Bayesian Inference on Mixture-of-Experts for Estimation of Stochastic Volatility

Alejandro Villagran and Gabriel Huerta [1]
Department of Mathematics and Statistics
University of New Mexico

Abstract

The problem of model mixing in time series, for which the interest lies in the estimation of stochastic volatility, is addressed using the approach known as Mixture-of-Experts (ME). Specifically, this work proposes a ME model where the *experts* are defined through ARCH, GARCH and EGARCH structures. Estimates of the predictive distribution of volatilities are obtained using a full Bayesian approach. The methodology is illustrated with an analysis of a section of US dollar/German mark exchange rates and a study of the Mexican stock market (IPC) index using the Dow Jones Industrial (DJI) index as a covariate.

Keywords: Mixture, stochastic volatility, covariates, ARCH/GARCH/EGARCH, MCMC.

JEL code: C1, C2, E4

[1]Corresponding author: Gabriel Huerta, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131; 505-821-9086; fax: 505-277-5505; ghuerta@stat.unm.edu

1 Introduction

In options trading and in foreign exchange rate markets, the estimation of volatility plays an important role in monitoring radical changes over time of key financial indexes. From a statistical standpoint, volatility refers to the variance of the underlying asset or return,conditional to all the previous information available until a specific time point.

It is well known that the volatility of a financial series tends to change over time and there are different types of models to estimate it: continuous-time or discrete-time processes. This paper discusses a model that falls in the second category. Discrete-time models are divided into those for which the volatility is ruled by a deterministic equation and those where the volatility has a stochastic behavior. Among the former, we have the autoregressive conditional heteroscedastic (ARCH) model introduced by Engle (1982) which provided a breakthrough in the modeling and estimation of time-varying conditional variance. Extensions to this seminal model are the generalized ARCH (GARCH) of Bollerslev (1986), the exponential GARCH (EGARCH) of Nelson (1991), the Integrated GARCH (IGARCH) and the GARCH in mean (GARCH-M). Additionally, models like the stochastic volatility (SV) of Melino and Turnbull (1990) give a discrete-time approximation for a continuous diffusion processes used in option pricing.

These variety of models gives a portfolio of options to represent volatility, but no agreement to decide which is the best approach. Given this difficulty, a new source of modeling arised: *mixture-of-models*. Since the only agreement seems to be that no real process can be completely explained by one model, the idea of model mixing, to combine different approaches into a unique representation, is very interesting. There are many types of methods of mixture models to estimate volatility. For instance, Wong and Li (2001) proposed a mixture of autoregressive conditional heteroscedastic models with an autoregressive component to model the mean (MAR-ARCH) and for which they use the EM algorithm to produce point estimation of the volatility. Another approach is given by Tsay (2002), who considered a mixture of ARCH and GARCH models by Markov Switching. Vrontos et. al. (2000) used *reversible jump* Markov chain Monte Carlo (MCMC) methods to predict a future volatility via model averaging of GARCH and EGARCH models. Both of these papers only considered a mixture of two models. Furthermore, Huerta et al.(2003) discussed the neural networks approach known as Hierarchical Mixture-of-Experts (HME) which is a very general and flexible approach of model mixing since it incorporates additional exogenous information, in the form of covariates or simply time, through the weights of the mixture. The example shown in that paper makes comparisons between a *difference-stationary* and a *trend-stationary* model using time as the only covariate. Additionally, Huerta et al.(2001), considers an HME model including AR, ARCH and EGARCH models and obtained point estimates of volatility using the EM algorithm. However, that paper does not report interval

estimates via a full Bayesian approach.

In this paper, we use *Mixture-of-Experts* (ME), which is a particular case of HME, to build a mixture of ARCH, GARCH and EGARCH models. Through a full Bayesian approach based on MCMC methods, we show how to estimate the posterior distribution of the parameters and the posterior predictive distribution of the volatility which is a very complicated function of the mixture representation. Additionally, we show how to obtain point estimates of the volatility, but also the usually unavailable forecast intervals. The paper proceeds in the following way. In section 2, we offer an introduction to ME and HME in the context of time series modeling. In section 3, we give the details of our volatility Mixture-of-Experts model along with the MCMC specifications to implement a full Bayesian approach to the problem of estimating volatility. Section 4 illustrates our methodology in the context of two financial applications and Section 5 provides some conclusions and extensions.

2 Mixture Modeling

Hierarchical Mixture of Experts (HME) was first introduced in the seminal paper by Jordan and Jacobs (1994) and it is based on mixing models to construct a *neural network* the using logistic distribution. This approach allows for model comparisons and a representation of the mixture weights as a function of time or other covariates. Additionally, the elements of the mixture, also known as experts, are not restricted to a particular parametric family which allows for very general model comparisons.

The model considers a response time series $\{y_t\}$ and a time series of covariates or exogenous variables $\{x_t\}$. Let $f_t(y_t|\mathcal{F}_{t-1},\chi;\theta)$ be the probability density function (pdf) of y_t conditional on θ, a vector of parameters, χ the σ-algebra generated by the exogenous information $\{x_t\}_0^n$, and for each t, $\mathcal{F}_{t-1}$ is the σ-algebra generated by $\{y_s\}_0^{t-1}$, the previous history about the response variable up to time $t-1$. Usually, it is assumed that this conditional pdf only depends on χ through x_t.

In the methodology of HME the pdf f_t is assumed to be a mixture of conditional pdfs of simpler models (Peng et. al., 1996). In the context of time series, the mixture could be represented by a finite sum

$$f_t(y_t|\mathcal{F}_{t-1},\chi;\theta) = \sum_J g_t(J|\mathcal{F}_{t-1},\chi;\gamma)\pi_t(y_t|\mathcal{F}_{t-1},\chi,J;\eta),$$

where the functions $g_t(\cdot|\cdot,\cdot;\gamma)$ are the mixtures weights; $\pi_t(\cdot|\cdot,\cdot,J;\eta)$ are the pdfs of simpler models defined by the label J; γ and η are sub-vectors of the parameter vector θ.

The models that are being mixed in HME are commonly denoted as *experts*. For example, in time series, one expert could be an AR(1) model, another expert could be

a GARCH(2,2) model. Also, the experts could be models that belong to the same class but with different orders or number of parameters. For example, all the experts are AR model but with different orders and different values of the *lag* coefficients. The extra hierarchy in HME partitions the space of covariates into O "overlays". In each overlay we have M competing models so that the most appropriate model will be assigned a higher weight.

For this hierarchical mixture, the expert index J, could be expressed as $J = (o, m)$, where the overlay index o takes a value in the set $\{1,\ldots,O\}$ and the model type index m takes a value in $\{1,\ldots,M\}$. The mixture model can be rewritten as

$$f_t(y_t|\mathcal{F}_{t-1}, \chi; \theta) = \sum_{o=1}^{O} \sum_{m=1}^{M} g_t(o, m|\mathcal{F}_{t-1}, \chi; \gamma)\pi_t(y_t|\mathcal{F}_{t-1}, \chi, o, m; \eta).$$

Within the neural network terminology, the mixture weights are known as *gating functions*. In difference to other approaches these weights have a particular parametric form that may depend on the previous history, exogenous information or exclusively on time. This makes the weights evolve across time in a very flexible way.

Specifically, it is proposed that the mixture weights have the form,

$$g_t(o, m|\mathcal{F}_{t-1}, \chi; \gamma) = \left\{ \frac{e^{v_o + u_o^T W_t}}{\sum_{s=1}^{O} e^{v_s + u_s^T W_t}} \right\} \left\{ \frac{e^{v_{m|o} + u_{m|o}^T W_t}}{\sum_{l=1}^{M} e^{v_{l|o} + u_{l|o}^T W_t}} \right\}$$

where the v's and u's are parameters which are components of γ; W_t is an input at time t, which is measurable with respect to the σ-algebra induced by $\mathcal{F}_{t-1} \cup \chi$. In this case, γ includes the following components: v_1,u_1,...,v_{O-1},u_{O-1},$v_{1|1}$,$u_{1|1}$,...,$v_{M-1|1}$,$u_{M-1|1}$, ...,$v_{M-1|O}$,$u_{M-1|O}$. For identifiability of the mixture weights , we set $v_O = u_O = v_{M|o} = u_{M|o} = 0$ for all $o = 1,\ldots,O$. This restriction guarantees that the gating functions are uniquely identified by γ as shown in Huerta et al.(2003). Both terms that define the mixture weights follow a multinomial logistic pdf where the first term describes the probability of a given overlay and the second term, the probability of a model within overlay. Each of these probabilities being a function of the input W_t. Mixtures of time series models for estimating volatility has also been considered in Wong and Li (2001). However, these authors only look at the problem from a point estimation perspective.

