A Comment on ”The Art of Data Augmentation”

Gabriel Huerta,
Department of Probability and Statistics, Centro de Investigación en Matemáticas,
Apartado Postal 402, Guanajuato, Gto. 36000, México
&
Wenxin Jiang and Martin A. Tanner[1]
Department of Statistics, Northwestern University, 2006 Sheridan Rd.,
Evanston IL 60208-4070, U.S.A.

[1] M. Tanner was supported by NIH grant CA35464.

This insightful paper by van Dyk and Meng (VM) makes the important point that advances in data augmentation algorithms offer a wide variety of tools for statistical inference. Time series methods are no exception and mixture modeling within this context may help to improve forecasting and to detect changes in structure across time.

The time series modeling approach that we adopt is based on the idea of mixing models through the neural network paradigm known as Hierarchical Mixtures-of-Experts (HME) - see Jordan and Jacobs (1994). The HME approach easily allows for model comparison and permits one to represent the mixture weights as a function of time or other covariables. With the additional hierarchy, it is possible to localize the comparisons to specific *regions* or *regimes*. Furthermore, the defining elements of the mixture do not have to be restricted to a particular class of models permitting very general comparisons. In this comment parameters are estimated via maximum likelihood using the EM-algorithm- extensions to a full Bayesian approach using MCMC may follow one or more of the many lines outlined by VM. We see this comment as a call to the Chagalls of this world to use their artist abilities to develop quick mixing stochastic algorithms for this important, yet complex class of HME models.

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

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

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

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

Furthermore, the HME models considered have an additional layer designed with the purpose of local time series modeling. The HME partitions the covariate space, which could include time, into O overlapping regions called "overlays". In each overlay, M models are to compete with each other, in the hope that the model most suitable to the specific region is favored by a high weight (see Figure 1). By having multiple overlays, the hierarchical mixture model allows for modeling multiple switching across regions.

Therefore, the expert index J can be expressed as $J = (o, m)$, where the overlay index o takes a value from $\{1, \ldots, O\}$, and the model-type index m from $\{1, \ldots, M\}$. We allow the same type of model m to assume different versions or more specifically different parameter values, at each possible overlay.

The mixing weights are often referred to as "gating functions". They can depend on the previous history, exogenous information (see McCulloch and Tsay, 1993), or can exclusively depend on t. The gating functions may have the form

$$\mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma) = \left\{ \frac{e^{v_o + \mathbf{u}_o^T \mathbf{w}_t}}{\sum_{s=1}^{O} e^{v_s + \mathbf{u}_s^T \mathbf{w}_t}} \right\} \left\{ \frac{e^{v_{m|o} + \mathbf{u}_{m|o}^T \mathbf{w}_t}}{\sum_{l=1}^{M} e^{v_{l|o} + \mathbf{u}_{l|o}^T \mathbf{w}_t}} \right\}, \qquad (2)$$

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

In the context where one is interested in how the weighting for individual models is assigned across different time periods, $\mathbf{w}_t$ can be taken as (t/n). Therefore, one can adopt the following parametric form for the gating functions:

$$\mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma) = g_{om}(t; \gamma) \equiv \left\{ \frac{e^{v_o + u_o(t/n)}}{\sum_{s=1}^{O} e^{v_s + u_s(t/n)}} \right\} \left\{ \frac{e^{v_{m|o} + u_{m|o}(t/n)}}{\sum_{l=1}^{M} e^{v_{l|o} + u_{l|o}(t/n)}} \right\}. \qquad (3)$$

Here γ includes all the following components: $v_1, u_1, \ldots, v_{O-1}, u_{O-1}, \ldots, v_{1|1}, u_{1|1}, \ldots, v_{M-1|1}, u_{M-1|1}, \ldots, v_{M-1|O}, u_{M-1|O}$. For identifiability, we set $v_O = u_O = v_{M|o} = u_{M|o} = 0$ for all $o = 1, \ldots, O$. The free vector of parameters γ in the gating functions automatically determines the location and the "softness" of the splitting of the regions.

Note that this framework defines the two-layer HME architecture of Jordan and Jacobs (1994), where the first layer of gating functions hypothesizes O overlays on the entire time axis, and the second layer of gating functions defines weights for each of the M model types within each overlay. When the input space for the gating functions is time, the hierarchical mixture model can identify the region over which a model or a set of models is (are) dominant in a data-adaptive manner. Thus, the present approach allows for modeling multiple regime switching. Further details of this approach, as well as related asymptotic theory are presented in Huerta, Jiang and Tanner (2000).

