathcal{X},J;\eta), \tag{18}$$

where $F_{\pi_t}(\cdot|\cdot,\cdot,J;\eta)$ is the distribution function of the simpler models each indexed by a label J. We note that, given the model parameters, the calculation of the overall distribution function is direct from the evaluation of the distribution function of the individual models.

It can be shown that if the model is correctly specified, u_t is uniformly distributed on the interval $(0,1)$ and defines an independent sequence (Rosenblatt(1952)). We can estimate u_t by $\hat{u}_t = F_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, \hat{\theta})$ where $\hat{\theta}$ can be the MLE estimate obtained with the EM algorithm or a posterior summary based on a MCMC scheme. In consequence, we can judge model adequacy by the serial correlation and the distributional shape of $\{\hat{u}_t\}$, or with transformed values via an inverse cdf, for example, the $N(0,1)$ cdf. Additionally, Kim, Shephard and Chib (1998) suggest focusing on $2|\hat{u}_t - 0.5|$ to explore the correlation and distributional form under a correct model.

Additionally, we can also compute the standardized forecast errors as

$$\frac{y_t - E(y_t|\mathcal{F}_{t-1}, \mathcal{X}, \hat{\theta})}{\sqrt{Var(y_t|\mathcal{F}_{t-1}, \mathcal{X}, \hat{\theta})}}; \quad t = 1, \ldots, n, \tag{19}$$

where $E(y_t|\mathcal{F}_{t-1}, \mathcal{X}, \hat{\theta})$ and $Var(y_t|\mathcal{F}_{t-1}, \mathcal{X}, \hat{\theta})$ are the conditional expectation and conditional variance of y_t given the history up to time $t-1$, the exogenous variables and a parameter estimate $\hat{\theta}$. The numerator of the standardized forecast errors is very similar to the generalized residual approach of Lai and Wong (2001), Section 4.3. However, the models of Lai and Wong (2001) are different from HME. Corresponding to their equation 9a, the numerator of equation 19 involves a combination of error terms from several different models with possible different variances. Distributional aspects of the standardized errors are harder to determine than for u_t, but they may provide a graphical aid to indentifying model misspecifications.

3. Application

3.1. Difference stationary-trend stationary modeling

We now consider the United States industrial production index as reported by the Federal Reserve Statistical Release G.17 (see McCulloch and Tsay (1994)). The data were obtained on a monthly basis from January 1947 to December 1993 and are seasonally adjusted. In total, the time series includes $n = 546$ observations that appear in Figure 2. From this picture, we can see that the production index exhibits a trend towards higher values with the progression of time. An important matter for economical policy is to determine if the trend has a stochastic or a deterministic nature. That is, which of the following is a more appropriate model:

$$y_t = \phi_0 + y_{t-1} + \phi_{1,1}(y_{t-1} - y_{t-2}) + \ldots + \phi_{1,u}(y_{t-u} - y_{t-u-1}) + \epsilon_{1,t},$$

$$y_t = \beta_0 + \beta_1 t/n + \phi_{2,1} y_{t-1} + \ldots + \phi_{2,v} y_{t-v} + \epsilon_{2,t},$$

where β_0, β_1, ϕ_0, $\phi_{1,i}$ and $\phi_{2,j}$ are constants, $\epsilon_{i,t}$ is the innovation error with variance σ_i^2; $i = 1, 2$. Note that u and v are non-negative integers and B is the lag operator of order 1, $By_t = y_{t-1}$.

