Journal of Forecasting
J. Forecast. **26**, 457–473 (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1008

The Extended Switching Regression Model: Allowing for Multiple Latent State Variables

ARIE PREMINGER,[1] URI BEN-ZION[2] AND DAVID WETTSTEIN[2*†]
[1] *CORE Université Catholique de Louvain, Louvain-la-Neuve, Belgium*
[2] *Department of Economics, Ben-Gurion University of the Negev, Beer-Sheva, Israel*

ABSTRACT

In this paper we extend the widely followed approach of switching regression models, i.e. models in which the parameters are determined by a latent discrete state variable. We construct a model with several latent state variables, where the model parameters are partitioned into disjoint groups, each one of which is independently determined by a corresponding state variable. Such a model is called an extended switching regression (ESR) model. We develop an EM algorithm to estimate the model parameters, and discuss the consistency and asymptotic normality of the maximum likelihood estimates. Finally, we use the ESR model to combine volatility forecasts of foreign exchange rates. The resulting forecast combination using the ESR model tends to dominate those generated by traditional procedures.

KEY WORDS algorithm; extended switching regression model; forecast combining

INTRODUCTION

Many economic and financial time series are characterized by dramatic changes over time, which can be attributed to various causes such as wars, financial panics, or significant changes in economic policy. To correctly analyze such data, and especially when focusing on prediction or simulation, it is important to model such behavior explicitly, and take into account the factors that might cause these changes.

One of the main approaches to explain this behavior of time series is the switching regression approach, in which the parameters changes are controlled by a latent state variable. The latent state variable is not a real variable that happens to be missing. Rather, this variable represents

* Correspondence to: David Wettstein, Department of Economics, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel. E-mail: wettst@bgu.ac.il
† Visiting professor, Università Bocconi, Milan, Italy.

underlying factors without a precise physical definition, but which often turn out to have a meaningful physical interpretation. More specifically, these models are based on the assumption that the data-generating process changes over time, and there is a latent model selection procedure dependent on a discrete state variable which randomly picks a parametric model each time. This procedure is characterized by defining a set or subset of the model parameters to be mutually dependent on the state variable.

Switching regression models have a rich history in econometrics; see Maddala and Kim (1998) for a comprehensive survey. Recently, a new class of switching regression models obtained as mixtures of linear autoregressive models has been considered by several authors, e.g. Le *et al.* (1996), Li and Wong (2000, 2001), and Lanne and Saikkonen (2003). These models can be defined by specifying the conditional distribution of the observed time series as a mixture of normal distributions with each component of the mixture similar to the conditional distribution of a Gaussian linear autoregressive model. Related classes of models are the Markov regression models (Hamilton 1994, Ch. 22), which differ from the switching regression models in that the unobserved state variable follows a latent Markov structure.

These models have proven to be extremely useful in modeling low-frequency (e.g. quarterly or monthly data) economic and financial time series (see, for example, Hamilton, 1990; Engel and Hamilton, 1990; Klaassen, 2005). In order to model high-frequency data (e.g. weekly or daily data) we may need to consider a larger number of states. However, one should consider the practical limitation that the number of parameters grows substantially with the cardinality of the state space. Therefore, we extend this modeling approach by positing the existence of several state variables, which independently influence the latent model selection procedure through the picking of a partial and disjoint group of the model parameters. We call this approach the extended switching regression (ESR) model. This approach relaxes the assumption that the model parameters are mutually dependent on one state variable, and allows independent switches in subsets of the model parameters over time.

The advantage of formulating our model in such a way is that we can consider many distinct states without substantial increase in the model parameters. This is done by imposing a tight set of restrictions on the switching parameters and assuming that the probability of each state can be factored into several components related to the probabilities of the latent state variables. Furthermore, the underlying assumption which usually motivates the usage of switching regression models in economic applications is that all economic agents switch to the new regime at once. However, if the economy consists of many individuals or firms, each of whom switches in different times, the ESR model seems more appropriate because it enables gradual switching by allowing some of the model parameters to remain the same while shifts occur in other parameters.

We apply the ESR model to combine forecasts of exchange rate volatility and compare out-of-sample performance of this model to traditional linear forecast combining methods and other methods for combining forecasts, derived from switching regression models. The motivation for using the ESR model is that when one uses a linear combination of several individual forecasts to obtain a single forecast, the weights which are given to each individual forecast may change over time. The changes in the weights may be associated with the realization of several independent state variables.