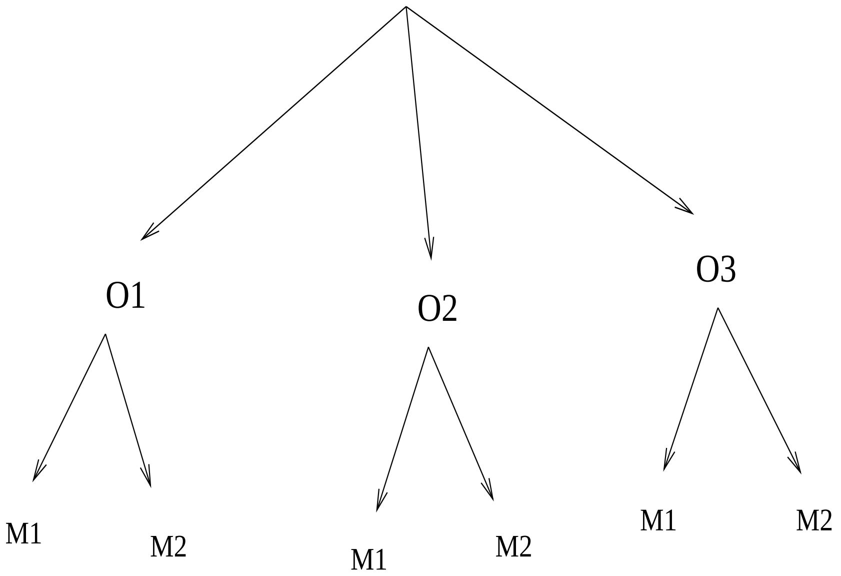

Figure 1: *A Graphical Representation of a Two-Layer HME.*

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

$$\mathcal{L}_n(\cdot) = n^{-1} \sum_{t=1}^{n} \log f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \cdot). \tag{4}$$

We denote the maximum likelihood estimate (MLE) of θ as $\hat{\theta} = \arg\max \mathcal{L}_n(\cdot)$. To obtain the MLE, the EM algorithm starts with an initial estimate of the parameters θ^0. Then a sequence $\{\theta^i\}$ is obtained by iterating between the following two steps:

For $i = 0, 1, 2, \ldots,$

E-step: Construct

$$Q^i(\theta) = \sum_{t=1}^{n} \sum_{o,m} h_{om}(t; \theta^i) \log\{\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta) \mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma)\}, \tag{5}$$

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

$$h_{om}(t; \theta) = \frac{g_{om}(t; \gamma)\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta)}{\sum_{s=1}^{O} \sum_{l=1}^{M} g_{sl}(t; \gamma)\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, s, l; \eta)} \tag{6}$$

is the "posterior probability" of choosing the expert (o, m) at time t.

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

Inference is greatly facilitated by the introduction of augmented data, resulting in the fact that the objective function Q^i has the form of a double sum of logarithms, instead of a "sum log sum" typical for the log likelihood function $\mathcal{L}_n$. For this reason, the maximization of the objective function can be decomposed into a number of smaller maximization problems which involve fewer parameters and usually define "known" maximizations of widely used models. For example, suppose the expert pdf has the form

$$\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta) = p_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, m; \eta_{om}), \tag{7}$$

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

$$\eta_{om}^{i+1} = \arg\max_{\eta_{om}} \sum_{t=1}^{n} h_{om}(t; \theta^i) \log p_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, m; \eta_{om}), \tag{8}$$

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

When the MLE $\hat{\theta}$ is obtained, we are interested in evaluating the relative weighting of each of the M model types at time t. Two estimates are of interest in this regard. One is an empirical Bayes estimate of the posterior probability / weight of model type-m at time t. This is the *conditional* probability regarding the history up to time t and defined by:

$$\hat{P}_t(m|y_t, \mathcal{F}_{t-1}, \mathcal{X}) \equiv \hat{h}_m(t) \equiv \sum_{o=1}^{O} h_{om}(t; \hat{\theta}), \tag{9}$$

where $\hat{\theta}$ is the MLE. Another approach for weighting is to consider an empirical Bayes-type estimate of the *unconditional* probability / weight of model m at time t (unconditional on the history of the endogenous process $\{y_t\}$):

$$\hat{P}_t(m|\mathcal{F}_{t-1}, \mathcal{X}) \equiv \hat{g}_m(t) \equiv \sum_{o=1}^{O} g_{om}(t; \hat{\theta}). \tag{10}$$