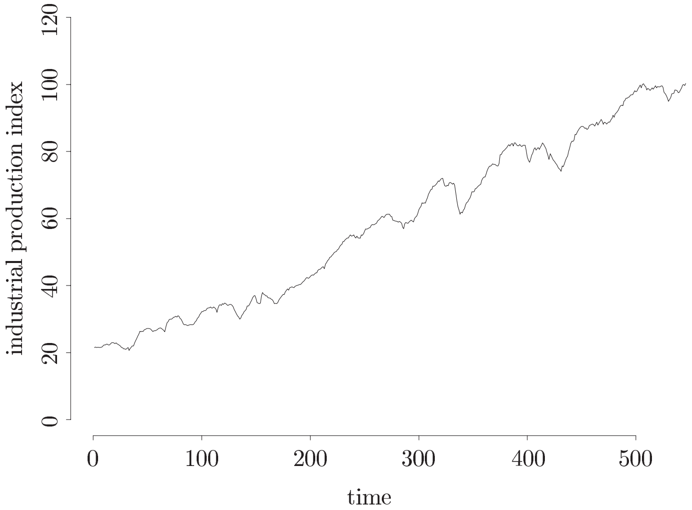
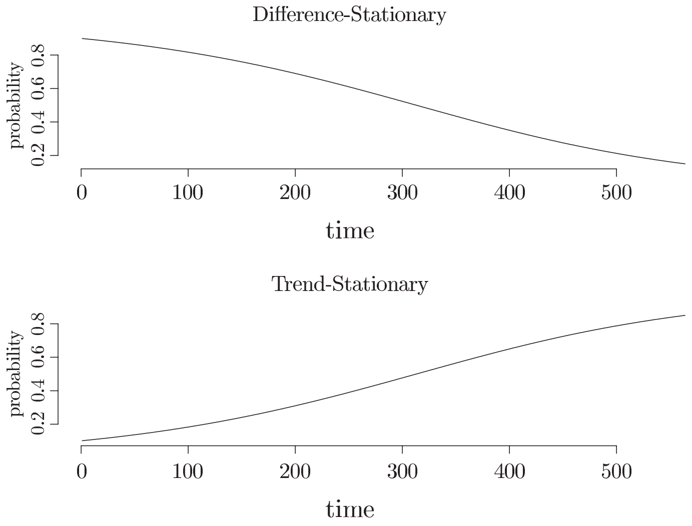


Figure 2. *US Industrial Production Index.* Monthly-seasonally adjusted observations from January 1947 to December 1993.

For these data, it has been suggested that autoregressions of order higher than two are unnecessary for adequate modeling, so we fix $u = v = 2$. For analysis using a HME we define $\pi_t|_{o,m=1}$ to be the pdf determined by the difference-stationary model and $\pi_t|_{o,m=2}$ to be the pdf defined by the trend-stationary model, so $M = 2$. Our first analysis uses $O = 2$, i.e., two overlays for each model-type combination. The EM was implemented by using 20 starting points, with parameters ϕ_0, $\phi_{1,i}$,$\phi_{2,j}$ and σ_1^2 fixed at the MLE based on fitting only a difference-stationary model. Similarly, β_0, β_1, $\phi_{2,i}$ and σ_2^2 were initialized at the MLE via a trend-stationary assumption. The initial values for the parameters in the mixing weights or gating functions, v_1, u_1, $v_{1|1}, u_{1|1}$ and $v_{1|2}, u_{1|2}$ were generated randomly from a uniform distribution on $[-a, a]$, with a large. In each case, the EM was run for 500 iterations and traces of individual parameters indicate that the EM finds a local mode at about 500 iterations. The 20 solutions were ranked by evaluating the log-likelihood of the HME, $\mathcal{L}_n(\cdot)$, and the solution that produced the maximum likelihood estimates (MLE) was used to obtain Figures 3-4.

Figure 3 presents our ML estimate of $g_m(t)$ for $m = 1, 2$ as a function of time, where 1 represents difference-stationary and 2 represents trend-stationary.

For the first part of the series, up to about 200 observations, the HME weights towards a difference-stationary model. For more recent information, it favors a trend-stationary process.


Difference-Stationary

Trend-Stationary

Figure 3. Maximum likelihood estimates of $g_m(t)$ for both model-types considered: Difference-stationary and Trend-stationary.

