Multivariate Time Series Modeling and Classification via Hierarchical VAR Mixtures

Raquel Prado [a], Francisco J. Molina [b] and Gabriel Huerta [c]

[a] *Applied Mathematics and Statistics*
Baskin School of Engineering
1156 High Street MS-SOE2
University of California
Santa Cruz, CA 95064, USA

[b] *Department of Mathematics*
University of California
Santa Cruz, CA 95064 USA

[c] *Department of Mathematics and Statistics*
University of New Mexico
Albuquerque, New Mexico 87131-1141, USA

Abstract

A novel class of models for multivariate time series is presented. We consider hierarchical mixture-of-expert (HME) models in which the experts, or building blocks of the model, are vector autoregressions (VAR). It is assumed that the VAR-HME model partitions the covariate space, specifically including time as a covariate, into overlapping regions called overlays. In each overlay a given number of VARs compete with each other so that the most suitable model for the overlay is favored by a large weight. The weights have a particular parametric form that allows the modeler to include relevant covariates. Maximum likelihood estimation of the parameters is achieved via the EM (expectation-maximization) algorithm. The number of overlays, the number of models and the model orders of the VARs that define a particular VAR-HME model configuration are chosen by means of an algorithm based on the Bayesian information criterion (BIC). Issues of model checking and inference of latent structure in multiple time series are investigated. The new methodology is illustrated by analyzing a synthetic data set and a 7-channel electroencephalogram data set.

Email addresses: raquel@ams.ucsc.edu (Raquel Prado), fjmolina@soe.ucsc.edu (Francisco J. Molina), ghuerta@stat.unm.edu (Gabriel Huerta).

1 Introduction

We propose a multivariate time series modeling approach based on the idea of mixing models through the paradigm known as hierarchical mixture-of-experts (HME) (Jordan and Jacobs, 1994). The HME approach easily allows for model mixing and permits the representation of the mixture weights as a function of time or other covariates. Our HME models assume that the components of the mixture are vector autoregressions (VAR). These models provide useful insight into the spatio-temporal characteristics of the data by modeling multiple time series jointly. In addition, the VAR-HME models can assess, in a probabilistic fashion, the different states of the multivariate time series over time by means of the estimated mixture weights.

Developments on univariate HME time series models can be found in Huerta *et al.* (2003). These authors show how to estimate the parameters of mixture-of-expert (ME) and HME models for univariate time series via the expectation-maximization (EM) algorithm and Markov chain Monte Carlo (MCMC) methods. Huerta *et al.* (2003) applied the HME methodology to model monthly US industrial production index from 1947 to 1993. Specifically, a HME model to discriminate between stochastic trend models and deterministic trend models was considered. In this analysis time was the only covariate included in the model. More recently, Villagran and Huerta (2004) showed that the inclusion of additional covariates leads to substantial changes in the estimates of some of the model parameters in univariate mixture-of-expert models. In particular, the authors consider ME models for stochastic volatility in a time series of returns where time and the Dow Jones index are both covariates.

We present an extension of the HME developments of Huerta *et al.* (2003) to handle multivariate time series. We propose a novel class of models in which the mixture components, usually called experts in the neural network terminology, are vector autoregressions. VAR-HME models extend the univariate mixture of autoregressive (AR) models presented in Wong and Li (2000) and Wong and Li (2001) to the multivariate framework. Related univariate models, in which single-layer stochastic neural networks are used to model non-linear time series, are also developed in Lai and Wong (2001). The hierarchical structure of the VAR-HME models developed here allows the construction of very flexible models to describe the non-stationarities and non-linearities often present in multiple time series. Such hierarchical structure is not present in the univariate models developed in Wong and Li (2000), Wong and Li (2001) and Lai and Wong (2001).

The time series applications that motivate the VAR-HME modeling approach arise mainly in the area of biomedical signal processing where the multiple time series have two main characteristics. First, the series consist of multiple signals recorded simultaneously from a system under certain conditions. Second, each individual signal has an underlying structure possibly, but not necessarily, quasi-periodic, that can adequately be modeled by a collection of univariate autoregressive models or AR models with parameters that vary over time (TVAR). These are the characteristics of the multi-channel electroencephalogram (EEG) data analyzed in Section 4.2. The VAR-HME models constitute a new class of multivariate time series

models that are non-linear and non-stationary and so, they are suitable for modeling highly complex and non-stationary signals such as EEG traces. It is important to emphasize that the multivariate nature of the VAR-HME models developed here is a key feature. These models are able to capture latent process that are common to several univariate time series by modeling them jointly. This could not be achieved by analyzing each series separately via univariate mixtures of AR models. Other potential areas of application for these models include seismic and speech signal processing and applications to environmental and financial data analysis.

The paper is organized as follows. Section 2 presents the mathematical formulation of the VAR-HME models and summarizes the EM algorithm for parameter estimation when the number of overlays, the number of models and the model orders of the VAR components are known. Section 3 describes an algorithm for selecting the number of overlays, models and model orders of the VARs using the Bayesian information criterion or BIC. Model checking issues are also discussed in Section 3. Section 4 presents the analyses of two datasets: a simulated data set and a 7-channel electroencephalogram data set. Finally, conclusions and future work are presented in Section 5.

2 Models and Methodology

Let $\{\mathbf{y}_t\}_1^T$ be a collection of T k-dimensional time series vectors, and let $\{\mathbf{x}_t\}_1^T$ be a collection of T l-dimensional vectors of covariates indexed in time. Let the conditional probability density function (pdf) of $\mathbf{y}_t$ be $f_t(\mathbf{y}_t|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector; $\mathcal{X}_T$ is the σ-field generated by $\{\mathbf{x}_t\}_1^T$ representing external information; and for each t, $\mathcal{F}_{t-1}$ is the σ-field generated by $\{\mathbf{y}_s\}_1^{t-1}$ representing the previous history at time $t-1$. Typically, the conditional pdf f_t is assumed to depend on $\mathcal{X}_T$ through $\mathbf{x}_t$ only. Our main interest lies in drawing inference on $\{\mathbf{y}_t\}_1^T$ conditional on $\{\mathbf{x}_t\}_1^T$, and in drawing inference on the model parameters $\boldsymbol{\theta}$ conditional on $\{\mathbf{y}_t\}_1^T$ and $\{\mathbf{x}_t\}_1^T$.

In the univariate hierarchical mixture-of-experts context of Jordan and Jacobs (1994), which was developed for time series modeling in Huerta *et al.* (2001) and Huerta *et al.* (2003), the pdf f_t is assumed to be a conditional mixture of the pdfs from simpler models or experts. We follow a similar approach in the multivariate time series case and so, we assume that the HME model partitions the covariate space, specifically including time as a covariate, into O overlapping regions called overlays. In each overlay, M models, or experts, are to compete with each other and the most suitable model for a given time region will be favored by a high weight. Since multiple overlays are available, the HME model allows for modeling multiple switching across regions. In a multivariate time series setting, the mixture can be represented by the finite sum

$$f_t(\mathbf{y}_t|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\theta}) = \sum_{o=1}^{O} \sum_{m=1}^{M} g_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\gamma}) \pi_t(\mathbf{y}_t|\mathcal{F}_{t-1}, \mathcal{X}_T, o, m; \boldsymbol{\eta}), \tag{1}$$

where the functions $g_t(\cdot,\cdot|\cdot,\cdot;\boldsymbol{\gamma})$ are the mixtures weights, usually known as *gating functions* in the neural network terminology; $\pi_t(\cdot|\cdot,\cdot,o,m;\boldsymbol{\eta})$ are the pdfs of simpler models defined by the labels o and m; and finally, $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are sub-vectors of the parameter vector $\boldsymbol{\theta}$.

Applying the approach of Huerta *et al.* (2003) to the multivariate case, we obtain that the mixture weights have a parametric form given by

$$g_t(o,m|\mathcal{F}_{t-1},\mathcal{X}_T;\boldsymbol{\gamma}) = g_t(o|\mathcal{F}_{t-1},\mathcal{X}_T;\boldsymbol{\gamma}) \times g_t(m|o,\mathcal{F}_{t-1},\mathcal{X}_T;\boldsymbol{\gamma}) \tag{2}$$

$$= \left\{ \frac{e^{u_o+\mathbf{V}_o'\mathbf{W}_t}}{\sum_{s=1}^{O} e^{u_s+\mathbf{V}_s'\mathbf{W}_t}} \right\} \times \left\{ \frac{e^{u_{m|o}+\mathbf{V}_{m|o}'\mathbf{W}_t}}{\sum_{l=1}^{M} e^{u_{l|o}+\mathbf{V}_{l|o}'\mathbf{W}_t}} \right\}, \tag{3}$$