Inferences on the parameter vector θ can be based in the log-likelihood function

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

To obtain the maximum likelihood estimator of θ, $\hat{\theta} = arg\ max \mathcal{L}_n(\cdot)$, it is possible to use the Expectation Maximization or EM algorithm as described in Huerta et al.(2003). A general presentation of the EM algorithm appears in Tanner (1996).

After the MLE, $\hat{\theta}$, is obtained, the interest focuses in the evaluation of the weights assigned to each of the M models as a function of time t. Primarily, there are two ways to achieve this, the first one is via the conditional probability of each model m defined by

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

where conditional refers to the actual observation at time t, y_t.

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

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

Point estimation of each probability can be obtained by evaluation at $\hat{\theta}$ or by computing the expected value with respect to the posterior distribution $\pi(\theta|\mathcal{F}_n, \chi)$.

The particular case of HME that we consider in this paper is $O = 1$ which is known as Mixture of Experts (ME). In the ME modeling it is assumed that the process that generates the response variable can be decomposed into a set of subprocesses defined over specific regions of the space of covariates. For each value of the covariate x_t, a 'label' r is chosen with probability $g_t(r|\chi,\mathcal{F}_{t-1},\gamma)$. Given this value of r, the response y_t is generated from the conditional pdf $\pi_t(y_t|r, \chi,\mathcal{F}_{t-1},\eta)$. The pdf of y_t conditional on the parameters, the covariate and the response history is given by

$$f(y_t|\chi, \mathcal{F}_{t-1}, \theta) = \sum_{r=1}^{M} g_t(r|\chi, \mathcal{F}_{t-1}, \gamma)\pi_t(y_t|r, \chi, \mathcal{F}_{t-1}, \eta).$$

and the likelihood function is

$$L(\theta|\chi) = \prod_{t=1}^{n} \sum_{r=1}^{M} g_t(r|\chi, \mathcal{F}_{t-1}, \gamma)\pi_t(y_t|r, \chi, \mathcal{F}_{t-1}, \eta).$$

As with HME, the mixture probabilities associated to each value of r are defined through a logistic function

$$g_t(r|\chi, \mathcal{F}_{t-1}, \eta) \equiv g_r^{(t)} = \frac{e^{\xi_r}}{\sum_{h=1}^{M} e^{\xi_h}},$$

where $\xi_r = v_r + u_r^T W_t$, W_t is an input that could be a function of time, history and $\{x_t\}$. For identifiability, ξ_M is set equal to zero.

Inference about the parameters in the ME is simplified by augmenting the data with non-observable indicator variables which determine the type of model expert. For each time t , $z_r^{(t)}$ is a binary variable such that $z_r^{(t)} = 1$ with probability

$$h_r^{(t)} = \frac{g_r^{(t)}\pi_t(y_t|r, \chi, \mathcal{F}_{t-1}, \eta)}{\sum_{r=1}^{M} g_r^{(t)}\pi_t(y_t|r, \chi, \mathcal{F}_{t-1}, \eta)}.$$

If χ'= $\{(x_t, z^{(t)})\}_{t=1}^n$, where $z^{(t)}$ is the vector that includes all the indicator variables, the *augmented likelihood* for the ME model is

$$L(\theta|\chi') = \prod_{t=1}^{n}\prod_{r=1}^{M}\{g_r^{(t)}\pi_t(y_t|r,\chi,\mathcal{F}_{t-1},\gamma)\}^{z_r^{(t)}}.$$

In the following section, we discuss how to estimate a ME model by a Bayesian approach and with the experts being ARCH, GARCH and EGARCH models. We picked these experts as our model building blocks since these are the main conditional heteroscedasticity models used in practice as pointed out by the seminal papers of Engle (1982), Engle (1995), Bollerslev (1986) and Nelson (1991). Also, Tsay (2002) and Vrontos, e. al. (2002) mention that these models are interesting in practice due to their parsimony.

3 Bayesian Inference on ME for Volatility

If the Bayesian paradigm is adopted, the inferences about θ are based on the posterior distribution $\pi(\theta|\underline{y})$. Bayes Theorem establishes that

$$\pi(\theta|\underline{y}) = \frac{f(\underline{y}|\theta)\pi(\theta)}{\int_\Theta f(\underline{y}|\theta)dF^\pi(\theta)}$$

which defines the way to obtain the posterior distribution of θ through the prior $\pi(\theta)$ and the likelihood function $f(\underline{y}|\theta)$. However, for a ME or HME approach the marginal distribution of $\underline{y}$, $\int_\Theta f(\underline{y}|\theta)dF^\pi(\theta)$ cannot be obtained analytically. We overcome this difficulty by using MCMC methods to simulate samples from $\pi(\theta|\underline{y})$. For more details about MCMC methods see Tanner (1996).

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

$$\pi(\theta) = \pi(\eta)\pi(\gamma),$$

so the expert parameter η and the weights or *gating* parameters γ are apriori independent. We define $\mathbf{Z} = \{\mathbf{z^{(t)}};\ t = 1,\ldots,n\}$ and for each t, $\mathbf{z^{(t)}} = \{z_r^{(t)};\ r = 1,\ldots,M\}$ is the set of indicator variables at t. Conditional on θ, $P(\mathbf{z^{(t)}}|\theta,\chi)$ is a Multinomial distribution with total count equal to 1 and cell probabilities $g_r(t,\gamma)$.

Our MCMC scheme is based on the fact that it is easier to obtain samples from the *augmented* posterior distribution $\pi(\theta,\mathbf{Z}|\mathcal{F}_n,\chi)$, instead of directly simulating values from $\pi(\theta|\mathcal{F}_n,\chi)$. This data augmentation principle was introduced by Tanner and Wong (1987). The MCMC scheme follows a Gibbs sampling format for which we iteratively simulate from the conditional distributions $\pi(\theta|\mathbf{Z},\mathcal{F}_n,\chi)$ and $\pi(\mathbf{Z}|\theta,\mathcal{F}_n,\chi)$.

The conditional posterior $\pi(\mathbf{Z}|\theta,\mathcal{F}_n,\chi)$ is sampled through the marginal conditional posteriors $\pi(\mathbf{z^{(t)}}|\theta,\mathcal{F}_n,\chi)$ defined for each value of t. Given θ, $\mathcal{F}_n$ and χ, it can be shown that

the vector $\mathbf{z^{(t)}}$ has a Multinomial distribution with total count equal to 1 and for which

$$P(z_r^{(t)} = 1|\theta, \mathcal{F}_n, \chi) = h_r(t;\theta) = \frac{g_r^{(t)}\pi_t(y_t|r, \mathcal{F}_{t-1}, \chi; \eta)}{\sum_{r=1}^{M} g_r^{(t)}\pi_t(y_t|r, \mathcal{F}_{t-1}, \chi; \eta)}.$$

The vector $\theta = (\eta,\gamma)$ is sampled in two stages. Firstly, η is simulated from the conditional posterior distribution $\pi(\eta|\gamma,\mathbf{Z},\mathcal{F}_n,\chi)$ and then γ is sampled from the conditional posterior $\pi(\gamma|\eta,\mathbf{Z},\mathcal{F}_n,\chi)$. By Bayes Theorem,

$$\pi(\eta|\gamma, \mathbf{Z}, \mathcal{F}_n, \chi) \propto \prod_{t=1}^{n}\prod_{r=1}^{M} f_t(y_t|\mathcal{F}_{t-1}, \chi, r; \eta)^{z_r^{(t)}}\pi(\eta).$$

Analogously,

$$\pi(\gamma|\eta, \mathbf{Z}, \mathcal{F}_n, \chi) \propto \prod_{t=1}^{n}\prod_{r=1}^{M} g_t(r|\mathcal{F}_{t-1}, \chi; \gamma)^{z_r^{(t)}}\pi(\gamma).$$

If η can be decomposed into a sub-collection of parameters η_r that are assumed apriori independent, the simulation for the full conditional for η is reduced to individual simulation of each η_r. If η_r is assigned a conjugate prior with respect to the pdf of the "r" expert, the simulation of η is straightforward. For γ, it is necessary to implement Metropolis-Hastings steps to obtain a sample from its full conditional distribution.