Figure 4 presents our ML estimate for $h_m(t)$, $m = 1, 2$, which for large samples sizes approximately gives the posterior probability for each model given all the information available up to time t (see Section 2). Due to this asymptotic behavior, comparisons of this picture to Figure 2 of McCulloch and Tsay (1994) (MT) are valid. Both HME and MT split the data in two parts, the first favors difference-stationarity, but the second HME favors trend-stationarity while for MT it is harder to distinguish the two models. On the other hand, HME seems to reflect a *transition* period for the second part of the information. In other words, as time passes, the data is moving toward a trend-stationary behavior and, particularly by the end of the series, there are many points for which the probability of trend-stationary is higher than 0.8. It is noted that Figure 3 offers a smooth version of Figure 4 which is less affected by individual, perhaps contrasting observations. At least for large sample sizes, this smooth behavior of the estimate of $g_m(t)$ relative to the estimate of $h_m(t)$ will be observed, since $g_m(t)$ is approximately the expected value of $h_m(t)$ with respect to the data.

For the full Bayesian analysis of the HME using the MCMC method of Section 2.4, we fixed $O = 2$ and $u = v = 2$. Define $\phi_1^o = (\phi_0^o, \phi_{1,1}^o, \phi_{1,2}^o)$ to be the vector of parameters for the difference-stationary model and $\phi_2^o = (\beta_0^o, \beta_1^o, \phi_{2,1}^o, \phi_{2,2}^o)$ to be the vector of parameters for the trend-stationary model, $o = 1, 2$.

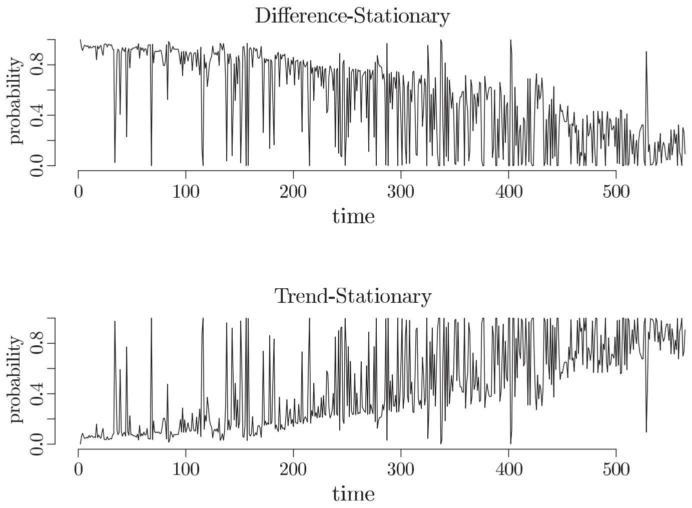


Figure 4. Maximum likelihood estimates of $h_m(t)$ for both model-types considered: Difference-stationary and Trend-stationary.

To obtain conditional posterior distributions for ϕ_1^o and ϕ_2^o that are conjugate, we assume independence a-priori and that each vector follows a multivariate normal distribution, i.e., $\phi_i^o \sim N(m_i, C_i)$; $i = 1, 2$; $o = 1, 2$.

Also to obtain conjugate conditional distributions, we assume that σ_1^2 and σ_2^2 are independent a-priori with an inverse gamma prior distribution, i.e., $\sigma_i^2 \sim IG(\alpha_i, \beta_i)$; $i = 1, 2$.

For the gating parameters, we assume that the vector (v_1, u_1) is independent of the vector $(v_{1|1}, u_{1|1})$ both with uniform/improper priors, i.e., $p(v_1, u_1) \equiv p(v_{1|1}, u_{1|1}) \propto 1$.

The form of the conditional posterior distributions is as follows. Assuming known initial values y_0, y_{-1}, y_{-2}, for $o = 1, 2$, let w^o be a column vector with entries $w_t^o = z_{o,1}(t)(y_t - y_{t-1})$, $t = 1, \ldots n$. For $o = 1, 2$, let X^o be a matrix with columns defined by $z_{o,1}(t)$, $z_{o,1}(t)(y_{t-1} - y_{t-2})$ and $z_{o,1}(t)(y_{t-2} - y_{t-3})$, $t = 1, \ldots, n$. Then it can be shown that the conditional posterior distribution for ϕ_1^o is $N(a_1^o, V_1^o)$, where

$$a_1^o = V_1^o \left(\frac{(X^o)^t w^o}{\sigma_1^2} + C_1^{-1} m_1 \right)^{-1} ; V_1^o = \left(\frac{(X^o)^t X^o}{{\sigma_1}^2} + C_1^{-1} \right)^{-1} .$$