where the u's and the $\mathbf{v}$'s are parameters which are components of $\boldsymbol{\gamma}$. The vector $\mathbf{w}_t$ is an input at time t which is measurable with respect to the σ-field induced by $\mathcal{F}_{t-1}\cup\mathcal{X}_T$. The vector $\boldsymbol{\gamma}$ includes the parameters $u_1,\ldots,u_{O-1},\mathbf{v}_1,\ldots,\mathbf{v}_{O-1}$, $u_{1|1},\ldots,u_{M-1|1},\ldots,u_{1|O},\ldots u_{M-1|O}$, $\mathbf{v}_{1|1},\ldots,\mathbf{v}_{M-1|1},\mathbf{v}_{M-1|O},\ldots,\mathbf{v}_{M-1|O}$. For identifiability of the $\boldsymbol{\gamma}$ vector, we set $u_O=0,\mathbf{v}_O=\mathbf{0}$, $u_{M|o}=0$ and $\mathbf{v}_{M|o}=\mathbf{0}$ for all $o=1,\ldots,O$. As noted in Huerta *et al.* (2003), if the interest lies on assessing how the weighting for individual models is assigned across different time periods, $\mathbf{w}_t$ can be taken as (t/T), and so, the following parametric function of the gating functions can be adopted

$$g_t(o,m|\mathcal{F}_{t-1},\mathcal{X}_T;\boldsymbol{\gamma}) = g_{o,m}(t;\boldsymbol{\gamma}) \equiv g_o(t;\boldsymbol{\gamma}) \times g_{m|o}(t;\boldsymbol{\gamma}),$$

with

$$g_o(t;\boldsymbol{\gamma}) = \frac{e^{u_o+v_o(t/T)}}{\sum_{s=1}^{O} e^{u_s+v_s(t/T)}} \quad \text{and} \quad g_{m|o}(t;\boldsymbol{\gamma}) = \frac{e^{u_{m|o}+v_{m|o}(t/T)}}{\sum_{l=1}^{M} e^{u_{l|o}+v_{l|o}(t/T)}}, \tag{4}$$

for $o=1,\ldots,O$ and $m=1,\ldots,M$. Other options for $\mathbf{w}_t$ include, for example, taking $\mathbf{w}_t=\mathbf{x}_t$, or $\mathbf{w}_t=(\mathbf{y}_{t-1}',\mathbf{y}_{t-2}',\ldots,\mathbf{y}_{t-q}')'$ for some value of q.

In the HME models considered here, we assume that each simpler model or expert is a vector autoregression. In other words, the pdfs defined by the experts have the form

$$\pi_{t|o,m} = \Phi_{o,m}\left(\mathbf{y}_t - \sum_{j=1}^{p_m} \mathbf{A}_j^{o,m}\mathbf{y}_{t-j}\right), \tag{5}$$

where $\mathbf{A}_j^{o,m}$ is a $k\times k$ matrix and $\Phi_{o,m}(\cdot)$ denotes the pdf of a multivariate k-dimensional normal distribution with zero mean vector and variance-covariance matrix $\Sigma_{o,m}$ Note that the order of the VAR depends on m, while the VAR coefficients, as well as the covariance matrix, depend on both, o and m. In addition, we assume that $\Sigma_{o,m}$ is diagonal for all o and m, and so, all the correlation structure across the k time series is accounted for in the VAR coefficients $\mathbf{A}_j^{o,m}$. This assumption can be relaxed, however, more general structures of $\Sigma_{o,m}$

are not considered here. In practice, we have found that models with diagonal $\Sigma_{o,m}$'s matrices are very flexible and in terms of capturing most of the non-linearities and non-stationarities present in a wide range of multiple time series processes.

For fixed values of O, M and model orders $p_1, \ldots, p_M$, inferences on $\boldsymbol{\theta}$ based on the observations $\mathbf{y}_{t_0}, \ldots, \mathbf{y}_{t_1}$ are made by maximizing the log-likelihood function

$$\mathcal{L}_{t_o:t_1}(\cdot) = \frac{1}{(t_1 - t_0 + 1)} \sum_{t=t_0}^{t_1} \log f_t(\mathbf{y}_t | \mathcal{F}_{t-1}, \mathcal{X}_T; \cdot). \tag{6}$$

Typically, for model configurations with a maximum model order given by $p_{\max,M}$, with $p_{\max,M} = \max\{p_1, \ldots, p_M\}$, inferences are made by maximizing (6) with $t_0 = p_{\max,M} + 1$ and $t_1 = T$ so the estimation is conditional on the first t_0 observations. Direct maximization of the log-likelihood is difficult hence, in order to obtain the maximum likelihood estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}} = \arg\max \mathcal{L}_{t_0:t_1}(\cdot)$, we use the Expectation Maximization (EM) algorithm. A general formulation of the EM algorithm can be found in Tanner (1996). In Section 2.1 we briefly outline the EM algorithm implemented for estimation of $\boldsymbol{\theta}$ in our multivariate time series framework. The algorithm proposed here follows a scheme similar to that of the EM algorithm described in Huerta *et al.* (2003) for the univariate case.

The components of the vector $\boldsymbol{\gamma}$ are the parameters that define the gating functions and so, they determine the location and softness of the splitting periods. The number of distinct model types M can be usually specified by the practitioner, depending on the number of models that are of interest to a particular application. The number of overlays O can be selected based on subjective information related to the application, or based on historical information. Alternatively, in Section 3, we propose an algorithm that searches for the numbers of overlays, models and VAR model orders that minimize the Bayesian information criterion or BIC (Schwarz, 1978).

Conditional on a given value of $\boldsymbol{\theta}$, it is possible to look at the weights assigned to each model as a function of time. Following Huerta *et al.* (2001), there are two ways of achieving this. One is based on computing the conditional probability of a given model m, with the current observation $\mathbf{y}_t$ being conditioned on, defined by

$$P_t(m | \mathbf{y}_t, \mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\theta}) \equiv h_m(t) \equiv \sum_{o=1}^{O} h_{o,m}(t; \boldsymbol{\theta}), \tag{7}$$

where $h_{o,m}(t; \boldsymbol{\theta})$ is given by

$$h_{o,m}(t; \boldsymbol{\theta}) = \frac{g_{o,m}(t; \boldsymbol{\gamma}) \pi_t(\mathbf{y}_t | \mathcal{F}_{t-1}, \mathcal{X}_T, o, m; \boldsymbol{\eta})}{\sum_{s=1}^{O} \sum_{l=1}^{M} g_{s,l}(t; \boldsymbol{\gamma}) \pi_t(\mathbf{y}_t | \mathcal{F}_{t-1}, \mathcal{X}_T, s, l; \boldsymbol{\eta})},$$

which is the conditional probability of choosing the expert (o, m) at time t. The second

approach is to consider the unconditional probability at time t given by

$$P_t(m|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\theta}) \equiv g_m(t) \equiv \sum_{o=1}^{O} g_{o,m}(t; \boldsymbol{\gamma}). \tag{8}$$

Point estimates of (7) and (8) can be obtained by evaluation of these expressions at the MLE $\hat{\boldsymbol{\theta}}$, or in a Bayesian framework, by computing the expected value with respect to the posterior distribution $\pi(\boldsymbol{\theta}|\mathcal{F}_T, \mathcal{X}_T)$. Point estimates of (7) may vary a lot over time due to the conditioning on $\mathbf{y}_t$, while point estimates of (8) are far more smoother if the regime switching described by the gating functions depends only on the time t and no other covariate is considered.

2.1 Estimation via the EM algorithm

We now describe how to estimate the parameters of a VAR-HME model via the expectation-maximization (EM) algorithm when the model configuration denoted by $\mathcal{M}$ is known, i.e., when O, M and the VARs model orders $\mathbf{p}_{\mathcal{M}} = (p_1, \ldots, p_M)'$ are assumed to be known. The algorithm starts with an initial estimate $\boldsymbol{\theta}^0_{\mathcal{M}}$ and then, a sequence $\{\boldsymbol{\theta}^i_{\mathcal{M}}\}$ is obtained by iterating between the following steps for $i = 1, 2, \ldots, n$, where n is a specific number of iterations.

(1) *E-step.* At each iteration i the function $Q^i(\boldsymbol{\theta}_{\mathcal{M}})$ is constructed, with

$$Q^i(\boldsymbol{\theta}_{\mathcal{M}}) = \sum_{t=1}^{T} \sum_{o,m} h_{o,m}(t; \boldsymbol{\theta}^i_{\mathcal{M}}) \log\{\pi_t(\mathbf{y}_t|\mathcal{F}_{t-1}, \mathcal{X}_T, o, m; \boldsymbol{\eta}_{\mathcal{M}}) g_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\gamma}_{\mathcal{M}})\},$$

where $\boldsymbol{\theta}_{\mathcal{M}} = (\boldsymbol{\gamma}_{\mathcal{M}}, \boldsymbol{\eta}_{\mathcal{M}})$, $\boldsymbol{\theta}^i_{\mathcal{M}} = (\boldsymbol{\gamma}^i_{\mathcal{M}}, \boldsymbol{\eta}^i_{\mathcal{M}})$, $h_{o,m}(t; \boldsymbol{\theta}^i_{\mathcal{M}}) = h_{o,m}(t; \boldsymbol{\theta}_{\mathcal{M}})|_{\boldsymbol{\theta}_{\mathcal{M}} = \boldsymbol{\theta}^i_{\mathcal{M}}}$.