As we shall see in an example, (9) can vary point-wise over time due to the conditioning on the specific history of the observations. The second weighting scheme (10) is smoother when describing a regional change of preference for model m. The term $\hat{h}_m(t)$ is an estimated "posterior" probability of model m, and $\hat{g}_m(t)$ is the corresponding estimate of the "prior" probability in the sense that the prior probability is not conditional y_t and the posterior probability is conditional on y_t. These estimates are not formal Bayesian priors and posterior probabilities for model m, since we have not assigned any prior $p(\theta)$ to the parameters θ, but instead are estimating the conditional or "unconditional" mixing weights at the MLE.

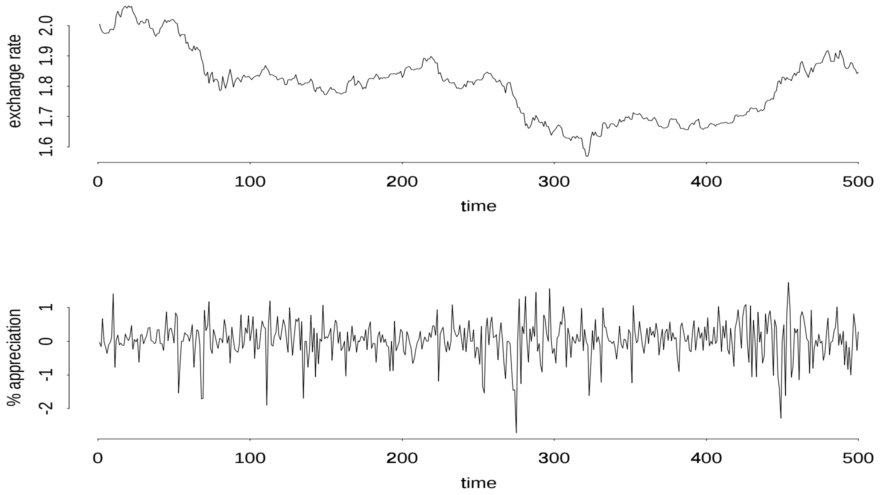

Figure 2: *Exchange Rate Data.* Top: 500 observations of spot rates between the German-mark and US-dollar beginning in October 1986. Bottom: Logarithm of the first difference of the spot rates between the German-mark and the US-dollar.

We consider a financial time series to illustrate this hierarchical mixture defined with GARCH and EGARCH models - see Bollerslev (1986) and Nelson (1991). The series consist of 500 daily observations of exchange rates between the German-mark and the US-dollar starting from October 9, 1986. In fact, Figure 2 presents a time plot of the 500 daily *spot rates* characterized by a non-stationary random walk behavior.

Also in the same figure, we present the logarithm of the first difference of spot rates as a function of time, a transformation which is widely used to induce covariance-stationarity and propose parametric models. Assuming that Y_t represents the log of the first difference in spot rates, as discussed in Andersen and Bollerslev (1998), 3 candidate models are used to obtain inferences on *volatilities* or innovation variances at each time t:

- An AR(1) model simply defined by:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t,$$

  where $\epsilon_t \sim N(0, \sigma^2)$.

- An AR(1)-GARCH(1,1) model represented by

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t,$$

but now $\epsilon_t \sim N(0, \sigma_t^2)$. That is, the innovation variance can change in time and according to the evolution equation

$$\sigma_t^2 = \theta_0 + \theta_1 \epsilon_{t-1}^2 + \theta_2 \sigma_{t-1}^2,$$

with non-negative parameters θ_0, θ_1 and θ_2.

- Finally, an AR(1)-EGARCH(1,1) is considered, where again

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t$$

but now the evolution in variance is defined in terms of the natural logarithm and the standardized innovations $z_t = \epsilon_t / \sigma_t$ through the expression

$$\log(\sigma_t^2) = \beta_0 + \beta_1 z_{t-1} + \beta_2(|z_{t-1}| - \sqrt{2/\pi}) + \beta_3 \log(\sigma_{t-1}^2),$$

with no restrictions on the parameters β_0, β_1, β_2 and β_3.