For $o = 1, 2$, redefine w^o as a column vector with entries $w_t^o = z_{o,2}(t) y_t$, $t = 1, \ldots n$, and X^o a four-dimensional matrix with columns $z_{o,2}(t)$, $z_{o,2}(t)t/n$, $z_{o,2}(t)y_{t-1}$ and $z_{o,2}(t)y_{t-2}$, $t = 1, \ldots, n$. The conditional posterior distribution for ϕ_2^o follows a $N(a_2^o, V_2^o)$ where

$$a_2^o = V_2^o \left(\frac{(X^o)^t w^o}{\sigma_2^2} + C_2^{-1} m_2 \right)^{-1} ; V_2^o = \left(\frac{(X^o)^t X^o}{{\sigma_2}^2} + C_2^{-1} \right)^{-1} .$$

Furthermore, the conditional posterior distribution for σ_i^2 is an $IG(\alpha_i^*, \beta_i^*)$; $i = 1, 2$, where

$$\alpha_i^* = \alpha_i + \sum_{t=1}^{n} \sum_{o=1}^{2} z_{o,i}(t)/2; \quad \beta_i^* = \beta_i + \sum_{t=1}^{n} \sum_{o=1}^{2} z_{o,i}(t)(y_t - \mu_{o,i}(t))^2$$

with $\mu_{o,1}(t) = \phi_0^o + y_{t-1} + \phi_{1,1}^o(y_{t-1} - y_{t-2}) + \phi_{1,2}^o(y_{t-2} - y_{t-3})$ and $\mu_{o,2}(t) = \beta_0^o + \beta_1^o t/n + \phi_{2,1}^o y_{t-1} + \phi_{2,2}^o y_{t-2}; \quad o = 1, 2.$

The conditional posterior distributions for (u_1, v_1) and $(u_{1|1}, v_{1|1})$ do not have a simple analytic form and we sample them with a Metropolis-Hastings step. In both cases, we implement the Metropolis Hastings algorithm using a random-walk proposal. Particularly, the proposal distribution is a bivariate normal centered at the previous sampled value of (u_1, v_1) or $(u_{1|1}, v_{1|1})$ respectively, with a covariance matrix of the form $\delta^2 I$. The accept-reject ratio is computed with the expression for a bivariate normal density function and the part of the augmented-likelihood (see (12)) that has the function $g_{om}(t; \theta)$.

For the initial value of the MCMC-data augmentation algorithm, we use the EM solution that produced the maximum likelihood estimates. The prior is set in a "vague" fashion to illustrate how our MCMC performs. We set m_i equal to a zero vector, C_i is a diagonal matrix with diagonal elements equal to 1,000, and $\alpha_i = \beta_i = 1$, $i = 1, 2$. Other priors may be used to express personal knowledge about parameters or models. We present posterior inference based on 5,000 samples, collected after a burn-in of 10,000 iterations and skipping every 10 iterations to break possible MCMC autocorrelations. Convergence diagnostics confirm that convergence is achieved after such a number of iterations.

Table 1 presents some parameter estimates for both EM and MCMC methods with posterior standard deviations. A referee noticed that in Table 1, the parameters at $(o, m) = (1, 1)$ and $(o, m) = (2, 2)$ are not precisely estimated (i.e., the difference-stationary model parameters in the first overlay (ϕ_0^1, $\phi_{1,1}^1$ and $\phi_{1,2}^1$), as well as the trend-stationary parameters in the second overlay (β_0^2, β_1^2, $\phi_{2,1}^2$ and $\phi_{2,2}^2$), have large posterior standard deviations). One possible reason is that at $o = 2$, model $m = 1$ (difference-stationary) is dominant, while at $o = 1$, model $m = 2$ (trend-stationary) is dominant. Therefore at $o = 1$, parameters for $m = 1$ (and also the parameters for $m = 2$ at $o = 2$) cannot be precisely estimated since the weight of this model is very small. Figures 5 and 6 present the posterior mean of $g_m(t)$ and $h_m(t), m = 1, 2$, respectively. In contrast to the EM approach based on a point estimate of θ, at each MCMC iteration we compute $g_m(t)$ and $h_m(t)$ with the sampled value of θ and average across samples. In comparison, the probabilities of Figures 3 and 5 are esentially the same. The probabilities of Figure 6 are affected by individual observations and have the general pattern of Figure 4, but seem to be smooth compared to Figure 4. In general, we find very small diferences between the EM and posterior mean approaches to estimate the relative weighting of each model.