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

In the VAR-HME framework, the partition of the $\boldsymbol{\theta}_{\mathcal{M}}$ vector is given in terms of the parameters related to the gating functions included in $\boldsymbol{\gamma}_{\mathcal{M}}$ and the VAR parameters, $\mathbf{A}^{o,m}_j$ and $\Sigma_{o,m}$ for all o, m and $j = 1, \ldots, p_M$ included in the vector $\boldsymbol{\eta}_{\mathcal{M}}$. In this case it is possible to decompose the maximization problem into a smaller number of maximization problems. It can be shown that the equations used in the M-step to estimate $\boldsymbol{\gamma}_{\mathcal{M}}$, when the only covariate is time —and so, $\mathbf{w}_t$ can be taken as a scalar with value (t/T)— are given by

$$\sum_t \sum_m h_{o,m}(t; \boldsymbol{\theta}_{\mathcal{M}}) = \sum_t \left[\frac{e^{u_o + v_o(t/T)}}{\sum_s e^{u_s + v_s(t/T)}} \right],$$

$$\sum_t \sum_m t \times h_{o,m}(t; \boldsymbol{\theta}_{\mathcal{M}}) = \sum_t \left[t \times \frac{e^{u_o + v_o(t/T)}}{\sum_s e^{u_s + v_s(t/T)}} \right],$$

$$\sum_t \sum_m h_{o,m}(t;\boldsymbol{\theta}_{\mathcal{M}}) = \sum_t \left[\left(\sum_q h_{o,q}(t;\boldsymbol{\theta}_{\mathcal{M}}) \right) \times \frac{e^{u_{m|o}+v_{m|o}(t/T)}}{\sum_s e^{u_{s|o}+v_{s|o}(t/T)}} \right],$$
$$\sum_t \sum_m t \times h_{o,m}(t;\boldsymbol{\theta}_{\mathcal{M}}) = \sum_t \left[t \times \left(\sum_q h_{o,q}(t;\boldsymbol{\theta}_{\mathcal{M}}) \right) \times \frac{e^{u_{m|o}+v_{m|o}(t/T)}}{\sum_s e^{u_{s|o}+v_{s|o}(t/T)}} \right].$$

Similarly, the equations used in the M-step to estimate $\boldsymbol{\eta}$ are

$$\Sigma_{o,m} = \frac{\sum_t h_{o,m}(t;\boldsymbol{\theta}_{\mathcal{M}}) \cdot \left[\left(\mathbf{y}_t - \sum_{j=1}^{p_m} \mathbf{A}_j^{o,m} \mathbf{y}_{t-j} \right) \left(\mathbf{y}_t - \sum_{j=1}^{p_m} \mathbf{A}_j^{o,m} \mathbf{y}_{t-j} \right)' \right]}{\sum_t h_{o,m}(t;\boldsymbol{\theta}_{\mathcal{M}})},$$

for the variance-covariance matrices and

$$\begin{pmatrix} \left[\mathbf{A}_{p_m}^{o,m}\right]' \\ \vdots \\ [\mathbf{A}_1^{o,m}]' \end{pmatrix} = \mathbf{B}_{o,m}^{-1}(t;\boldsymbol{\theta}_{\mathcal{M}}) \times \left(\sum_t h_{om}(t;\boldsymbol{\theta}_{\mathcal{M}}) \mathbf{C}_m(t) \right),$$

for the VAR coefficients, with

$$\mathbf{B}_{o,m}(t;\boldsymbol{\theta}_{\mathcal{M}}) = \sum_t \left[h_{o,m}(t,\boldsymbol{\theta}_{\mathcal{M}}) \begin{pmatrix} [\mathbf{y}_{t-p_m}\mathbf{y}'_{t-p_m}] & \cdots & [\mathbf{y}_{t-p_m}\mathbf{y}'_{t-1}] \\ \vdots & \ddots & \vdots \\ \left[\mathbf{y}_{t-1}\mathbf{y}'_{t-p_m}\right] & \cdots & \left[\mathbf{y}_{t-1}\mathbf{y}'_{t-p_m}\right] \end{pmatrix} \right],$$

and

$$\mathbf{C}_m(t) = \begin{bmatrix} \mathbf{y}_{t-p_m}\mathbf{y}'_t \\ \vdots \\ \mathbf{y}_{t-1}\mathbf{y}'_t \end{bmatrix}.$$

The limit of the sequence $\{\boldsymbol{\theta}^i_{\mathcal{M}}\}$, denoted by $\hat{\boldsymbol{\theta}}_{\mathcal{M}}(\boldsymbol{\theta}^0)$, is a root of the likelihood equation $\nabla_{\boldsymbol{\theta}_{\mathcal{M}}} \mathcal{L}_{t_0:t_1} = 0$ corresponding to a stationary point (Tanner, 1996). If the likelihood is multimodal, the limit depends on the different modes. Multiple starting points are used to find the corresponding limits via the EM algorithm. For a given model configuration $\mathcal{M}$, the result with the largest likelihood value is chosen as the limit point estimate of $\boldsymbol{\theta}_{\mathcal{M}}$.

3 Model selection and model checking

3.1 Model selection

We now describe a general algorithm that searches for the optimal VAR-HME for a given data set. Optimality here will be defined in terms of the Bayesian information criterion or BIC (Schwarz, 1978). In other words, the optimal VAR-HME model configuration $\mathcal{M}$, will be the one whose number or overlays, number of expert models, VAR model orders and associated parameter values minimize the BIC, which is defined in this framework as

$$BIC(\mathcal{M}, \boldsymbol{\theta}_{\mathcal{M}}) = -2\mathcal{L}_{t_0:t_1}(\boldsymbol{\theta}_{\mathcal{M}}) + \dim(\boldsymbol{\theta}_{\mathcal{M}}) \log(kT^*),$$

where $\dim(\boldsymbol{\theta}_{\mathcal{M}})$ is the dimension of the parameter vector $\boldsymbol{\theta}_{\mathcal{M}}$. Here $T^* = T - p_{\max}$, where $p_{\max}$ is the maximum VAR model order considered in the search, or equivalently, the maximum VAR model order for all the model configurations considered in the search.

Since it is not possible to explore the space of all VAR-HME models and also, since there is no guarantee that running the EM a finite number of times for a given model configuration would lead to the global minimum of the BIC, a heuristic is used to select both, the number of model configurations that the algorithm visits and the number of iterations the EM algorithm is run for each particular model configuration. This selection is made at running time, and it depends to some extent on the complexity of the structure of the models that are local minima. At every iteration, a model configuration $\mathcal{M}$ with O overlays, M expert models, $\mathbf{p}_{\mathcal{M}}$ model orders and some initial parameter values are chosen. Then, the EM algorithm is applied to obtain the estimates of the model parameters. If the same model configuration is visited some iterations later, the algorithm is started again with different initial values. In the search of the model that minimizes the BIC, BIC values obtained at previous iterations determine the selection of the next points to be visited. After a certain number of model configurations have been visited, possibly more than once, the search continues through those points where it is most likely that the configuration that minimizes the BIC could be found. The search method considered here assumes that model configurations that produce the smallest BIC values after a certain number of iterations are more likely to minimize the BIC than those that lead to larger BIC values. It also assumes that model configurations that are similar to the one that minimizes the BIC after a certain number of iterations are more likely to lead to the global minimum than those that are not similar. Search methods based on these assumptions yielded very good results in several simulation studies.

In order to fully describe the search algorithm it is necessary to define the concept of similar or neighboring model configurations. Assume that a model configuration $\mathcal{M}$ has O overlays and M expert models. Let $\mathbf{p}_{\mathcal{M}} = (p_1, \ldots, p_M)'$ be the M-dimensional vector of VAR model orders of such model. Let $\mathbf{p}^*_{\mathcal{M}}$ be the $O \times M$-dimensional vector containing the VAR model orders for all the possible combinations of overlays and expert models ordered in increasing order. So, for example, let $\mathcal{M}_1$ be a model with $O = 1$, $M = 2$ and model orders $p_1 = 1$

and $p_2 = 1$. Then, $\mathbf{p}_{\mathcal{M}_1} = (1,1)'$ and $\mathbf{p}^*_{\mathcal{M}_1} = (1,1)'$. Similarly, a model $\mathcal{M}_2$ with $O = 2$, $M = 2$ and $p_1 = 3$, $p_2 = 1$ leads to $\mathbf{p}_{\mathcal{M}_2} = (3,1)'$ and $\mathbf{p}^*_{\mathcal{M}_2} = (1,1,3,3)'$. Note that O, M and $p_1, \ldots, p_M$ determine $\mathbf{p}^*_{\mathcal{M}}$, however, a given vector $\mathbf{p}^*$ can be associated to various model configurations. For instance, consider a model $\mathcal{M}_3$ with $O = 2$, $M = 1$ and $p_1 = 1$. Then, $\mathbf{p}^*_{\mathcal{M}_3} = (1,1)'$ and so, $\mathbf{p}^*_{\mathcal{M}_1} = \mathbf{p}^*_{\mathcal{M}_3}$ even though $\mathcal{M}_1 \neq \mathcal{M}_3$. Now, using $\mathbf{p}^*$ it is possible to define a class of model configurations that are equivalent to a given model configuration $\mathcal{M}$, say $\mathbf{EQ}(\mathcal{M})$, as follows

$$\mathbf{EQ}(\mathcal{M}) = \{\text{All models } \mathcal{M}_i \text{ such that } \mathbf{p}^*_{\mathcal{M}_i} = \mathbf{p}^*_{\mathcal{M}}\}.$$