The layout of this paper is as follows: the next section introduces the extended switching regression (ESR) model. The third section discusses the large sample properties of the model. In the fourth section we develop an EM algorithm for estimation. An empirical application is provided in the fifth section, where we use the ESR model to combine forecasts and compare it with other forecast combining models. The sixth section summarizes the main findings and outlines several extensions.

DOI: 10.1002/for

THE ESR MODEL

Switching regression models are motivated by the realization that time series may have parameters that are themselves changing over time.

This is in contrast to the situation expressed by a regression model:

$$y_t = \mu_t(x_t, \psi_0) + \varepsilon_t \tag{1}$$

where $\mu_t : \mathrm{X} \times \Psi \to R$ are known functions measurable on X for each ψ_0 in Ψ, a compact subset of R^d, and continuous on Ψ a.s. for all t. The error is a zero-mean Gaussian white noise and $x_t \in \mathrm{X}$ is a vector of explanatory variables. While these models have been popular in describing some 'behavioral law', they are not flexible enough to account for situations where the researcher believes the parameters are not constant over time.

Switching regression (SR) was developed as a way of allowing data to arise from a combination of two or more distinct data generation processes. An equivalent description of the models presented above is based on the assumption that there is a single unobserved random discrete variable s_t, which will be called a state variable. This variable selects a specific conditional distribution of y_t. So, if there are k possible distributions, $s_t = 1$, when the process distributes according to the first distribution; $s_t = 2$, the process distributes according to the second distribution, and so on, i.e. $s_t \in \{1, \ldots, k\}$. The state variable is unobserved and the distribution of s_t is multinomial; therefore, in the switching regression model we have

$$y_t = \mu_t(x_t, \psi(s_t)) + \varepsilon_t \tag{2}$$

This model has the important property that $\psi(s_t) \in \Psi$, the parameters governing the data generation process, change over the set $\{\psi_1, \psi_2 \ldots, \psi_k\}$ according to the values of the state variables.[1] For example in linear switching regression models $\mu_t = x_t' \cdot \beta(s_t)$ and the parameter vector $\beta(s_t)$ change over the set $\{\beta_1, \beta_2 \ldots, \beta_k\}$ according to the random values of the state variable.

In the ESR model we propose, the existence of p discrete switches in disjoint groups of the model parameters is assumed. The changes in the ith group depend only on the realization of s_t^i, unobserved i.i.d. state variables which can assume one of k_i integer values $\{1, 2, \ldots, k_i\}$. In order to simplify our discussion we will assume without loss of generality that $k = k_i$, although the number of elements in each of the subsets can be different.

The ESR model can be described as

$$y_t = \mu_t(x_t, \psi_1(s_t^1), \ldots, \psi_P(s_t^p)) + \varepsilon_t \tag{3}$$

where $\mu_t : \mathrm{X} \times \Psi_1 \ldots \times \Psi_p \to R$. The parameter set Ψ is being partitioned into p distinct subsets such that $\underset{i=1}{\overset{p}{\times}} \Psi_i = \Psi \subset R^d$ and the function $\psi_i(s_t^i) \in \Psi_i \subset R^{d_i}$ ($d = \sum_{i=1}^{p} d_i$) associates with each realization of the state variable s_t^i, a parameter vector which is chosen from the set $\{\psi_{i1}, \psi_{i2} \ldots, \psi_{ik}\}$. The ESR models we consider in this work are linear with individual coefficients that are allowed to shift independently over time. The full specification of the model and the estimation procedure are given in what follows.

[1] It may be possible to allow the error term variance to shift between states; however, for simplicity, we assume it remains constant.

DOI: 10.1002/for

Note that if we impose the restriction that $s_t^1 = s_t^2 = \ldots s_t^p$ or assume $p = 1$ and $\Psi_1 = \Psi$ the process described by equation (3) is equivalent to the switching regression model presented in equation (2). On the other hand, the ESR model can be estimated inefficiently by a switching regression model. That is, if there are p state variables, where each state variable can assume one of k states, we have to estimate a switching regression model with k^p states, which will be estimated with $k^p - 1$ probabilities as well as other model parameters. This creates an identification and estimation problem as we increase p and k. In the ESR model we can parameterize the same state space with only a few parameters by imposing independence among the state variables. Furthermore, if an ESR model describes the true data-generating process, it is better to estimate it directly, rather than resort to the conventional SR approach that nests it, with the cost of having to estimate an exponentially increasing number of parameters. Also note that under the ESR interesting qualitative information may result from the nature of the state variables, whereas this information will be lost if we estimate a model with one state variable.