Table 1. Parameter estimates with EM and MCMC for both model-types considered: Difference-stationary and Trend-stationary.

	ϕ_0^o	$\phi_{1,1}^o$	$\phi_{1,2}^o$	β_0^o	β_1^o/n	$\phi_{2,1}^o$	$\phi_{2,2}^o$
MLE $(o = 1)$	0.160	0.284	0.273	1.279	0.008	0.697	0.249
Posterior Mean $(o = 1)$	-0.531	0.209	0.034	1.253	0.011	0.968	-0.035
Posterior SD $(o = 1)$	31.707	31.841	31.862	0.360	0.003	0.038	0.034
MLE $(o = 2)$	-0.110	0.601	0.023	0.709	0.013	2.106	-1.187
Posterior Mean $(o = 2)$	0.026	0.429	0.205	-0.309	-0.315	0.606	0.256
Posterior SD $(o = 2)$	0.025	0.058	0.052	31.582	31.205	31.786	31.523

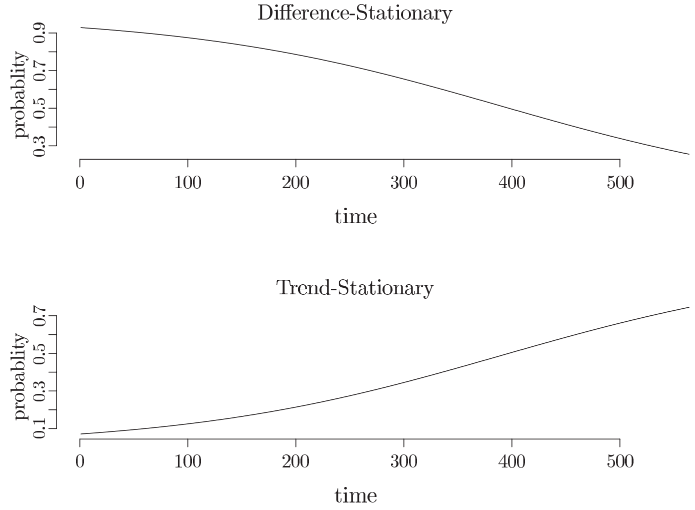


Figure 5. Posterior mean estimates of $g_m(t)$ for both model-types considered: Difference-stationary and Trend-stationary.

One advantage of implementing a full Bayesian analysis with respect to the EM algorithm is to visualize the complete behavior of posterior distributions. Diagnostic summaries are included in Figures 7 and 8. Figure 7 presents the autocorrelation function and a qqplot for $\hat{u}_t$ and $2|\hat{u}_t - 0.5|$, where $\hat{u}_t$ is computed with equation (17) of Section 2.6 and $\hat{\theta}$ is the posterior mean. In both cases, to explore the distributional assumption of u_t under a correct model, we transformed the sequence using the inverse CDF of a standard normal. The qq-plot of the transformed values of $\hat{u_t}$ and $2|\hat{u_t} - 0.5|$ shows most points at the qqline with some deviation from a Normal. Figure 8 presents a time plot of the standardized forecast errors or residuals. We notice there are some time points with large errors. On the other hand, these standardized forecast errors are not as informative for checking the model assumption on individual error terms as in the case of Lai and Wong (2001). Even though the HME model may not be completely satisfactory for modeling the data, it gives insight on model weighting and comparison. The analysis of $\hat{u_t}$ based on the EM solution results in the same conclusions. A more complicated structure for HME may be needed for a larger data set resulting in a heavier computational burden. For this example, the computing time for 250 iterations of the EM algorithm implemented in Splus and in a Sun workstation is approximately 45 minutes. Also, 10000 iterations of the MCMC method implemented in Fortran and running in a PC with a Pentium III processor is approximately 10 minutes. If we fix $O = 3$ instead of $O = 2$, the estimates of $h_m(t)$ and $g_m(t)$ are practically the same. This gives some evidence that a third overlay is unnecessary for this example. Some comments on the selection of O appear in the next section.