Similarly, it is possible to define the class of neighboring model configurations of $\mathcal{M}$, $\mathbf{NB}(\mathcal{M})$, given by

$$\mathbf{NB}(\mathcal{M}) = \{\text{All models } \mathcal{M}_i \text{ such that } \mathbf{p}^*_{\mathcal{M}_i} \in \mathbf{C}\},$$

where $\mathbf{C}$ is the set of vectors $\mathbf{p}^*_{\mathcal{M}_i}$ that are derived from $\mathbf{p}^*_{\mathcal{M}}$ in one of the following ways:

- By adding 1 or 0 to one of the components of $\mathbf{p}^*(\mathcal{M})$ and subtracting 1 or 0 to one of the components of $\mathbf{p}^*(\mathcal{M})$, i.e., $\mathbf{p}^*_{\mathcal{M}_i}(j_0) = \mathbf{p}^*_{\mathcal{M}}(j_0) + a_0$, for some component j_0 and $\mathbf{p}^*_{\mathcal{M}_i}(j_1) = \mathbf{p}^*_{\mathcal{M}}(j_1) - a_1$ for some other component j_1 such that $j_0 \neq j_1$ and $\mathbf{p}^*_{\mathcal{M}_i}(j) = \mathbf{p}^*_{\mathcal{M}}(j)$ for all the other components j, such that $j \neq j_0, j \neq j_1$. Here $a_0 = 0, 1$ or $a_1 = 0, 1$.
- By adding 1 or 2 to one of the components of $\mathbf{p}^*(\mathcal{M})$ and subtracting 1 or 2 to one of the components of $\mathbf{p}^*(\mathcal{M})$, i.e., $\mathbf{p}^*_{\mathcal{M}_i}(j_0) = \mathbf{p}^*_{\mathcal{M}}(j_0) + b_0$, for some component j_0 and $\mathbf{p}^*_{\mathcal{M}_i}(j_1) = \mathbf{p}^*_{\mathcal{M}}(j_1) - b_1$ for some other component j_1 such that $j_0 \neq j_1$ and $\mathbf{p}^*_{\mathcal{M}_i}(j) = \mathbf{p}^*_{\mathcal{M}}(j)$ for all the other components j, such that $j \neq j_0, j \neq j_1$. Here $b_0 = 1, 2$ or $b_1 = 1, 2$.
- By removing one component in the mixture and so, $\dim(\mathbf{p}^*_{\mathcal{M}_i}) = \dim(\mathbf{p}^*_{\mathcal{M}}) - 1$.
- By adding one new component to the mixture and so, $\dim(\mathbf{p}^*_{\mathcal{M}_i}) = \dim(\mathbf{p}^*_{\mathcal{M}}) + 1$. It is assumed that the model order that is added differs in at most 1 from the model orders in $\mathbf{p}^*_{\mathcal{M}}$.

Increasing or decreasing by one or two a given VAR model order is equivalent to adding or eliminating at least one real or complex root to the characteristic polynomial associated to such VAR mixture component. At a given iteration, the algorithm keeps the lowest BIC found so far, as well as the model configuration and the estimates of the parameters associated to this configuration. All of them are updated when a lower value of the BIC is found.

3.1.1 Stopping and reinitialization rules

The search algorithm stops or is restarted at a given iteration when one of the following happens:

- Every neighboring model configuration of the current model configuration and corresponding parameter estimates that minimize the BIC has been visited a number n_1 of times.

In this case the algorithm reports the optimal model configuration and its corresponding parameter estimates.

- If one of the elements of the diagonal matrix $\Sigma_{o,m}$ is smaller than some threshold τ_1, with τ_1 very close to zero or, if $\sum_{t=p_{\max}+1}^{T} h_{o,m}(t;\boldsymbol{\theta}) < \tau_2$, with τ_2 close to zero for some m, o then a new run of the algorithm is started with initial estimates build in the following way:
  - the parameters $\mathbf{A}_1^{o,m}, \ldots, \mathbf{A}_p^{o,m}$ and $\Sigma_{o,m}$ are initialized at random values;
  - the rest of the VAR parameters, $\mathbf{A}_j^{s,l}$ for $s \neq o$ and $l \neq m$, are set at the values they had at the previous iteration and all the u's and $\mathbf{v}$'s parameters are set to zero.

  If the number of partial reinitializations exceeds a threshold n_2 then a complete reinitialization of the algorithm takes place so that the initial values of this reinitialization are unrelated to the results of the previous convergence. If the number of complete reinitializations reaches n_3 then the EM algorithm stops without convergence to a local minimum.
- The search stops when the relative differences of the VAR parameters at consecutive iterations is smaller than τ_3, or if the number of iterations from the last time the EM was reinitialized is greater than n_4. In both cases the algorithm reports that a local minimum has been successfully found.

3.2 Model checking

As in Huerta *et al.* (2003), and following Kim *et al.* (1998) and Elerian *et al.* (2001), we use the one-step-ahead predictive distribution function for model checking. In our case this distribution is given by

$$F_t(\mathbf{y}_t|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\theta}) = \sum_{o=1}^{O} \sum_{m=1}^{M} g_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\gamma}) F_{\pi_t}(\mathbf{y}_t|\mathcal{F}_{t-1}, \mathcal{X}_T, o, m; \boldsymbol{\eta}),$$

where $F_{\pi_t}(\cdot|\cdot, \cdot, o, m; \boldsymbol{\eta})$ is the distribution function of the simpler models, each indexed by the pair (o, m). In the univariate HME context, Huerta *et al.* (2003) use a transformation proposed by Rosenblatt (1952) to show that, if the model is correctly specified then $\{u_t\}$, with $u_t = F_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\theta})$ and y_t a univariate time series process, is a sequence of independent random variables uniformly distributed in the interval $(0, 1)$.

Here we consider Rosenblatt's transformation for each component of the time series, and so, if the model is correct, each sequence of random variables $\{u_{t,j}\}$, for $j = 1, \ldots, k$, will be a sequence of independent random variables uniformly distributed in the interval $(0, 1)$. In this case each $u_{t,j}$ is defined as $u_{t,j} = F_{t,j}(y_{t,j}|\mathcal{F}_{t-1}, \mathcal{X}_T; \boldsymbol{\theta})$, where $F_{t,j}(\cdot|\cdot, \cdot; \boldsymbol{\theta})$ is the marginal distribution function for the j-th component of the time series. Each $u_{t,j}$ can be estimated by $\hat{u}_{t,j} = F_{t,j}(y_{t,j}|\mathcal{F}_{t-1}, \mathcal{X}_T; \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the ML estimate obtained via the EM algorithm. As in Huerta *et al.* (2003), we judge model adequacy by looking at distributional characteristics of $\{\hat{u}_{t,j}\}$, for $j = 1, \ldots, k$, or with transformed values via the Normal inverse cdf. We also look at the correlation and distributional forms of $2|\hat{u}_{t,j} - 0.5|$, as suggested by Kim *et al.*

(1998). Model checking techniques will be illustrated in the data analyses presented in the following Section.

4 Applications

In order to show the performance of the proposed models and search algorithm, two examples are considered. In the first example we analyze a simulated bivariate time series. In the second example we apply the VAR-HME models to a 7-channel electroencephalogram data recorded during electroconvulsive therapy.

4.1 A simulation study

A bivariate time series of 1,000 observations was generated according to a VAR-HME with two overlays ($O = 2$) and two expert models ($M = 2$). The VAR models were parameterized as follows:

$$O = 1,\ \ M = 1,\ \ \mathbf{y}_t = \underbrace{\begin{pmatrix} -0.750 & -0.250 \\ -0.250 & -0.250 \end{pmatrix}}_{\mathbf{A}_1^{1,1}} \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t^{(1,1)},$$

$$O = 2,\ \ M = 1,\ \ \mathbf{y}_t = \underbrace{\begin{pmatrix} 0.250 & 0.500 \\ 0.750 & 0.250 \end{pmatrix}}_{\mathbf{A}_1^{2,1}} \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t^{(2,1)},$$

$$O = 1,\ \ M = 2,\ \ \mathbf{y}_t = \underbrace{\begin{pmatrix} -0.250 & 0.000 \\ -0.250 & -0.500 \end{pmatrix}}_{\mathbf{A}_1^{1,2}} \mathbf{y}_{t-1} + \underbrace{\begin{pmatrix} 0.125 & 0.125 \\ 0.125 & 0.250 \end{pmatrix}}_{\mathbf{A}_2^{1,2}} \mathbf{y}_{t-2} + \boldsymbol{\epsilon}_t^{(1,2)},$$

$$O = 2,\ \ M = 2,\ \ \mathbf{y}_t = \underbrace{\begin{pmatrix} 0.500 & 1.000 \\ 0.750 & 1.250 \end{pmatrix}}_{\mathbf{A}_1^{2,2}} \mathbf{y}_{t-1} + \underbrace{\begin{pmatrix} -0.125 & -0.500 \\ -0.375 & -0.563 \end{pmatrix}}_{\mathbf{A}_2^{2,2}} \mathbf{y}_{t-2} + \boldsymbol{\epsilon}_t^{(2,2)},$$

where $\boldsymbol{\epsilon}_t^{o,m} \sim N(\mathbf{0}, \Sigma_{o,m})$, with $\Sigma_{o,m} = \text{diag}(7^2, 7^2)$ for all o and m. In addition, the following parameters were used for the gating functions: $u_1 = 259.540, v_1 = -519.081$, $u_{1|1} = 130.532, v_{1|1} = -522.128$, $u_{1|2} = -63.054$ and $v_{1|2} = 84.072$.