The specific details for the MCMC implementation depends on the type of expert models and prior distributions on model parameters. For example, Huerta et al.(2003) discussed a HME model with a full Bayesian approach where the experts are a 'difference-stationary' and a 'trend-stationary' model. The priors used on the parameters of their HME model were non-informative. Here, we consider the case of experts that allow volatility modeling.

It is well known that the volatility of a financial time series can be represented by ARCH, GARCH and EGARCH models. The properties of these models make them attractive to obtain forecasts in financial applications. We propose a ME that combines the models AR(1)-ARCH(2), AR(1)-GARCH(1,1) and AR(1)-EGARCH(1,1). Although the order of the autoregressions for the observations and volatility is low for these models, in practice it is usually not necessary to consider higher order models.

The elements of our ME model are, the time series of returns $\{y_t\}_1^n$, the series of covariates $\{x_t\}_1^n$, which in one of our applications presented in the next section it is simply time and in the other, it is the Dow Jones index. In any case, $\xi_r = v_r + u_r^T W_t$, where W_t is an input that depends on the covariates.

Our expert models will be parameterized in the following way,

AR(1)-ARCH(2)

$$y_t = \phi_1 y_{t-1} + \epsilon_{1,t} \qquad \epsilon_{1,t} \sim N(0, \sigma_{1,t}^2)$$
$$\sigma_{1,t}^2 = \omega_1 + \alpha_{11}\epsilon_{1,t-1}^2 + \alpha_{12}\epsilon_{1,t-2}^2$$

AR(1)-GARCH(1,1)

$$y_t = \phi_2 y_{t-1} + \epsilon_{2,t} \qquad \epsilon_{2,t} \sim N(0, \sigma^2_{2,t})$$
$$\sigma^2_{2,t} = \omega_2 + \alpha_{21}\epsilon^2_{2,t-1} + \alpha_{22}\sigma^2_{2,t-2}$$

AR(1)-EGARCH(1,1)

$$y_t = \phi_3 y_{t-1} + \epsilon_{3,t} \qquad \epsilon_{3,t} \sim N(0, \sigma^2_{3,t})$$
$$ln(\sigma^2_{3,t}) = \omega_3 + \alpha_{31} ln(\sigma^2_{3,t-1}) + \alpha_{32}\epsilon_{3,t-1} + \alpha_{33}(|\epsilon_{3,t-1}| - E(|\epsilon_{3,t-1}|))$$

For each expert $m = 1, 2, 3$, the ME will be represented by the following pdfs and gating functions,

Expert 1

$$g_t(1|\mathcal{F}_{t-1}, \chi; \gamma) = \frac{exp\{\xi_1\}}{\sum_{r=1}^{3} exp\{\xi_r\}} = \frac{exp\{v_1 + u_1 W_t\}}{\sum_{r=1}^{3} exp\{v_r + u_r W_t\}}$$

$$\pi_t(y_t|1, \mathcal{F}_{t-1}, \chi; \eta_1) = \frac{1}{\sqrt{2\pi\sigma^2_{1,t}}} exp\{-\frac{1}{2\sigma^2_{1,t}}(y_t - \phi_1 y_{t-1})^2\}$$

Expert 2

$$g_t(2|\mathcal{F}_{t-1}, \chi; \gamma) = \frac{exp\{\xi_2\}}{\sum_{r=1}^{3} exp\{\xi_r\}} = \frac{exp\{v_2 + u_2 W_t\}}{\sum_{r=1}^{3} exp\{v_r + u_r W_t\}}$$

$$\pi_t(y_t|2, \mathcal{F}_{t-1}, \chi; \eta_2) = \frac{1}{\sqrt{2\pi\sigma^2_{2,t}}} exp\{-\frac{1}{2\sigma^2_{2,t}}(y_t - \phi_2 y_{t-1})^2\}$$

Expert 3

$$g_t(3|\mathcal{F}_{t-1}, \chi; \gamma) = \frac{exp\{\xi_3\}}{\sum_{r=1}^{3} exp\{\xi_r\}} = \frac{1}{\sum_{r=1}^{3} exp\{v_r + u_r W_t\}}$$

$$\pi_t(y_t|3, \mathcal{F}_{t-1}, \chi; \eta_3) = \frac{1}{\sqrt{2\pi\sigma^2_{3,t}}} exp\{-\frac{1}{2\sigma^2_{3,t}}(y_t - \phi_3 y_{t-1})^2\}$$

As mentioned before, the vector $\theta = (\eta,\gamma)$ can be decomposed into two subsets, one that includes the expert parameters, $\eta = (\alpha_{11},\alpha_{12},\alpha_{21},\alpha_{22},\alpha_{31},\alpha_{32},\alpha_{33},\omega_1,\omega_2,\omega_3,\phi_1,\phi_2,\phi_3)$ and another subvector that includes the gating function parameters, $\gamma = (u_1,u_2,v_1,v_2)$.

The likelihood function of the model is expressed as,

$$\mathcal{L}(\theta|\chi) = \prod_{t=1}^{n} \sum_{r=1}^{3} g_r^{(t)} exp\{-\frac{1}{2\sigma^2_{r,t}}(y_t - \phi_r y_{t-1})^2\},$$

and the augmented likelihood function is,

$$\mathcal{L}(\theta|\chi') = \prod_{t=1}^{n} \prod_{r=1}^{3} \left\{ g_r^{(t)} exp\{-\frac{1}{2\sigma^2_{r,t}}(y_t - \phi_r y_{t-1})^2\} \right\}^{z_r^{(t)}}.$$

The MCMC scheme to obtain posterior samples for this ME is based on the principle of "divide and conquer". For *Expert 1*, we assume that $\eta_1 = (\phi_1,\omega_1,\alpha_{11},\alpha_{12})$ has prior distribution with components that are apriori independent,$\pi(\eta_1) = \pi(\phi_1)\pi(\omega_1)\pi(\alpha_{11})\pi(\alpha_{12})$ where $\phi_1 \sim N(0,0.1)$, $\omega_1 \sim U(0,\infty)$, $\alpha_{11} \sim U(0,1)$ and $\alpha_{12} \sim U(0,1)$.

For the *Expert 2*, $\eta_2 = (\phi_2,\omega_2,\alpha_{21},\alpha_{22})$ also has components that are apriori independent where, $\phi_2 \sim N(0,0.1)$, $\omega_2 \sim U(0,\infty)$, $\alpha_{21} \sim U(0,1)$ and $\alpha_{22} \sim U(0,1)$.

In an analogous way, for *Expert 3* the vector $\eta_3 = (\phi_3,\omega_3,\alpha_{31},\alpha_{32},\alpha_{33})$ has independent components with marginal prior distributions given by $\phi_3 \sim N(0,0.1)$, $\omega_3 \sim N(0,10)$, $\alpha_{31} \sim U(-1,1)$, $\alpha_{32} \sim N(0,10)$ and $\alpha_{33} \sim N(0,10)$.