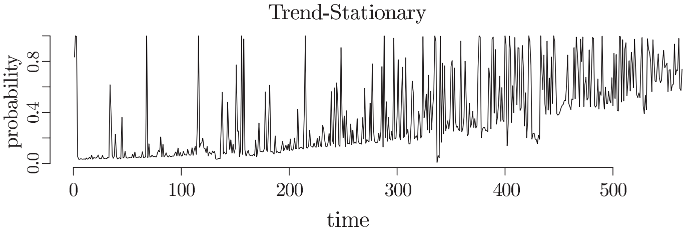


Figure 6. Posterior mean estimates of $h_m(t)$ for both model-types considered: Difference-stationary and Trend-stationary.

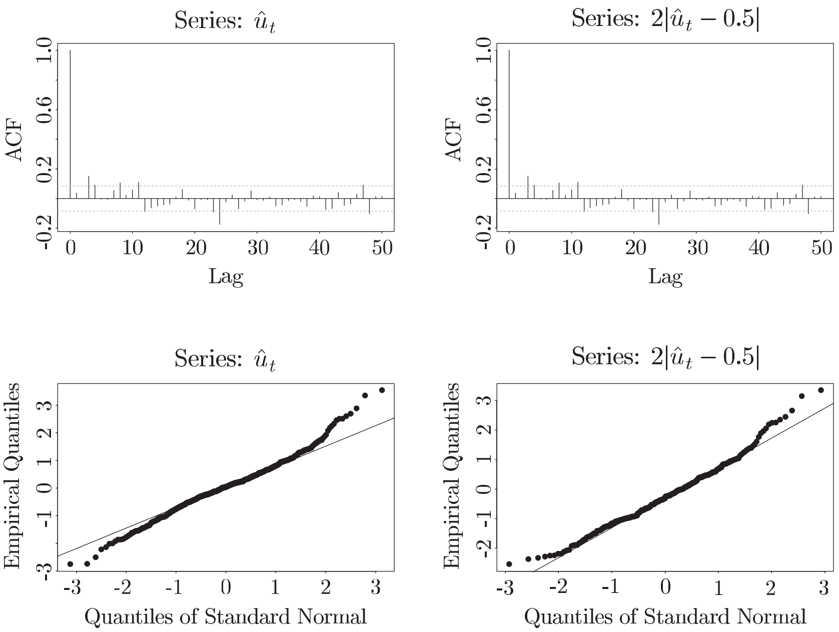


Figure 7. Graphical summaries for $\hat{u}_t$ and $2|\hat{u}_t - 0.5|$: Autocorrelation function up to lag 50 and *qqplot* of transformed series with $\Phi^{-1}(\cdot)$.

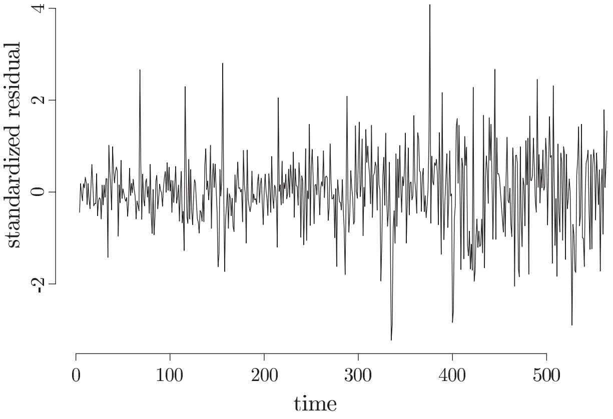


Figure 8. Time series plot of standardized residuals evaluated at the posterior mean of θ.

4. Extensions

We have not considered the situation in which the number of overlays O is uncertain. Selection of O can be performed using a criteria such as AIC or BIC which we discuss in more detail in further work. Alternatively, we could adddress this issue by using extensions to our current MCMC approach based on *reversible jump* or *variable selection* schemes that include model uncertainty on O or model uncertainty on the parameters of the defining models, such as the orders of the autoregressions in the trend−difference stationary example. These extensions will be studied in the future.