Within an HME framework taking $M = 3$: $\pi_t|_{o,m=1}$ denotes the pdf of Y_t given the history based on an AR(1); $\pi_t|_{o,m=2}$ the same pdf with an AR(1)-GARCH(1,1); and $\pi_t|_{o,m=3}$ denotes the pdf with an AR(1)-EGARCH(1,1). As before, the index o is added to model parameters and our initial exploration is based on a value of $O = 2$, i.e. allowing for 2 overlays. We ran the EM-algorithm with 20 different starting points. Parameters for the pdfs $\pi_t|_{o,m}$ were initialized at the individual model MLE's and initial parameters for the gating functions were generated from uniform distributions. Each EM was run for 500 iterations and solutions were ranked using the log-likelihood function $\mathcal{L}_n(\cdot)$.

Figure 3 presents the estimates $\hat{g}_m(t)$ for $m = 1, 2, 3$. In general, the model assigns a very low weight to the AR(1), with competing weights for AR(1)-GARCH(1,1) and AR(1)-EGARCH(1,1), for approximately the first 100 observations of the series. For the remaining segment of the time series, the preferred model is the AR(1)-EGARCH(1,1).

In Figure 4, we present the estimates of $\hat{h}_m(t)$ for $m = 1, 2, 3$. Although the general "smoothed" pattern is similar to that exhibited by Figure 3, the model posterior probabilities have jumps of high probability for the three competing alternatives. This example reflects how $h_m(t)$ can be highly impacted by single observations. The "ups and downs" in volatility experienced by exchange rate data across time lead to these model switches. Periods of almost constant variance can be well represented by an AR(1) model but when the data present periods of non-constant variance, the GARCH or EGARCH structure dominates producing the large jumps in the functions $h_m(t)$.

Figure 5 presents a comparison of the estimated volatilities of the HME with those based on the individual models. The MLE, computed with a numerical optimizer, was

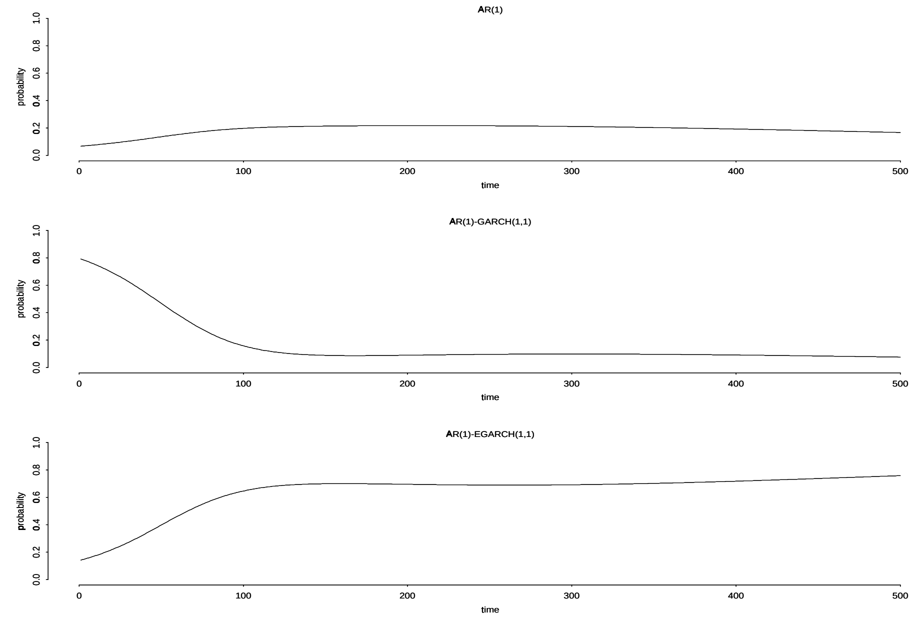

Figure 3: *Exchange Rate Data Example.* Maximum likelihood estimates of $g_m(t)$ for the 3 models considered: AR(1), AR(1)-GARCH(1,1) and AR(1)-EGARCH(1,1).

used to obtain the volatilities for AR(1), GARCH(1,1) and EGARCH(1,1). For the present hierarchical mixture model, the volatilities were estimated using the EM-solution, recognizing that for this mixture model, the variance of Y_t given the history is the expectation with respect to the mixing weights $g_m(t)$ of individual model-variances plus the variance of the expectation functions for each defining model. We note that the volatility for the HME smoothes some of the high volatility peaks induced by other models but recognizes the overall pattern suggested by the AR(1)-GARCH(1,1) and the AR(1)-EGARCH(1,1) processes.