Finally, each of the entries of the vector of parameters appearing in the mixture weights, $\gamma = (u_1,u_2,v_1,v_2)$, is assumed to have a $U(-l,l)$ prior distribution with a large value for l. This prior specification was chosen to reflect vague (flat) prior information and to facilitate the calculations of the different steps inside our MCMC scheme. The N(0,0.1) prior on the AR(1) coefficients was proposed with the idea of containing most of its mass in the region defined by the stationarity condition. Instead of a N(0,0.1) prior, we also used a $U(-1,1)$ prior on the coefficients and the results obtained were essentially the same as with the Normal prior. In fact, the non-informative priors were suggested by Vrontos, et al.(2000) in the context of parameter estimation, model selection and volatility prediction. Also, these authors show that under these priors and for pure GARCH/EGARCH models, the difference between classical and Bayesian point estimation is minimal. The restrictions on these priors in the different parameter spaces is to satisfy the stationarity conditions of the expert models. For our ME model, these type of non-informative priors were key to produce good MCMC convergence results which could not be obtain with other classes of (informative) priors. Since the parameters of ARCH/GARCH/EGARCH models are not very meaningful in practice, the most typical prior specification for these parameters is to adopt non-informative prior distributions. To our knowledge, there is no study on the effects of using informative priors in this context. Furthermore, Villagran (2003) discuses another aspect of our prior specification in terms of simulated data. If the true data follows an AR(1)-GARCH(1,1) structure, the non-informative prior allows to estimate the parameters of true model with a maximum absolute error of 0.003. The maximum posterior standard deviation for all the model parameters is 0.0202 and the posterior mean for $g_t^{(t)}$ is practically equal one for the true model (in this case GARCH) and for all time t.

Our MCMC algorithm can be summarized as follows:

- Assign initial values for θ and with these values calculate the volatilities for each expert, $\sigma_{1,t}^{2(0)}$, $\sigma_{2,t}^{2(0)}$ and $\sigma_{3,t}^{2(0)}$ for all t.
- Evaluate the probabilities $g_r^{(t)}$ at $\gamma^{(0)}$ and compute the conditional probabilities $h_r^{(t)}$ for all t.

- Generate $z_r^{(t)}$ conditional on θ, $\mathcal{F}_n$, χ from a Multinomial distribution with total count equal to 1 and cell probabilities $h_r^{(t)}$. Across this step, we are generating vectors $(z_1^{(t)},z_2^{(t)},z_3^{(t)})$ for all values of t.

- For the augmented posterior distribution, we generate each of the expert parameters and each of the mixture weights or gating function parameters via Metropolis-Hastings (M-H) steps. A general description of the M-H algorithm with several illustrative examples appears in Tanner (1996).

- At each M-H step, we propose a new value $\theta^{(j)}$ for the parameters from a candidate distribution and than accept or reject this new value with probability $\alpha(\theta^{(j-1)},\theta^{(j)}) = min\big[\frac{\pi(\theta^{(j)}|\cdot)q(\theta^{(j-1)})}{\pi(\theta^{(j-1)}|\cdot)q(\theta^{(j)})},1\big]$.

- After generating all the model parameters at iteration j, we update the volatilities $\sigma_{1,t}^{2(j)}$, $\sigma_{2,t}^{2(j)}$ and $\sigma_{3,t}^{2(j)}$, the probabilities $g_r^{(t)}$, $h_r^{(t)}$ and the indicator variables $\mathbf{Z}^{(j)}$.

- The algorithm is iterated until Markov Chain convergence is reached. An initial section of the iterations is considered a burn-in period and the remaining iterations are kept as posterior samples of the parameters.

Given that mixture models are highly multimodal, to improve on the convergence of our MCMC method, it is convenient to propose several starting points for θ and run the algorithm for a few iterations. The value that produces the maximum posterior density is used as the initial point $\theta^{(0)}$ to produce longer runs of the Markov chain. In the applications that are presented in the next section, we used 20 overdispersed starting values for θ and calibrated our proposal distribution so that the acceptance rates of all the M-H steps were around 45%. We mantained this acceptance rates relatively low to allow full exploration of the parameter space and to avoid getting stuck around a local mode.

For a specific MCMC iteration, the volatility or conditional variance of our ME model is computed as

$$V_t^{(j)} = \sum_{r=1}^{3} g_{r,t}^{(j)} \sigma_{r,t}^{2(j)} + \sum_{r=1}^{3} g_{r,t}^{(j)} (\mu_{r,t}^{(j)} - \overline{\mu}_t^{(j)})^2,$$

$$\overline{\mu}_t^{(j)} = \sum_{r=1}^{3} g_{r,t}^{(j)} \mu_{r,t}^{(j)} = \sum_{r=1}^{3} g_{r,t}^{(j)} \phi_r^{(j)} y_{t-1}.$$

where the index t represents time, the index j represents iteration and $\mu_{r,t}$ represent the mean of expert r at time t. These expressions follow from well known results to compute the variance of a mixture distribution with 3 components. Given a value of model parameters at iteration j, $V_t^{(j)}$ can be directly evaluated from these equations. The expression for the

volatility of our ME model is formed by two terms. The first term represents the dependency of the conditional variance with respect to past volatilities and the second term, represents changes in volatility of the mixture model due to the differences in conditional mean between experts.

In the next section, we show that using time as a covariate allows one to detect structural changes in volatility so ME is able to determine if the process generating the data corresponds to a unique expert.

4 Applications

4.1 Exchange rates US dollar/German Mark

[2] Figure 1 (a) shows 500 daily observations between the American dollar and the German mark starting on October of 1986 and Figure 1 (b) shows the corresponding returns of these exchange rates.

Figure 1 about here.

Using the returns as our response time series, we implemented the ME model with time being our covariate. Figure 2 shows posterior means and 95 % credible intervals for the unconditional probabilities of each expert, i.e., $g_r^{(t)}; r = 1, 2, 3$. Both $h_r^{(t)}$ and $g_r^{(t)}$ are functions of the unknown parameter vector θ, so it makes absolute sense to assess measures of uncertainty to these probabilities.

Figure 2 about here.

We can appreciate that the expert that dominates in terms of probability is the AR(1)-GARCH(1,1) model. Furthermore, Figure 2 also shows the relative uncertainty of the different experts across time. For the period covering October 86 to January 87, there is a significant weight associated to the AR(1)-EGARCH(1,1) model and the credible band for the weight of this model can go as high as 0.8 and as low as 0.2. In Figure 3 we report posterior means of the conditional probabilities of each model, $h_r^{(t)}; r = 1, 2, 3$.

Figure 3 about here.

This figure shows a similar pattern compared to the description of probabilities given

[2] The code to fit the models used for this section is available under request from *avhstat@unm.edu*. Also, this code can be downloaded from *http://www.stat.unm.edu/∼avhstat*

by Figure 2. Since these are conditional probabilities, individual observations may produce high fluctuations in probability. However, this figure confirms that the models that dominate in the ME, at least in the initial time periods, are the AR(1)-GARCH(1,1) and the AR(1)-EGARCH(1,1). Towards the end of the considered time periods, the dominant model is the AR(1)-GARCH(1,1) but the AR(1)-ARCH(2) model has a significant weight of 0.4.

In Figure 4 we present posterior mean estimators of the volatility for the ME model and for the individual expert models.

Figure 4 about here.

The ME model has a higher volatility estimate at the beginning of the time series. This is due to the influence of the EGARCH models in the first part of the series as shown by Figures 2 and 3. A referee and the editors suggested that we compared the square of the residuals of a pure AR(1) model fitted to the return series, with the volatilities presented in Figure 4. These residuals seem to be better charaterized by the AR(1)-ARCH(2) volatilities towards the end of the time period covered by the data. However at the beginning, the residuals are more closely followed by the AR(1)-EGARCH(1,1) volatilities. Additionally, we think that the ME is at least as good as any individual expert model since it is pooling information from different models. A great advantage of the ME model is that it shows, as a function of time t, how different expert models are competing with each other conditional on the information at $t-1$ and how the volatilities change according to time. Notice that from January 1987, the volatilities of the AR(1)-ARCH(2) are consistent with the volatilities of the ME model.

Figure 5 considers a "future" period starting from March 1988 and that covers 100 daily exchange rate values.

Figure 5 about here.