In this context it is important to note a few things. The independence assumption reduces the number of parameters in the model. However, if there is a priori information or a reasonable assumption regarding the probability structure of the state variables, we can use it. So, although we use the independence assumption as a basis for our estimation in the rest of this paper, nothing prevents us from assuming some form of dependence among the state variables. It should, however, be stressed that such assumptions entail a decrease in the degrees of freedom.

We rely on maximum likelihood estimation techniques and in the next section discuss the asymptotic properties of the estimates derived in such a manner.

ASYMPTOTIC THEORY

In this section we discuss large-sample properties of the maximum likelihood estimates of the ESR model. Let $\{Z_t\}_{t\geq 0}$ be a discrete-time, real-valued, stochastic process. The vector Z_t is partitioned into $Z_t = (Y_t, X_t)$, where Y_t is the dependent variable and X_t is the vector of explanatory variables which could include lagged variables of Y_t. We are interested in a parametric family of conditional distributions[2] $\{P^t_{Y|X}(\psi), \psi \in \Psi\}$ of Y_t given X_t, which exists by Jirina's theorem (Bauer, 1972), where the conditional distributions have Radon–Nikodym derivatives $g(y_t|x_t, \psi)$ with respect to some σ-finite measure ν on the real line (these derivatives usually correspond to the usual notion of a density functions).

A stochastic process is said to follow an extended switching regression or ESR model if there exists a latent model selection procedure which picks an element from the set $\{g(y_t|x_t, \psi), \psi \in \Psi\}$ of conditional densities for each t. The selected (probability) model provides a complete description of the stochastic behavior of the data for each point in time.[3] This selection process is unobserved and characterized by independent selections from the disjoint parameter sets $\{\Psi_i\}$, where $\Psi = \overset{p}{\underset{i=1}{\times}} \Psi_i$. From each set Ψ_i we select one element among k possible ones. The selection is dependent on the realization of p unobserved independent state variables $s_t^i \in \{1, \ldots, k\}$ as described in the previous

[2] In the previous section we assumed for simplicity that only the parameters of the conditional expectation switch over time; this assumption does not prevent us from dealing with the general case when discussing the large sample properties of the maximum likelihood estimator.

[3] Note that in the ESR model it is possible, even though the switches in the model parameters are independent, for the resulting explanatory variables to be dependent.

DOI: 10.1002/for

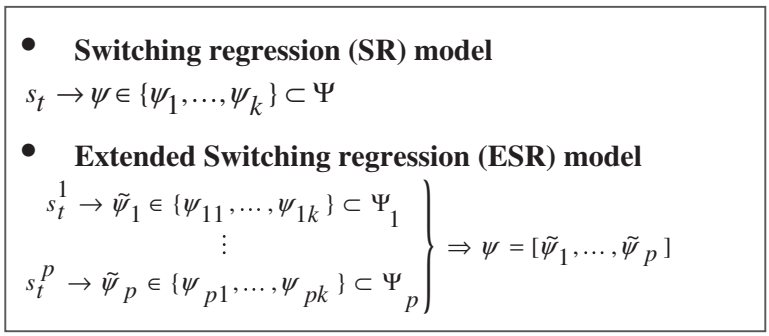

Figure 1. The latent structure of the ESR model and the SR model

section. A concise comparison between the latent structures of the ESR and the switching regression (SR) models is given in Figure 1.

Let $\{p_{i1}, \ldots, p_{ik}\}$ denote the probability of the ith selection and let $\varphi_i \subset \Psi_i$ be a set of k distinct values determined by s_t^i, where each element of this set is denoted by φ_{ij_i} and is chosen with probability p_{ij_i}, where $j_i \in \{1, \ldots, k\}$. The conditional density of y_t is given by[4]

$$f(y_t|x_t, \theta) = \sum_{j_1=1}^{k} \ldots \sum_{j_p=1}^{k} \left[\left(\prod_{i=1}^{p} p_{ij_i} \right) \cdot g(y_t|x_t, \varphi_{1j_1}, \ldots, \varphi_{pj_p}) \right] \tag{4}$$