The search algorithm described in the Section 3 was initialized at a model configuration $\mathcal{M}_0$ with $O_{\mathcal{M}_0} = 2$, $M_{\mathcal{M}_0} = 4$ and $\mathbf{p}_{\mathcal{M}_0} = (1, 2, 3, 4)$. The values of the parameters that define the stopping and reinitialization rules in the search algorithm were set as follows: $n_1 = 20$, $n_2 = 20$, $n_3 = 500$, $n_4 = 500$, $\tau_1 = 10^{-5}$, $\tau_2 = 0.001$, $\tau_3 = 0.01$. The initial values for the $\mathbf{A}$'s and Σ's matrices were randomly generated from uniform distributions on the intervals $[-1, 1]$ and $[1, 100]$, respectively. The maximum VAR model order considered in the search was $p_{\max} = 10$.

The optimal VAR-HME model configuration, denoted by $\mathcal{M}_1$ found by the algorithm has the form $O_{\mathcal{M}_1} = 2$, $M_{\mathcal{M}_1} = 2$ and $\mathbf{p}_{\mathcal{M}_1} = (1, 2)$, which corresponds to the correct model configuration. The total number of model configurations visited was 955.

The parameter estimates for the optimal VAR-HME model configuration are

$$\hat{\mathbf{A}}_1^{1,1} = \begin{pmatrix} -0.764 & -0.275 \\ -0.239 & -0.259 \end{pmatrix}, \quad \hat{\mathbf{A}}_1^{2,1} = \begin{pmatrix} 0.240 & 0.533 \\ 0.704 & 0.234 \end{pmatrix},$$

$$\hat{\mathbf{A}}_1^{1,2} = \begin{pmatrix} -0.339 & 0.021 \\ -0.238 & -0.480 \end{pmatrix}, \quad \hat{\mathbf{A}}_2^{1,2} = \begin{pmatrix} 0.164 & 0.147 \\ 0.074 & 0.289 \end{pmatrix},$$

$$\hat{\mathbf{A}}_1^{2,2} = \begin{pmatrix} 0.511 & 0.973 \\ 0.750 & 1.251 \end{pmatrix}, \quad \hat{\mathbf{A}}_2^{2,2} = \begin{pmatrix} -0.227 & -0.400 \\ -0.339 & -0.594 \end{pmatrix},$$

for the VAR coefficients and $\hat{\Sigma}_{1,1} = \text{diag}(6.905^2, 6.105^2)$, $\hat{\Sigma}_{2,1} = \text{diag}(7.512^2, 6.516^2)$, $\hat{\Sigma}_{1,2} = \text{diag}(7.548^2, 6.939^2)$ and $\hat{\Sigma}_{2,2} = \text{diag}(5.628^2, 6.708^2)$ for the variance-covariance matrices.

Figure 1 shows the trajectories of the gating functions $g_t(o, m)$ used to generate the data (solid lines) and their estimates (dotted lines) for the four combinations of overlays and models. These plots show that the optimal VAR-HME model adequately captures the structure of the simulated series. Graph (a) in Figure 2 displays the trajectory of $g_{m=1}(t)$ (solid line) and its estimate (dotted line) as a function of time. Here $m = 1$ corresponds to a VAR(1) model-type. $g_{m=1}(t)$ is the unconditional probability of choosing a VAR(1) model-type at time t computed using (8). Panel (b) in this figure shows the estimate for $h_{m=1}(t)$, the conditional probability of model $m = 1$ given in equation (7). These graphs split the data in two parts, the initial portion and final portion of the series in which a VAR(1)-type of model is favored, and the middle portion of the series in which a VAR(2)-type of model is favored. Graph (a) is a smoother version of graph (b), which is less affected by individual observations.

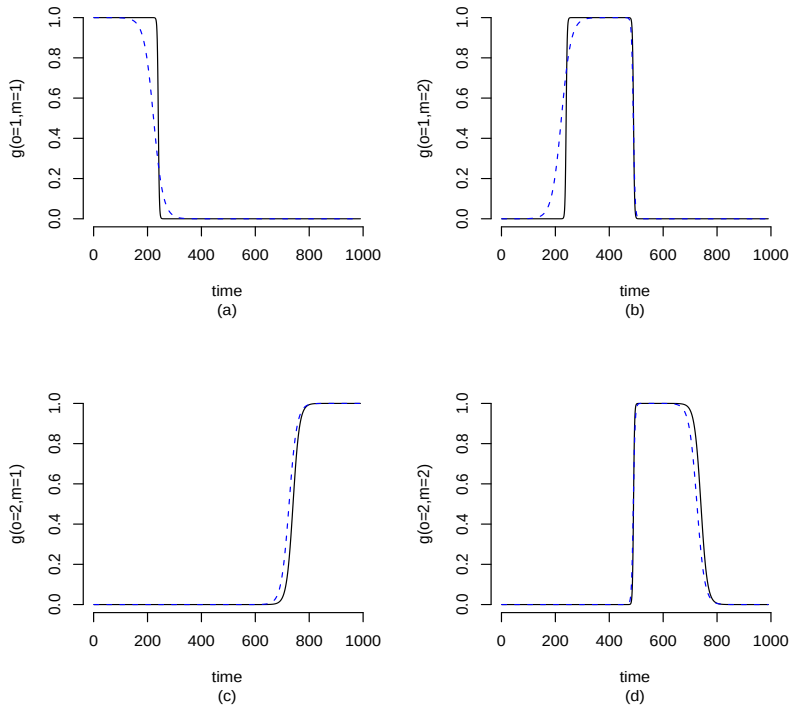

Fig. 1. The gating functions $g_t(o, m)$ (solid lines) used in the simulation and their maximum likelihood estimates (dotted lines) for all the combinations of overlays and models considered: two overlays and two model types, a VAR(1) and a VAR(2).

Diagnostic summaries appear in Figures 3 and 4. These graphs show the autocorrelation functions and qqplots for $\hat{u}_{t,1}$ and $2|\hat{u}_{t,1} - 0.5|$ in Figure 3, and for $\hat{u}_{t,2}$ and $2|\hat{u}_{t,2} - 0.5|$ in Figure 4. Here the estimates, $\hat{u}_{t,j}$ for $j = 1, 2$ are computed as $\hat{u}_{t,j} = F_{t,j}(y_{t,j}|\mathcal{F}_{t-1}, \mathcal{X}_T; \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$ obtained from the EM algorithm. In order to explore the distributional assumptions of $u_{t,j}$ for $j = 1, 2$ under the correct model, we transformed the sequences using the inverse cdf of a standard Normal distribution. The qqplots in Figures 3 and 4 show that most points lie at the qqline, with a few displaying some deviation from the standard Normal distribution. A careful look at the points that deviate from the Normal shows that such points correspond to observations around the transition times when jumps from one of the overlay/expert models to another one occurs. This is expected as the series were simulated from a model in which the switches between various expert models and overlays occur in a very abrupt manner (see Figure 1).

4.2 Analysis of multichannel EEG data

We consider the analysis of 7 EEG signals recorded on a patient who received electroconvulsive therapy (ECT), a treatment for major depression. The 7 signals are part of a multichannel recording of 19 EEG channels located over the patient's scalp. Figure 5 displays a scheme of the approximate location of the 19 electrodes over the scalp. The seven sites

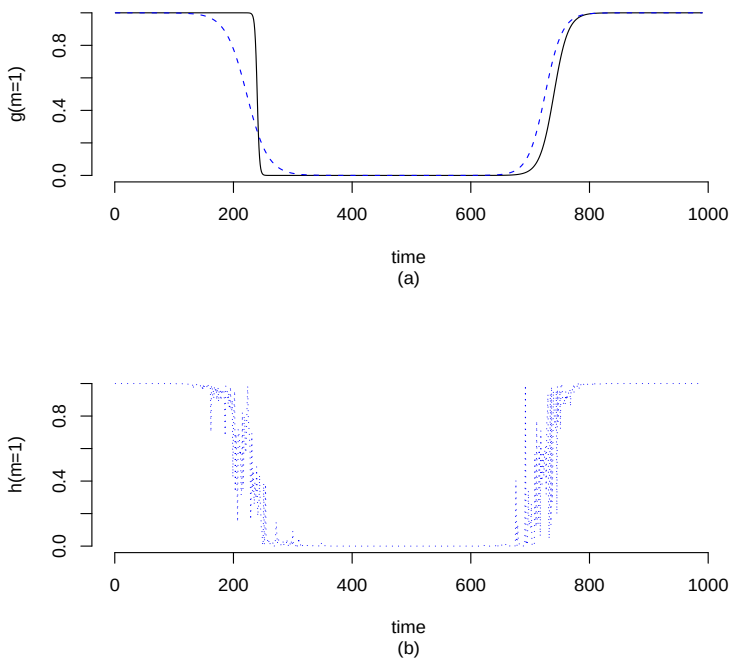

Fig. 2. (a): $g_t(m = 1)$ (solid line) and its maximum likelihood estimate (dotted line). (b) Maximum likelihood estimates of $h_m(t)$ for model-type VAR(1)