Figure 5 (a) shows the time series of returns for this future period and Figure 5 (b) shows the one-step-ahead predictive posterior means and the 95 % one-step-ahead forecast intervals for volatility based on the ME model that only uses previous data from October 86 to March 88 and with a forecasting horizon of 100 time steps. This Figure illustrates one of the main features of our model. The MCMC approach allows us to compute samples of future or past volatilities values that can be summarized in terms of predictive means and credible intervals. In our ME model, the volatility is a very complicated function of the parameters and producing non-Monte Carlo estimates, especially predictive intervals, is practically impossible.

4.2 Analysis of the Mexican stock Market

In this application, we studied the behavior of the Mexican stock market (IPC) index using as covariates the Dow Jones Industrial (DJI) index from January 2000 to September 2004 and also using time. Figure 6 shows both the IPC index and the DJI index time series with their corresponding returns.

Figure 6 about here.

It is obvious that the IPC index and the DJI index have the same overall pattern over time. In fact, some Mexican financial analysts accept that the IPC index responds to every 'strong' movement of the DJI index. Our Bayesian ME approach adds some support to this theory.

Figure 7 about here.

In Figure 7 we show the posterior distributions of the parameters for the mixture weights or gating functions when W_t was set equal to the DJI index. The posterior distribution for the 'slope' parameters u_1 and u_2 have most of their posterior mass away from 0, which means that the effect of the covariate in our ME analysis is 'highly significant'.

In Figure 8 we show the mixture weights of the ME using different covariates.

Figure 8 about here.

The left column presents the posterior mean estimates of $g_r(t); r = 1, 2, 3$ using the DJI index as covariate and right column shows the estimates as a function of time. The right column shows a shift on the regime since the AR(1)-EGARCH(1,1) expert rules the evolution of the volatility of the IPC index from January 2000 to March 2002. After this date, the AR(1)-ARCH(2) model is the one with higher probability. As a function of the DJI index, the mixture weights behavior is quite different. Now the AR(1)-EGARCH(1,1) expert rules all the time in comparison to the other experts. This is a result due to the common high volatility shared by the IPC index and the DJI index.

5 Conclusions and Extensions

In this paper we present a mixture modeling approach based on the Bayesian paradigm and with the goal of estimating stochastic volatility. We illustrate the differences of our mixture methodology versus a sole model approach in the context of ARCH/GARCH/EGARCH

models for two different financial series. The two main aspects of our ME model are: (1) the comparison of different volatility models as a function of covariates and (2) the estimation of predictive volatilities with their corresponding measure of uncertainty given by a credible interval. On the other hand, we had only considered ME and not HME. The difficulty with HME is that it requires the estimation of the number of overlays O which poses challenging computational problems in the form of reversible jump MCMC methods. Additionally, we had not considered any other experts beyond ARCH, GARCH and EGARCH models. An extension to our approach considers other competitors or experts like the stochastic volatility models of Jacquier, et al.(1994). This leads into MCMC algorithms combining mixture modeling approaches with *Forward Filtering Backward simulation*. These extensions are part of future research.

6 Acknowledgments

We wish to express our thanks to Professors Thomas Fomby and Carter Hill, editors of this volume, for all their considerations about this paper. During the preparation of this paper, A. Villagran was partially supported by CONACyT-Mexico grant 159764 and by The University of New Mexico.

References

[1] Bollerslev, T.,1986. Generalized Autoregressive Conditional Heteroskedasticity. Journal of Econometrics 31, 307–327.

[2] Engle, R.F.,1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica 50, 987-1007.

[3] Engle, R.F., 1995. ARCH Selected Readings. Oxford University Press: New York.

[4] Huerta, G., Jiang, W. and Tanner, M.A.,2003. Time Series Modeling via Hierarchical Mixtures. Statistica Sinica 13, 1097-1118.

[5] Huerta, G., Jiang, W. and Tanner, M.A.,2001. Mixtures of Time Series Models. Journal of Computational and Graphical Statistics 10, 82-89.

[6] Jacquier, E., Polson N. and Rossi, P.,1994. Bayesian Analysis of Stochastic Volatility Models. Journal of Business & Economic Statistics 12, 371-389.

[7] Jordan, M. and Jacobs, R.,1994. Hierarchical Mixture of Experts and the EM Algorithm. Neural Computation 6, 181-214.

[8] Melino, A. and Turnbull, S.M.,1990. Pricing foreign currency options with stochastic volatility. Journal of Econometric 45, 239-265.

[9] Nelson, D.B.,1991. Conditional Heteroskedasticity in Asset Returns. Econometrica 59, 347-370.

[10] Peng, F., Jacobs, R.A. and Tanner, M.A.,1996. Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With an Application to Speech Recognition. Journal of the American Statistical Association 91, 953-960.

[11] Tanner, M.A.,1996. Tools for Statistical Inference. Springer-Verlag, Third Edition, New York.

[12] Tanner, M.A. and Wong, W.H.,1987. The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association 82, 528-550.

[13] Tsay, R.S.,2002. Analysis of Financial Time Series. John Wiley & Sons, New York.

[14] Villagran, A., 2003. Modelos Mezcla para Volatilidad. Unpublished MSc. Thesis. Universidad de Guanajuato, Mexico.

[15] Vrontos, D., Dellaportas, P. and Politis, D.N.,2000. Full Bayesian Inference for GARCH and EGARCH Models. Journal of Business & Economic Statistics 18, 187-197.

[16] Wong, Ch.S. and Li, W.K.,2001. On a Mixture Autoregressive Conditional Heteroscedastic Model. Journal of the American Statistical Association 96, 982-995.

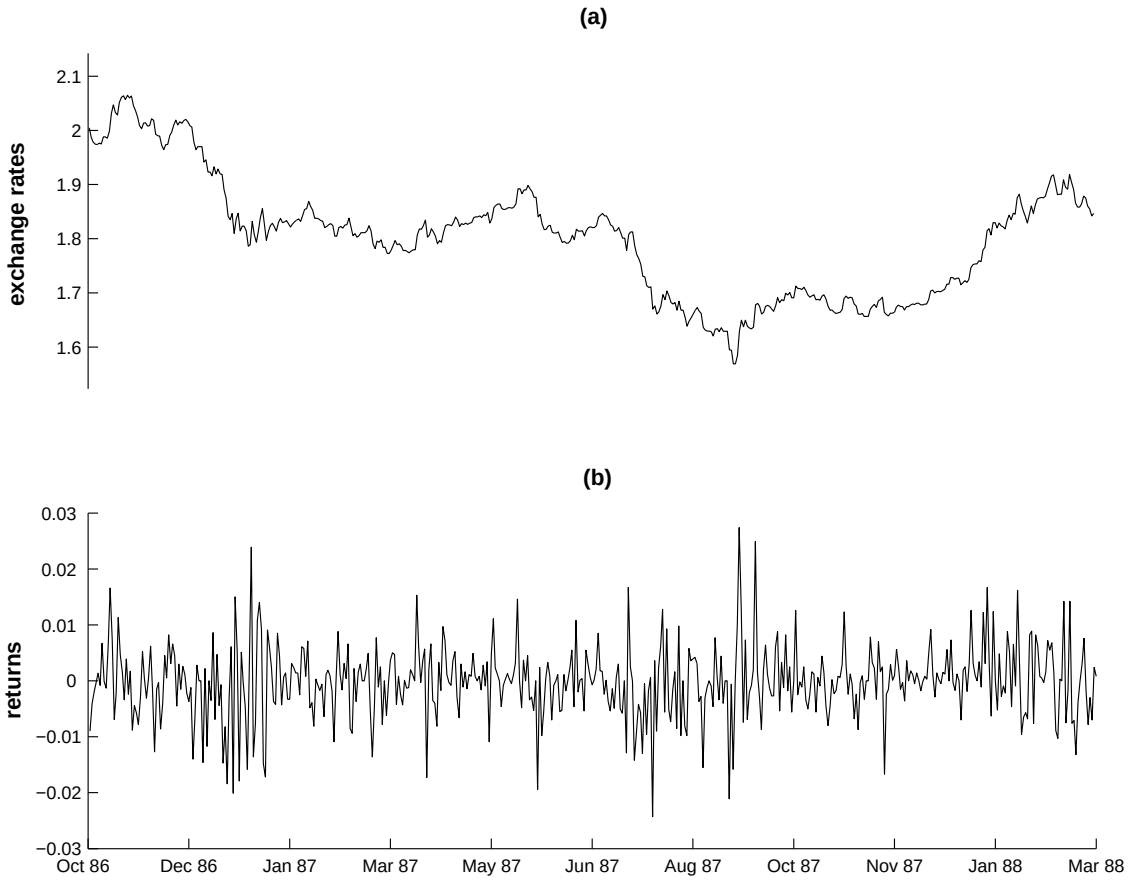

Figure 1: (a) Exchange rates between U.S. dollar and German Mark starting from October 1986. (b) Returns of the exchange rates.