where $\theta = (\varphi_1, \ldots, \varphi_p, p_{11}, \ldots, p_{1k-1}, \ldots, p_{p1}, \ldots, p_{pk-1})$ is the vector of the model parameters. Let $\hat{\theta}_T$ be the maximum likelihood estimator which maximizes the following log-likelihood function:

$$L_T(\theta) = \sum_{t=1}^{T} \log f(y_t|x_t; \theta) \tag{5}$$

Since under each state the density functions are from the same parametric family, the ESR model in (4) is invariant under permutation of the labels in $(\varphi_{i1}, p_{i1}, \ldots \varphi_{ik}, p_{ik})$ for each i. Hence for this model there exist $(k!)^p$ distinct parameter values, which give the same likelihood due to 'label switching'. For more details on the 'label switching' problem in mixture models, see Render (1981) and Render and Walker (1984). In addition, when $p_{ij} = 0$ and/or $\varphi_{ij} = \varphi_{ij'}$ for some $i, j \neq j'$, the structure of the ESR model is degenerate in the sense that there exists another model with fewer states which gives the same value to the log-likelihood function, and therefore some of the model parameters are not identified. To see it, assume that one of the model parameters has the same value in all states, i.e. it is constant but has been modeled as a switching parameter, and then its corresponding probability parameters are not identified. Furthermore, in this case, the likelihood function is flat, the Fisher information matrix is singular and the regularity conditions needed to establish the asymptotic properties (consistency, asymptotic normality) of the maximum likelihood estimator are not satisfied.

In order to avoid the identification problem, we assume that the probabilities of the state variables are bounded away from zero and that the values of the parameters are strictly increasing over states

[4] Note that when $p = 1$ the conditional density is the same as for the switching regression model.

DOI: 10.1002/for

with the size of the jump bounded away from zero. That is, we impose the following restrictions on the parameter set; $p_{ij} \in [\tau_1, 1 - \tau_1]$ and $\varphi_{ij} + \tau_2 \leq \varphi_{ij+1}$ for $i = 1, 2, \ldots, p$ and $j = 1, \ldots, k - 1$, where $\tau_1, \tau_2 > 0$ are fixed small numbers. Let Θ be the parameter space which satisfies these restrictions and contains θ_0 (the underlying parameter vector of the ESR model[5]), in its interior. By examining only those parameters in this set we may say that θ_0 is *identifiably unique* (White, 1994, def. 3.3) on this set.

It can be shown, similarly to Jeffries (1998) and Rahbek and Shepherd (2002), that for stationary and ergodic processes and under mild regularity conditions which take into account the identifiability and singularity problems of the ESR model as mentioned above, we can apply Theorem 2.1 and 2.2 of Billingsley (1961, pp. 10–14) to establish that the maximum likelihood estimator is consistent and asymptotically normal, so that

$$\hat{\theta}_T \to {}_p\theta_0 \quad \text{and} \quad \sqrt{T}(\hat{\theta}_T - \theta_0) \Rightarrow N(0, Q^{-1}) \tag{6}$$

where

$$Q_{ij} = \frac{1}{T}\sum_{t=1}^{T} E\left[\frac{\partial \text{Log} f(Y_t|X_t, \hat{\theta}_T)'}{\partial \theta_i} \cdot \frac{\partial \text{Log} f(Y_t|X_t, \hat{\theta}_T)}{\partial \theta_j}\right]$$

and $\to_p$ denotes convergence in probability and $\Rightarrow$ denotes convergence in distribution. The empirical counterpart of the Fisher information matrix is easily obtained by taking expectation with respect to the empirical distribution. For dependent and heterogeneous data, see the discussion in Preminger *et al.* (2003). Under model dynamic misspecification we can use the results of Domowitz and White (1982, 1984); see also Gallant and White (1988) and White (1994) for a comprehensive discussion.

ESTIMATION

The linear ESR model we consider in this work is given by

$$y_t = \alpha + \sum_{i=1}^{d} \beta_{it} x_{it} + \varepsilon_t \tag{7}$$

where $\varepsilon_t \sim i.i.N(0, \sigma)$ and $\beta_{it} \in \{\beta_{i1}, \ldots, \beta_{ik}\}$ for $1 \leq i \leq d$, although it may be possible to allow the intercept and the error variance to shift between several states over time; we assume for simplicity that they remain constant over the whole period. The parameter β_{it} is a random variable. Its distribution is discrete and is determined by a state variable s_t^i as described in the previous section. An equivalent description of (7) is given as follows:

$$y_t = \alpha + \sum_{i=1}^{d} \beta_i S_t^i x_{it} + \varepsilon_t \tag{8}$$

[5] By θ_0, we mean one of the parameter vectors which are obtained from the true parameters by label switching.