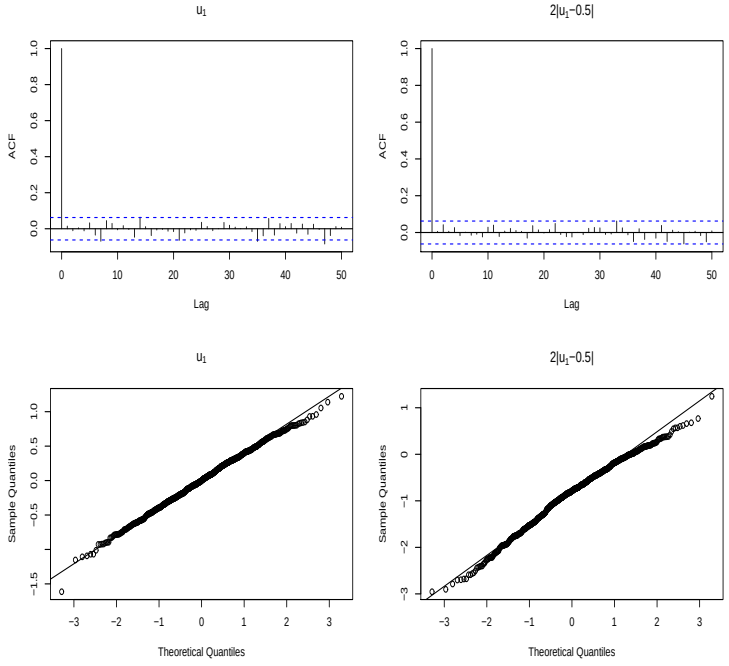

Fig. 3. Graphical summaries for $\hat{u}_{t,1}$ and $2|\hat{u}_{t,1} - 0.5|$.

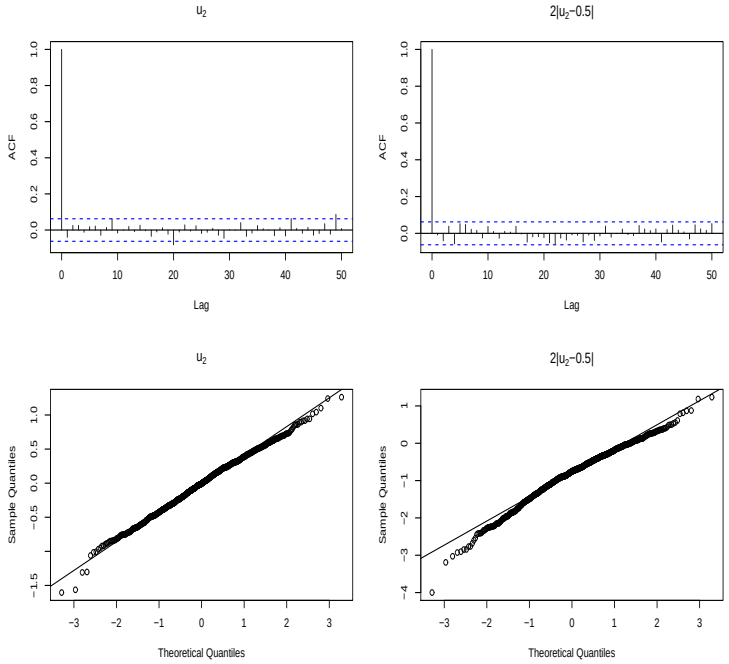

Fig. 4. Graphical summaries for $\hat{u}_{t,2}$ and $2|\hat{u}_{t,2} - 0.5|$.

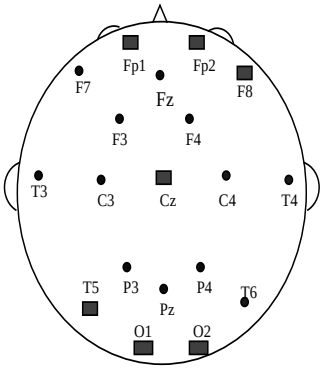

Fig. 5. Schematic Ictal-19 electrode placement. The electrodes marked with a rectangle correspond to those analyzed here using HME models.

marked with a rectangle correspond to those analyzed here with a VAR-HME model. Univariate analyses of the 19 EEG signals via time-varying autoregressive (TVAR) models can be found in Prado (1998); West *et al.* (1999) and Prado *et al.* (2001). The EEG signals studied here were recorded at channels labeled as $Fp_1, Fp_2, F_8, C_z, T_5, O_1$ and O_2 (from the top down and left to right in Figure 5). According to EEG nomenclature Fp, F, C, T and O stand for *prefrontal, frontal, central, temporal* and *occipital* regions, respectively. Electrodes located down the center of the scalp are labeled with the letter z. Even numbered channels are located to the right while odd numbered channels are located to the left.

Previous analyses of individual channels via TVAR models with time-varying orders suggested that models with different orders were appropriate to describe the complexity of the time series over time (Prado and Huerta, 2002). In particular, such analyses found that rel-

search	$\mathbf{p}^*_{\mathcal{M}_0}$	$\mathbf{p}^*_{\mathcal{M}_1}$	*#models visited*
1	$(2,5)$	$(1,1,1,1,1,2,2,2)$	1,243
2	$(1,2,3)$	$(1,1,1,1,1,1,1,2,2)$	848
3	$(1,2,3,4)$	$(1,1,1,1,1,1,1,2,2)$	811
4	$(1,1,1,1,1,1,2,2,2,2)$	$(1,1,2,2,2,2,2)$	75
5	$(1,1,1,1,1,2,2,2,2)$	$(1,1,1,1,1,1,1,1,2,2)$	400
6	$(1,1,1,1,1,2,2,2)$	$(1,1,1,1,1,2,2,2)$	640

search	$O_{\max}$	$M_{\max}$	$p_{\max}$	BIC
1	5	10	6	199276.8
2	3	10	4	198955.2
3	4	11	6	199122.0
4	3	10	4	199826.0
5	4	11	5	199185.8
6	4	11	4	199026.6

Table 1
Results obtained from six runs of the search algorithm when applied to the 7-channel EEG data

atively high model orders (around 6-12) were needed in earlier and middle portions of the signals while lower order models (around 2-4) were appropriate for the latter portions of the signal.

In this Section we simultaneously study the 7 channels listed above using multivariate VAR-HME models. Several trials of the algorithm described in Section 3 were initialized at different model configurations and multiple starting points were chosen for the parameters of each initial model configuration. Table 1 summarizes some of the results obtained after running the algorithm 6 times using different model configurations as starting points. The parameters that define the reinitialization and stopping criteria of the algorithm were set as follows: $n_1 = n_2 = 20$, $n_3 = n_4 = 500$, $\tau_1 = 10^{-5}$, $\tau_2 = 10^{-3}$ and $\tau_3 = 10^{-2}$. A maximum model order $p_{\max} = 30$ was allowed. All the final configurations had one overlay, even though the algorithm visited several model configurations with 2,3,4 and even 5 overlays. Model configurations with a minimum of $M = 1$ experts and a maximum of $M = 11$ experts were visited. The maximum model order for all the VAR components visited was 6. Table 1 displays the initial $\mathbf{p}^*_{\mathcal{M}_0}$ vectors that contain the model orders for each expert, as well as the vector $\mathbf{p}^*_{\mathcal{M}_1}$ containing the model orders for each expert model of the final model configuration found by the algorithm before stopping. The number of model configurations, the maximum number of overlays, the maximum number of models and the maximum VAR model orders, as well as the optimal BIC for each search are also shown in Table 1. The optimal VAR-HME model configuration from the six trials was obtained in the second search. We denote

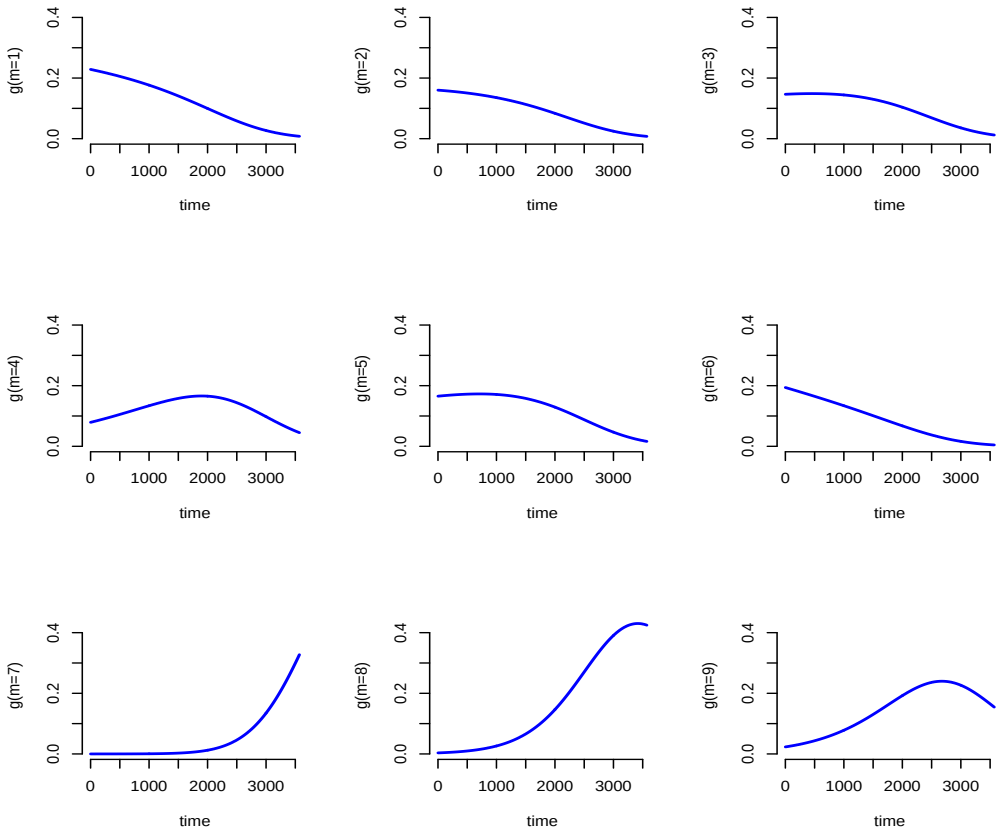

Fig. 6. Maximum likelihood estimates of $g_m(t)$ for $m = 1, \ldots, 9$

this configuration as $\mathcal{M}_1^*$. This model configuration is defined by $O_{\mathcal{M}_1^*} = 1$, $M_{\mathcal{M}_1^*} = 9$ and $\mathbf{p}^*_{\mathcal{M}_1^*} = (1, 1, 1, 1, 1, 1, 1, 2, 2)$.