Acknowledgements

We thank the editor, associate editors and referees for their comments that greatly improved this paper. Also, we thank Andreas Weigend for useful comments. G. Huerta was supported by CONACYT grant J34413-E. M. A. Tanner was supported by NIH grant CA35464.

References

Crowder, M. (1976). Maximum likelihood estimation for dependent observations. *J. Roy. Statist. Soc. Ser. B* **38**, 45-53.

Elerian, O., Chib, S. and Shephard, N. (2001). Likelihood inference for discretely observed non-linear diffusions. *Econometrica* **69**, 959-993.

Filardo, A. J. (1994). Business cycle phases and their transitional dynamics *J. Bus. Econom. Statist.* **12**, 299-308.

Gallant, A. R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models.* Basil Blackwell, Oxford, UK.

Gerlach, R., Carter, C. and Kohn, R. (1999). Diagnostics for Time Series Analysis. *J. Time Ser. Anal.* **20**, 309-330.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357-384.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting (with discussion). *J. Roy. Statist. Soc. Ser. B* **38**, 205-247.

Huerta, G. and West, M. (1999). Priors and component structures in autoregressive time series models. *J. Roy. Statist. Soc. Ser. B* **61**, 881-899.

Huerta, G., Jiang, W. and Tanner, M. A. (2001) Discussion article: A comment on the art of Data Augmentation. *J. Comput. Graph. Statist.* **10**, 82-89.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comp.* **3**, 79-87.

Jiang, W. and Tanner, M. A. (1999). On the identifiability of mixtures-of-experts. *Neural Networks* **12**, 1253-1258.

Jordan, M. I., Ghahramani, Z. and Saul, L. K. (1996). Hidden Markov decision trees. MIT Computational Cognitive Science Technical Report 9606.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comp.* **6**, 181-214.

Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Rev. Econom. Stud.* **65**, 361-393.

Lai, T. L. and Wong, S. (2001) Stochastic neural networks with applications to nonlinear time series. *J. Amer. Statist. Assoc.* **96**, 968-981.

Lehmann, E. L. (1991). *Theory of Point Estimation.* Wadsworth, Monterey, CA.

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *J. Amer. Statist. Assoc.* **88**, 968-978.

McCulloch, R. E. and Tsay, R. S. (1994). Bayesian inference of trend- and difference-stationarity. *Econom. Theory* **10**, 596-608.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**, 347-370.

Peng, F., Jacobs, R. A. and Tanner, M. A. (1996). Bayesian inference in mixture-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *J. Amer. Statist. Assoc.* **91**, 953-959.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23**, 470-472.

Sarma, Y. R. (1986). Asymptotic properties of maximum likelihood estimators from dependent observations. *Statist. Probab. Lett.* **4**, 309-311.

Troughton, P. T. and Godsill, S. J. (1997). A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves. Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR.304.

Tanner, M. A. (1996). *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions.* (Third edition.) Springer-Verlag, New York.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82**, 528-550.

Weigend, A. S., Mangeas, M. and Srivastava, A. N. (1995). Nonlinear gated experts for time series: discovering regimes and avoid overfitting. *Internat. J. Neural Systems* **6**, 373-399.

Weiss, L. (1973). Asymptotic properties of maximum likelihood estimators in some nonstandard cases, II. *J. Amer. Statist. Assoc.* **68**, 428-430.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models.* (Second edition.) Springer-Verlag, New York.

Zeevi, A., Meir, R. and Adler, R. (1998). Nonlinear models for time series using mixtures of autoregressive models. http://www-isl.stanford.edu/∼azeevi/

Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131, U.S.A.

E-mail: ghuerta@stat.unm.edu

Department of Statistics, Northwestern University, 2006 Sheridan Rd., Evanston IL 60208-4070, U.S.A.

E-mail: wjiang@northwestern.edu

Department of Statistics, Northwestern University, 2006 Sheridan Rd., Evanston IL 60208-4070, U.S.A.

E-mail: tanm@neyman.stats.nwu.edu

(Received May 2001; accepted July 2003)