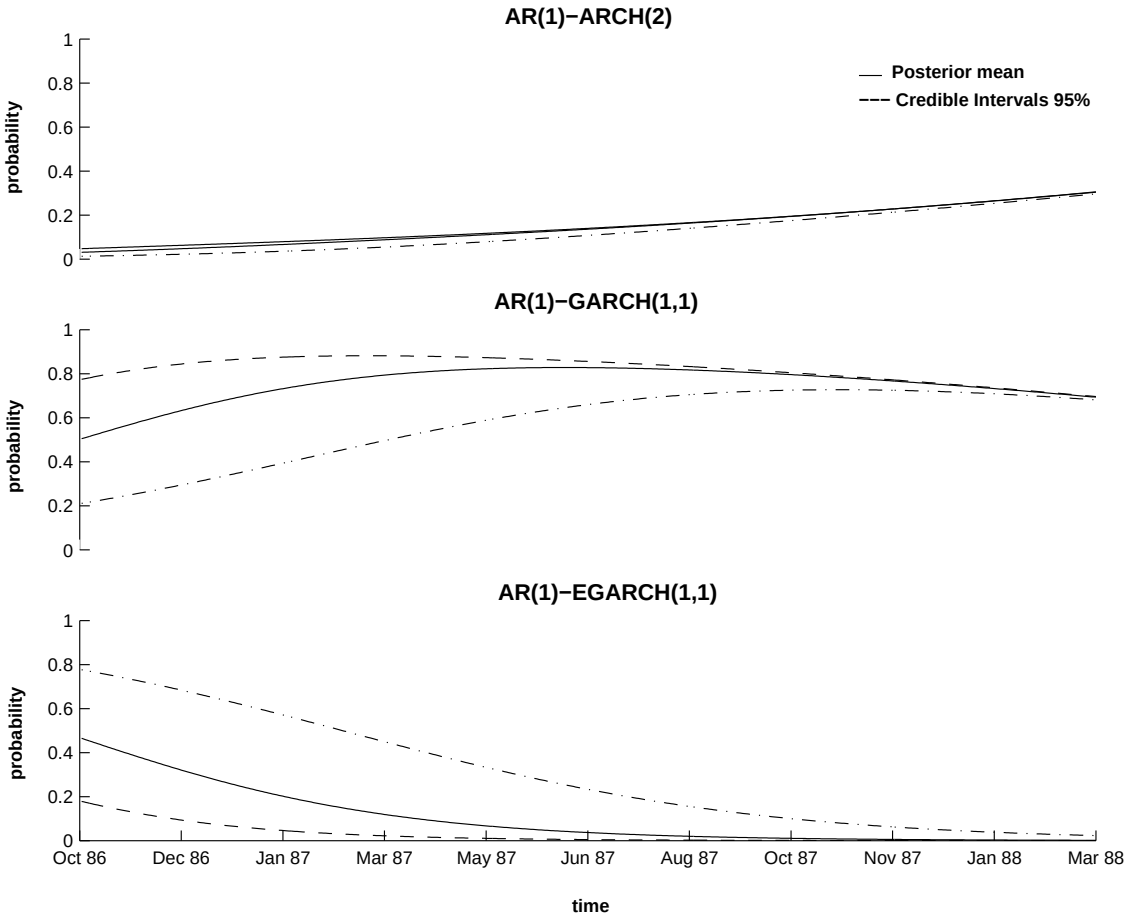

Figure 2: Exchange rates example. Posterior means of $g_r^{(t)}; r = 1, 2, 3$ (solid lines) and 95% credible intervals (dashed lines).