Figures 6 and 7 show the ML estimates of the traces of $g_m(t)$ and $h_m(t)$, respectively, for $m = 1, \ldots, 9$. From these figures we see that expert models $m = 1$ to $m = 6$ have more weight in initial and middle portions of the series, while expert models $m = 7, m = 8$ and $m = 9$ have more weight toward the end of the series. The fact that six model components are used to describe initial and middle portions of the series, while three components are enough to explain the behavior of the series toward the end of the seizure indicates that the latent structure of the multiple series is more complex at initial and middle portions of the EEGs. Figure 8 shows the added trajectories of the estimated gating functions for the expert models $m = 1, \ldots, 6$. It is obvious from this picture that components 1 to 6 do a very good job in explaining initial and middle portions of the multiple series. These findings are consistent with those summarized in Prado and Huerta (2002) in the analysis of a single EEG channel. The latent structure of the series has a greater complexity before the seizure starts to dissipate and therefore, more components are needed to capture such structure at initial and middle parts.

The variance-covariance ML estimates obtained from the algorithm for the optimal model

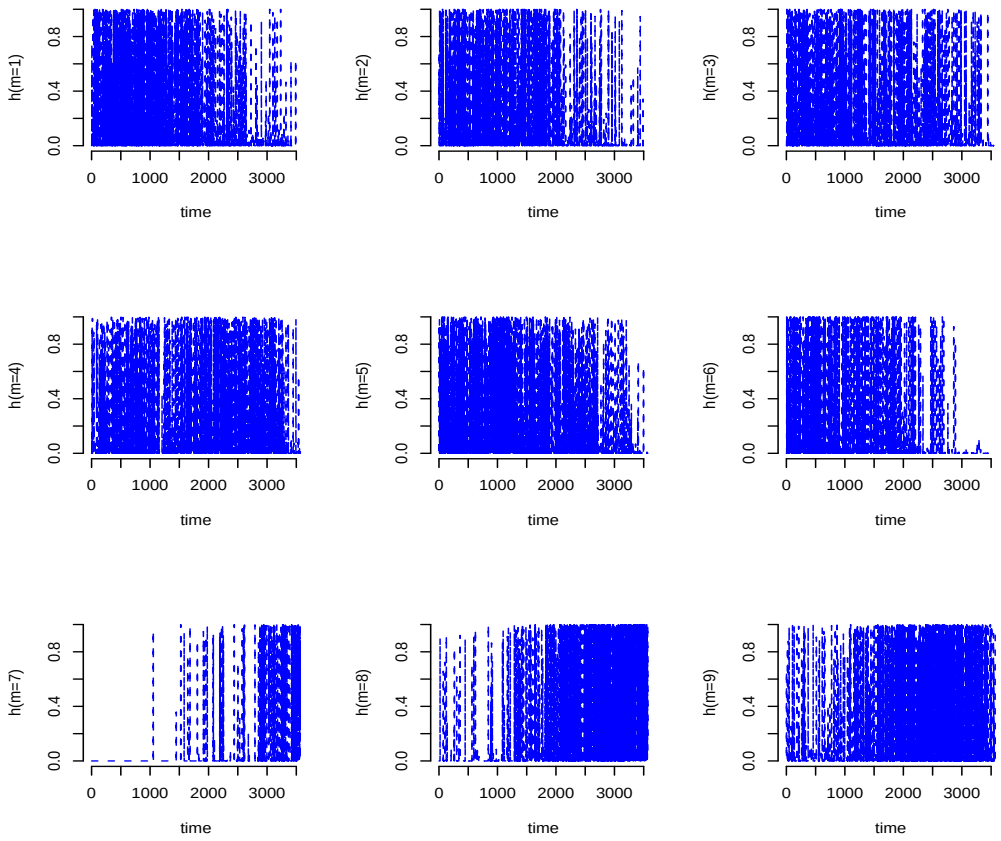

Fig. 7. Maximum likelihood estimates of $h_m(t)$ for $m = 1, \ldots, 9$

configuration are given by

$$\hat{\Sigma}_{1,1} = \text{diag}(1165.3, 1148.3, 1105.2, 1305.5, 516.0, 2693.3, 363.0),$$
$$\hat{\Sigma}_{1,2} = \text{diag}(928.0, 1183.7, 758.0, 1197.7, 603.6, 2708.0, 407.6),$$
$$\hat{\Sigma}_{1,3} = \text{diag}(2413.9, 1259.7, 2101.5, 2529.7, 706.2, 2504.7, 474.9),$$
$$\hat{\Sigma}_{1,4} = \text{diag}(795.2, 953.3, 564.0, 835.3, 300.7, 993.9, 271.4),$$
$$\hat{\Sigma}_{1,5} = \text{diag}(553.9, 1319.3, 546.9, 887.0, 422.2, 2041.8, 346.9),$$
$$\hat{\Sigma}_{1,6} = \text{diag}(5504.1, 3749.6, 4719.5, 10082.0, 3362.3, 6955.8, 1864.2),$$
$$\hat{\Sigma}_{1,7} = \text{diag}(29.8, 49.6, 33.6, 20.4, 18.4, 51.0, 16.5),$$
$$\hat{\Sigma}_{1,8} = \text{diag}(213.6, 176.8, 221.7, 134.9, 83.7, 254.7, 69.0),$$
$$\hat{\Sigma}_{1,9} = \text{diag}(278.4, 405.5, 245.9, 229.2, 127.5, 357.5, 109.9).$$

The variance is a measure of the amplitude of the signal and so, these results are consistent with the fact that, on average, the EEGs have higher amplitude at initial and central portions of the series — represented by variance covariance matrices $\hat{\Sigma}_{1,i}$ for $i = 1, \ldots, 6$ — and lower amplitudes toward the end — represented by variance-covariance matrices $\hat{\Sigma}_{1,i}$ for $i = 7, 8, 9$

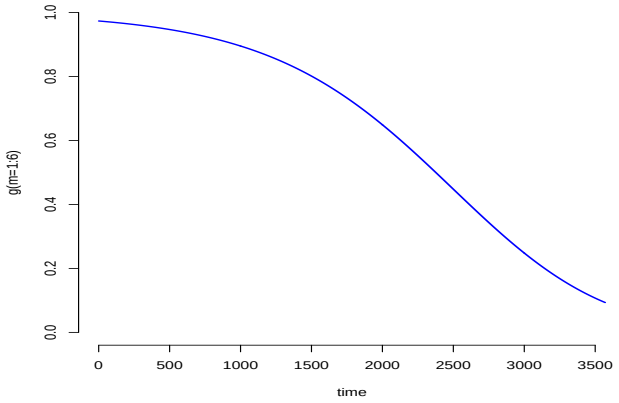

Fig. 8. Estimated trace $\sum_{j=1}^{6} \hat{g}_t(j)$ for the EEG data

— when the seizure starts to dissipate. In addition, ordering the channels by increasing variance for each of the components it is possible to see that channels F_3 and Fp_2 (frontal-left and pre-frontal right, respectively) consistently display the lowest amplitude over the seizure course. The channels displaying the highest amplitude vary over the seizure course, with C_z and O_2 having the highest amplitudes at the beginning of the series and O_2, O_1 and Fp_1 displaying the highest amplitude toward the end. These findings are consistent with those reported in the multi-channel univariate analysis of Prado *et al.* (2001). In addition, the VAR-HME analysis presented here provides further insight into the structure of the multiple EEG series over time. The estimates of $g_m(t)$ and $h_t(m)$ for $m = 1, \ldots, 9$ indicate a change in the structure of the multiple time series that is associated to seizure dissipation. In particular, Figure 8 shows a decrease in the probabilities related to components $m = 1$ to $m = 6$ starting at around $t = 1500$ to $t = 2000$. Detecting the beginning of seizure dissipation is believed to be relevant in assessing the efficacy of ECT treatment (Prado and Huerta, 2002). The multivariate VAR-HME models allow us to estimate seizure dissipation in a probabilistic fashion using several EEG channels to make such inferences instead of a single channel. This is clearly an advantage of using these multivariate models instead of simpler univariate alternatives.