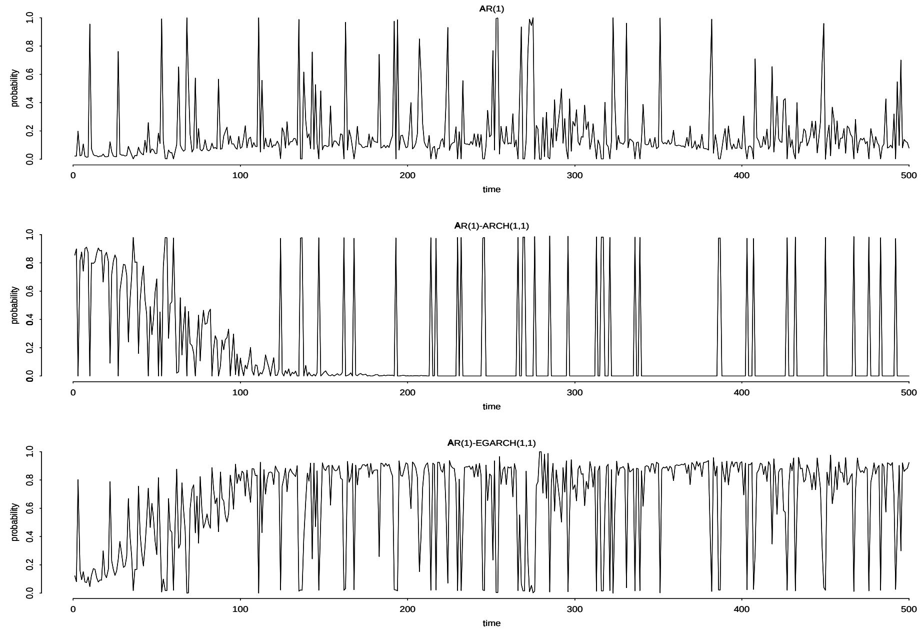

Figure 4: *Exchange Rate Data Example.* Maximum likelihood estimates of $h_m(t)$ for the 3 models considered: AR(1), AR(1)-GARCH(1,1) and AR(1)-EGARCH(1,1).

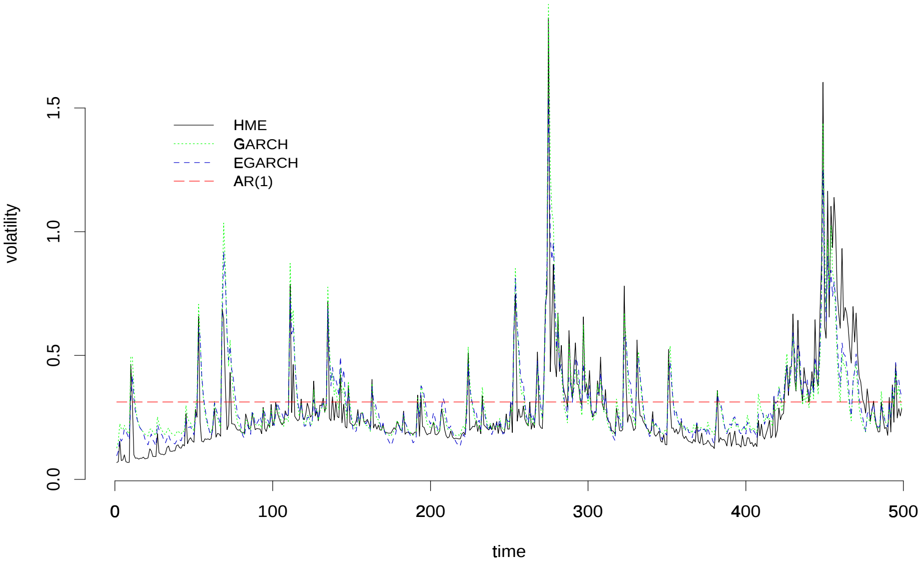

Figure 5: *Exchange Rate Data Example.* Comparison of estimated volatilities for the HME and the individual models AR(1), AR(1)-GARCH(1,1) and AR(1)-EGARCH(1,1).

References

Andersen, T. G. and Bollerslev, T. (1998). ARCH and GARCH models. *Encyclopedia of Statistical Sciences. Update Vol.* **2**, 6-16. Edited by Kotz, S., Read, C. B. and Banks, D. L., Wiley, New York.

Bollerslev T. (1986). Generalized autoregressive conditional heteroskedasticity *Journal of Econometrics* **31**, 307-327.

Huerta, G., Jiang, W., and Tanner, M. A. (2000). Time series modeling via hierarchical mixtures. Technical Report, Northwestern University.

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

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association* **88**, 968-978.

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