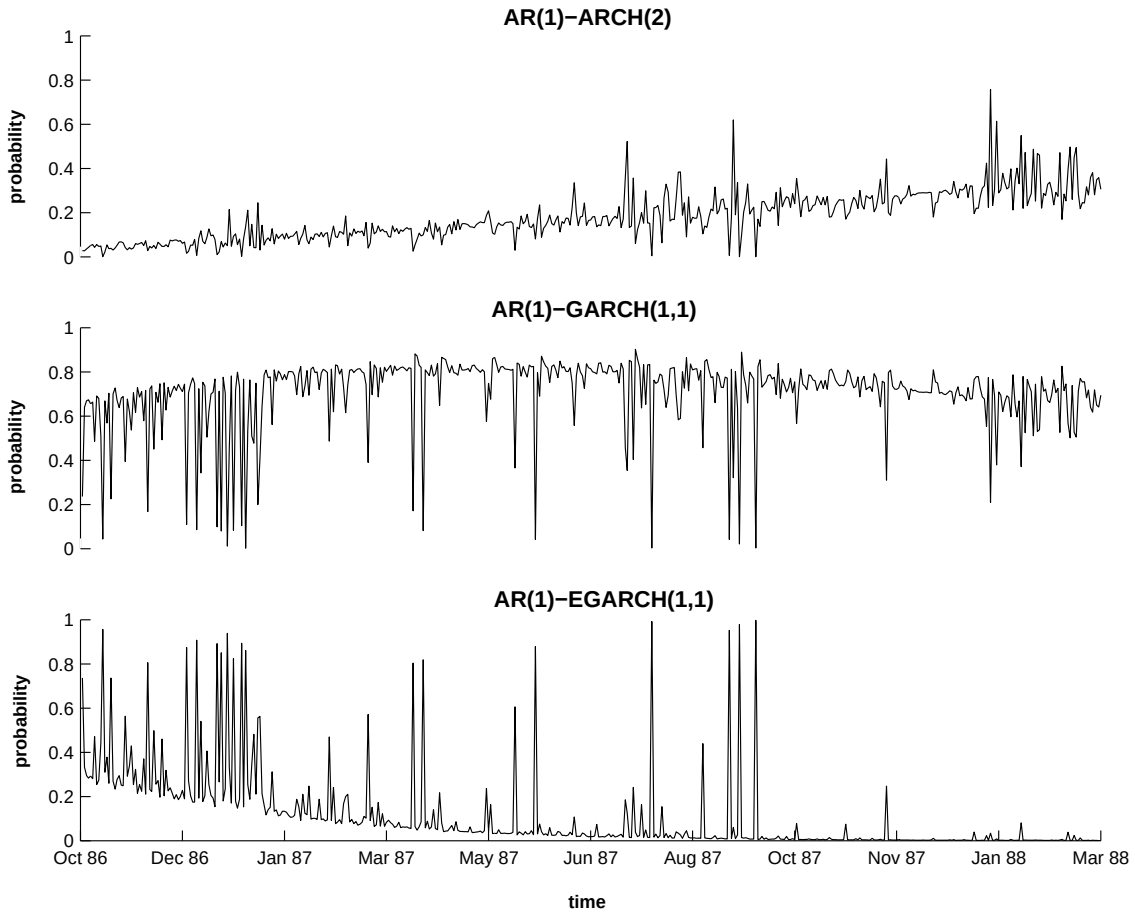

Figure 3: Exchange rates example. Posterior means of $h_r^{(t)}; r = 1, 2, 3$

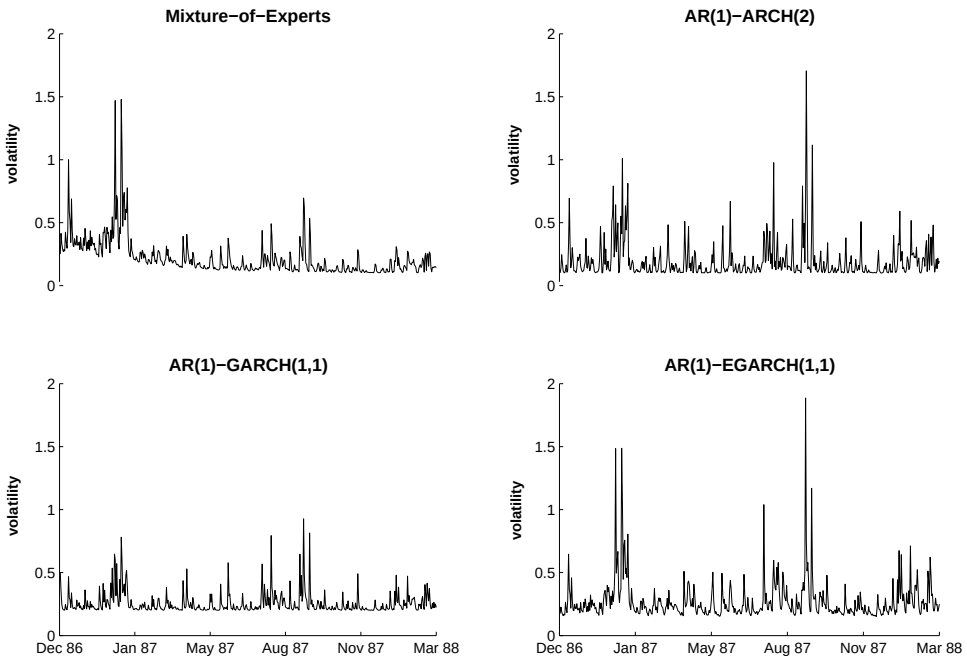

Figure 4: Exchange rates example. Posterior mean estimate of volatility for mixture-of-experts model and posterior estimates of volatility for individual models.

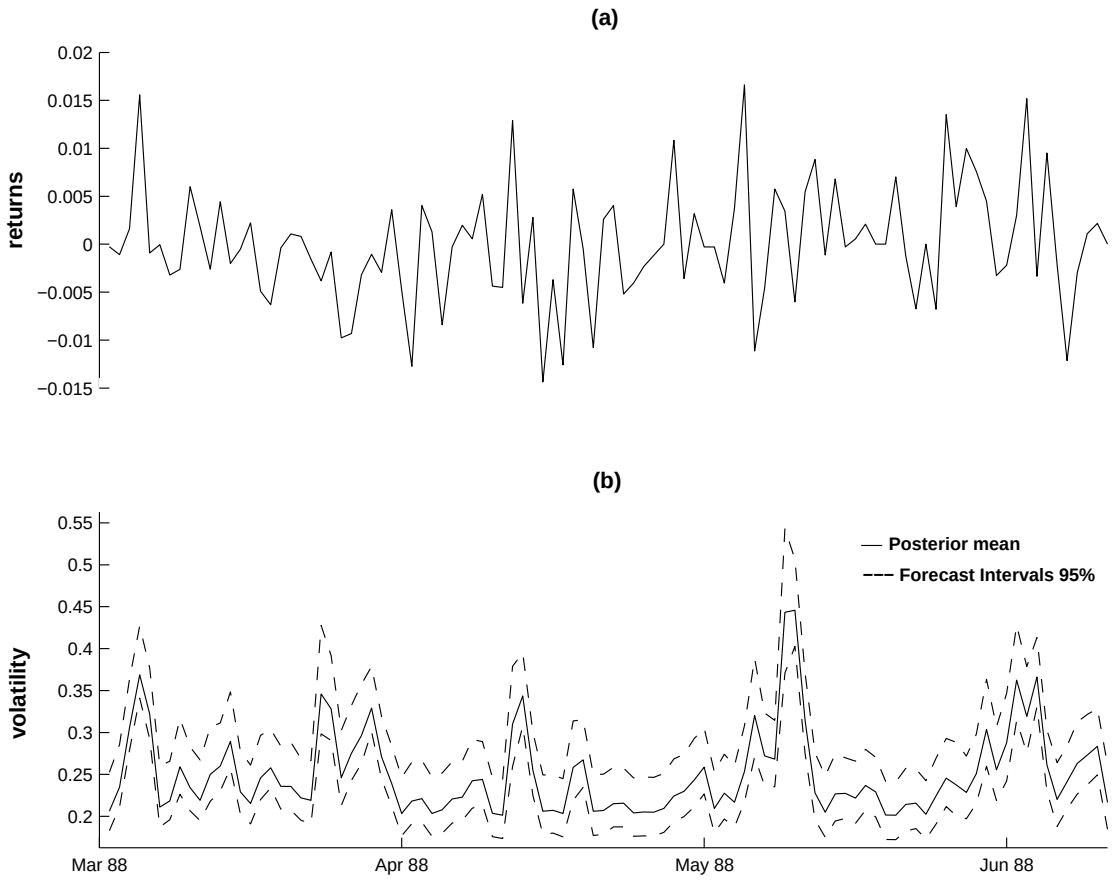

Figure 5: Exchange rates example. (a) Time series of returns starting from March 1988. (b) Predictive posterior means and 95% forecast intervals for volatility.

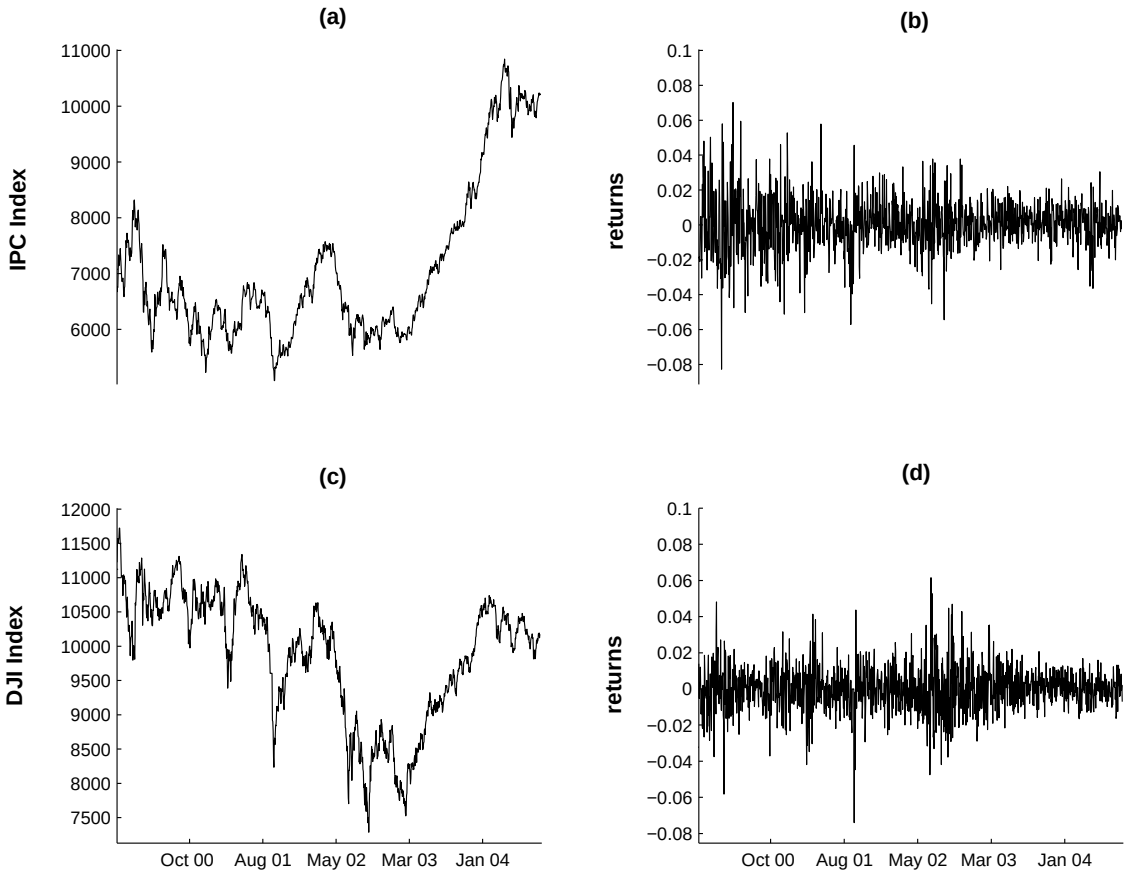

Figure 6: (a) The Mexican stock market (IPC) index from January 2000 to September 2004. (b) Time series of returns for the IPC index. (c) The Dow Jones Index from January 2000 to September 2004. (d) Time series of returns for the Dow Jones Index.