The latent structure of the multivariate time series can be investigated by looking at the moduli and wavelengths (or frequencies) of the characteristic polynomials of each of the VAR models present in the mixture. A time series decomposition for each of the seven series can also be computed using the multivariate decomposition results presented in Prado (1998). Here we only summarize the characteristic root structure. Specifically, Table 2 shows the moduli and wavelengths of the characteristic roots whose moduli were higher than 0.8, for each of the VAR components in the VAR-HME model. Expert $m = 7$ has a very high modulus component with a wavelength that is higher than any of the wavelengths in expert models $m = 4$ and $m = 1$. This is an indication that the dominant frequency components in the series tend to decrease toward the end of the seizure.

Finally, Figure 9 displays the qqplots of the transformed values of $2|\hat{u}_{t,j} - 0.5|$ for the EEG

Component	*Real roots* r_j	*Complex roots* (r_j, λ_j)
$m = 1$	—	$(0.89, 18.71)$
$m = 2$	1.05	—
$m = 3$	1.78; 0.94	—
$m = 4$	0.99	$(0.85, 79.36)$
$m = 5$	1.00	—
$m = 6$	1.05	—
$m = 7$	1.07; 0.94; 0.87	$(0.98, 112.13)$
$m = 8$	0.98	$(0.90, 19.73)$
$m = 9$	0.99	$(0.93, 15.08)$

Table 2
Characteristic roots with moduli higher than 0.8 for each of the VAR-HME components in the model configuration $\mathcal{M}_1^*$

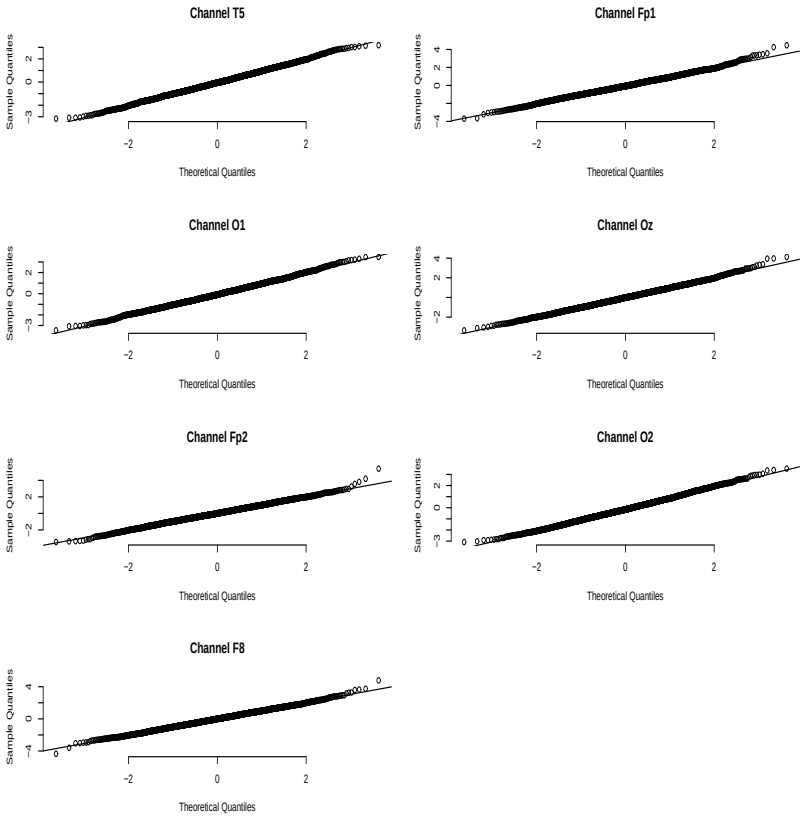

Fig. 9. qqplots of $2|\hat{u}_{t,j} - 0.5|$ for the EEG components $j = 1, \ldots, 7$.

series $j = 1, \ldots, 7$. The graphs show that most of the points for the 7 series lie at the qqline and only a few display some deviation from the Normal.

5 Conclusions and future directions

This paper provides advancement in multivariate time series methodology that combines hierarchical mixtures and vector autoregressive components. Our modeling approach is illustrated with the analysis of synthetic data and a 7-channel EEG multiple times series. We propose an innovative search algorithm based on BIC that determines the number of overlays, experts and model orders for our VAR-HME framework. This is a step beyond other previous analyses done with HME time series models. Both data examples exhibit that this search algorithm is efficient and determines good model structures. It is also shown, particularly in the EEG data analysis, that VAR-HME constitute a very flexible class of models to describe the structure of multiple time series, offering advantages with respect to simpler modeling alternatives such as univariate mixture of autoregressive models.

On the other hand, this paper only addresses point estimation of the model parameters via an EM algorithm and not a full Bayesian analysis. A complete Bayesian approach can be achieved using MCMC/Data Augmentation algorithms as exposed in Huerta *et al.* (2003). However, in this context it is not trivial how to generalize the model search algorithm. One possible road is to include model uncertainty in the number of overlays, number of model parameters and model orders via a sophisticated Reversible jump Markov chain Monte Carlo method (Green, 1995). This will require careful thought of the prior specification since default uniform priors in mixture models usually lead to improper posteriors. It is also well known that specifying sensible prior distributions for multivariate time series models is, in general, a very difficult task (Huerta and Prado, 2003). Another option is to compare models for different values of overlays and for experts with different model orders with Bayes factors obtained via marginal likelihood approximations as in Chib and Jeliazkov (2001). These approaches will be pursued elsewhere.

Although the focus of this paper is not on forecasting, this is another issue that should be explored in the future. In this paper we deal with multivariate time series for which the main goal is to infer key features of the latent processes over a given period of time, however, in many time series applications forecasting is also relevant. Possibly, the simplest way to do forecasting would be to condition on a particular set of parameters estimated, for example, in a Bayesian fashion, and then, sample a future observation q steps ahead, say $\mathbf{y}_{T+q}$, from $f_{T+q}(\mathbf{y}_{T+q}|\mathcal{F}_{T+q-1}, \mathcal{X}_{T+q}; \boldsymbol{\theta})$ in two steps. First, a pair of indexes (o, m) would be sampled with probabilities given by the mixture weights, $g_{T+q}(o, m|\mathcal{F}_{T+q-1}, \mathcal{X}_{T+q}; \boldsymbol{\gamma})$. Then, conditional on (o, m), $\mathbf{y}_{T+k}$ could be sampled from the basic density $\pi_{T+q}(\mathbf{y}_{T+q}|\mathcal{F}_{T+q-1}, \mathcal{X}_{T+q}; \boldsymbol{\eta})$.

Finally, although the VAR-HME models described in Section 2 are very general, we did not explore any examples in which the gating functions depend on other relevant covariates that are not exclusively time. For example, an interesting extension in the 7-channel EEG application would be to consider models in which the gating functions include the channel location as a covariate. These type of models will be considered in the future.

Bibliography

Chib, S. and Jeliazkov, I. (2001) Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96,** 270–281.

Elerian, O., Chib, S. and Shephard, N. (2001) Likelihood inference for discretely observed non-linear difussions. *Econometrica*, **69,** 959–963.

Green, P.J. (1995) Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82,** 711–32.

Huerta, G., Jiang, W. and Tanner, M.A. (2001) Discussion article: Mixtures of time series models. *Journal of Computational and Graphical Statistics*, **10,** 82–89.

Huerta, G., Jiang, W. and Tanner, M.A. (2003) Time series modeling via hierarchical mixtures. *Statistica Sinica*, **4,** 1097–1118.

Huerta, G. and Prado, R. (2003) Structure priors for multivariate time series. Technical Report. , AMS, UC Santa Cruz and Department of Mathematics and Statistics, University of New Mexico. *To appear in JSPI.*

Jordan, M.I. and Jacobs, R.A. (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **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.P. (2001) Stochastic neural networks with applications to nonlinear time series. *Journal of the American Statistical Association*, **96,** 968–981.

Prado, R. (1998) Latent structure in non-stationary time series. Ph.D. Thesis. Duke University, Durham, NC.

Prado, R. and Huerta, G. (2002) Time-varying autoregressions with model order uncertainty. *Journal of Time Series Analysis*, **23,** 599–618.

Prado, R., West, M. and Krystal, A.D. (2001) Multi-channel EEG analyses via dynamic regression models with time-varying lag/lead structure. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **50,** 95–109.

Rosenblatt, M. (1952) Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, **23(3),** 470–472.

Schwarz, G. (1978) Estimating the dimension of the model. *Annals of Statistics*, **6,** 461–464.

Tanner, M.A. (1996) *Tools for statistical inference.* Springer-Verlag.

Villagran, A. and Huerta, G. (2004) Bayesian inference on mixture-of-experts for estimation of stochastic volatility. Technical Report. Department of Mathematics and Statistics, University of New Mexico.

West, M., Prado, R. and Krystal, A.D. (1999) Evaluation and comparison of EEG traces: Latent structure in nonstationary time series. *Journal of the American Statistical Association*, **94,** 1083–1095.

Wong, C.S. and Li, W.K. (2000) On a mixture of autoregressive models. *J. R. Statist. Soc.* B, **62,** 95–115.

Wong, C.S. and Li, W.K. (2001) On a logistic mixture autoregressive model. *Biometrika*, **88,** 833–846.