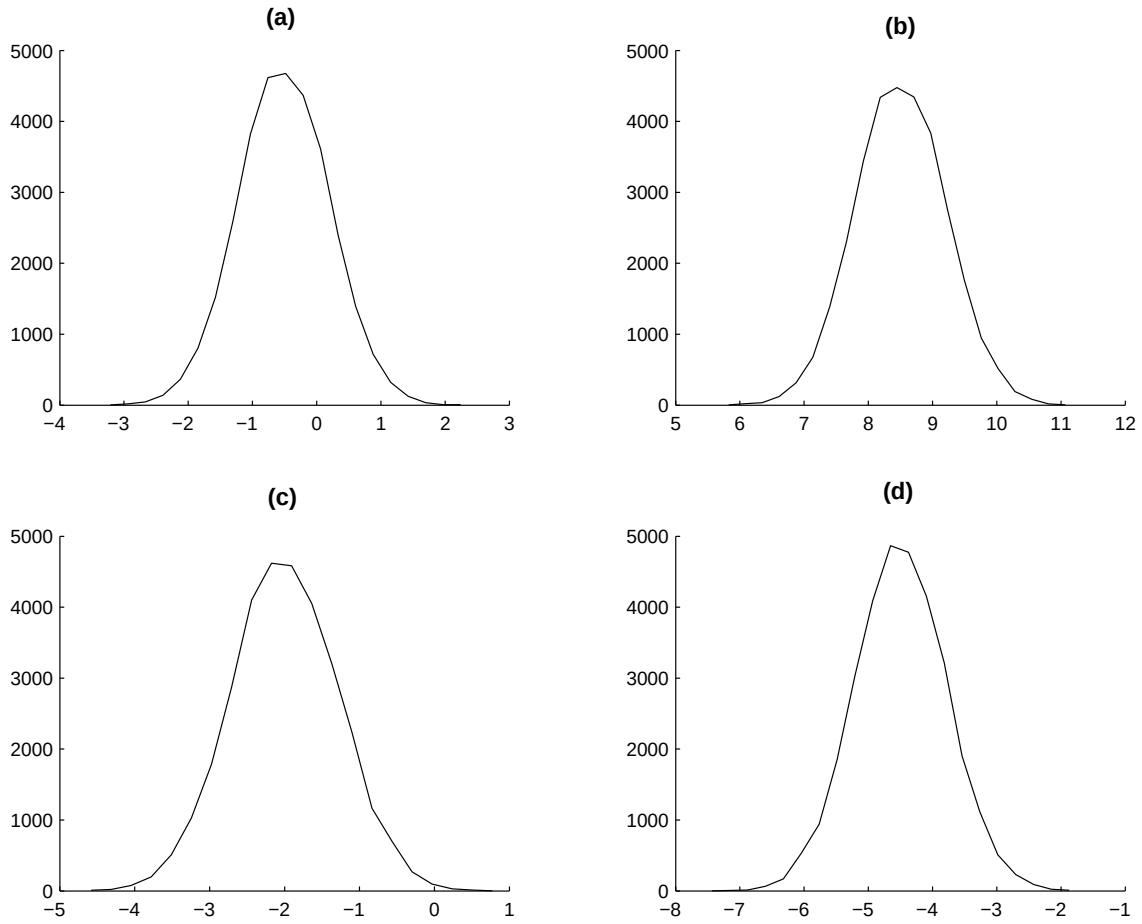

Figure 7: (a) Posterior distribution for gating parameter v_1. (b) Posterior distribution for gating parameter u_1. (c) Posterior distribution for gating parameter v_2. (a) Posterior distribution for gating parameter u_2.

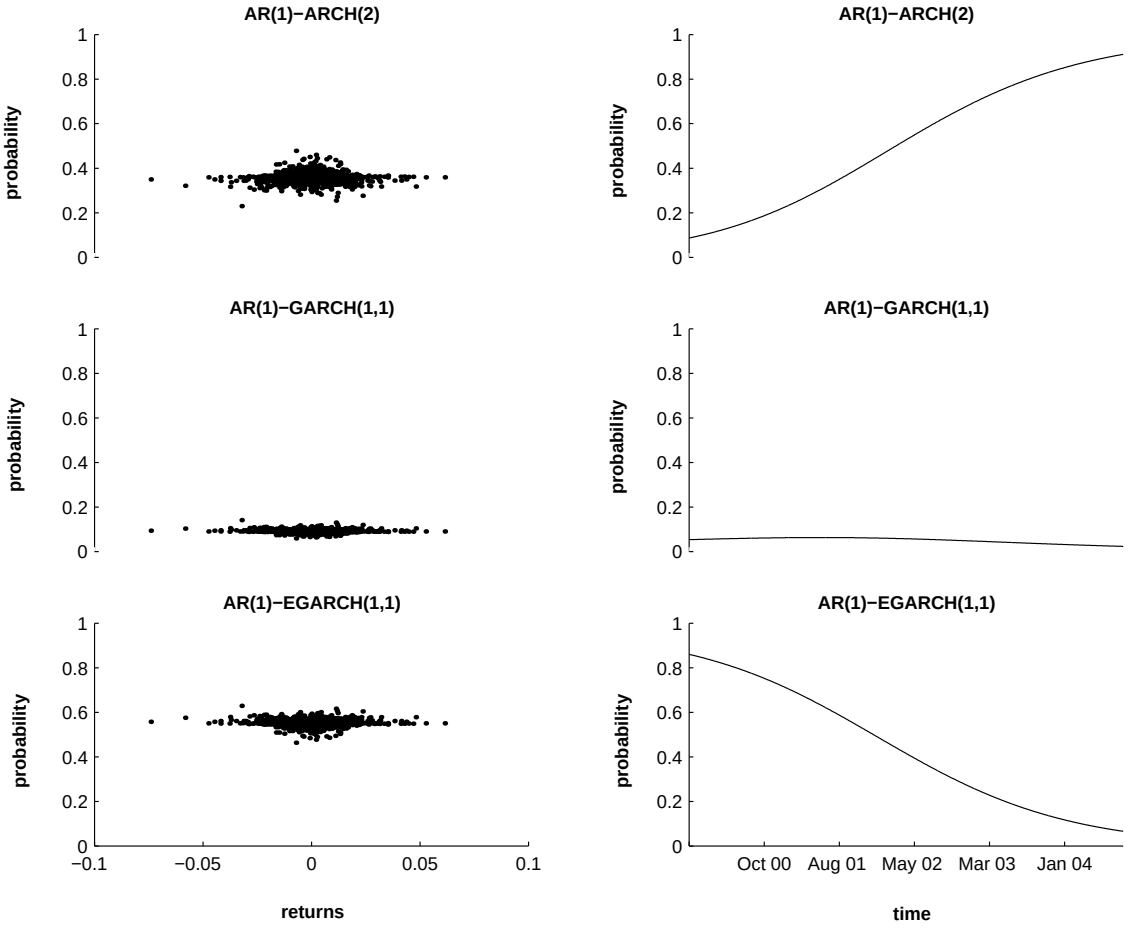

Figure 8: *Left column.* Probabilities of each expert as a function of the returns of the DJI. *Right column.* Probabilities of each expert as a function of time.