DOI: 10.1002/for

where $\beta_i = [\beta_{i1}, \ldots, \beta_{ik}]$ and $\{S_t^i\}_{i=1}^d$ is a set of k-dimensional d vectors with the j element in the vector S_t^i equaling one if $s_{it} = j$ and zero otherwise, hence

$$S_t^i = \begin{cases} [1, \ldots, 0]' & \text{if} \quad s_t^i - 1 \\ \vdots & \vdots \\ [0, \ldots, 1]' & \text{if} \quad s_t^i = k \end{cases} \tag{9}$$

The estimation of the linear ESR model is done via the EM algorithm popularized by Dempster *et al.* (1977); see also McLachlan and Krishnan (1997). It should be noted that an EM approach is not necessary—the parameters could be estimated by other minimization or maximization routines. We want to point out, though, that the likelihood function will not be a concave function because it is a mixture likelihood, and thus Newton–Raphson-based optimizations may not work well. The advantage of using the EM algorithm lies in the fact that the likelihood values increase (weakly) in each iteration, thus ensuring the algorithm will converge to a local maximum in almost all cases. This means that if θ^ℓ denotes the EM algorithm's estimate of the true parameters after ℓ iterations, then the EM estimates satisfy $L_T(\theta^{\ell+1}) > L_T(\theta^\ell)$. This monotone property of the EM estimates becomes more important as the number of estimated parameters increases. As more parameters are added, it is more likely that the standard maximization routines will fail. There are pathological constructions in which the EM estimates may converge to a critical point other than a local maximum (Wu, 1983), but such aberrations are usually overcome by changing the starting values of the algorithm. In order to avoid the identifiability problem due to the structure of the ESR model, we impose the following restrictions on the parameter space; $p_{ij} \in [\varepsilon_1, 1 - \varepsilon_1]$ and $\beta_{ij} + \varepsilon_2 \leq \beta_{ij+1}$ for $1 \leq i \leq d$ and $1 \leq j \leq k - 1$, where $\varepsilon_1, \varepsilon_2 > 0$ are fixed small numbers.

Now, let $S_t = (1, S_t^{1'}, S_t^{2'}, \ldots, S_t^{d'})'$, $B = (\alpha, \beta_{11}, \ldots, \beta_{1k}, \ldots, \beta_{d1}, \ldots, \beta_{dk})'$, $\tilde{X}_t = (1, x_{1t}, \ldots, x_{1t}, x_{2t}, \ldots, x_{d-1,t}, \ldots, x_{dt}, \ldots, x_{dt})'$. We first write the ESR model in a more compact way that would simplify calculations later on:

$$y_t = (B \odot \tilde{X}_t)' \cdot S_t + \varepsilon_t \tag{10}$$

The symbol '$\odot$' denotes the Hadamard product, which means element-by-element multiplication. Also $\theta^\ell = [B^\ell, \sigma^\ell, p_{11}^\ell, \ldots, p_{1k}^\ell, \ldots, p_{d1}^\ell, \ldots, p_{dk}^\ell]$ denotes the parameters estimated in the ℓth iteration and let $\Lambda_t(\theta^{\ell-1}) = E(S_t S_t' | y_t, x_t; \theta^{\ell-1})$, $\hat{S}_t(\theta^{\ell-1}) = E(S_t | y_t, x_t; \theta^{\ell-1})$.

The EM procedure

The M-step

The (conditional) likelihood function of the complete data, assuming that the values of the state variables are known, is given by

$$L_T^C(\theta) = \sum_{t=1}^{T} \sum_{i=1}^{d} \sum_{j=1}^{k} S_t^{ij} \log(p_{ij}) - \sum_{t=1}^{T} \frac{\left(y_t - (B \odot \tilde{X}_t)' S_t\right)^2}{2\sigma^2} - 0.5T \log(2\pi\sigma^2) \tag{11}$$

The expectation is taken with respect to the distribution of the state variables given the data and the parameters estimated in the previous iteration (we describe the exact mode of calculation of the E-step later on) and yields

DOI: 10.1002/for

$$E\left(L_T^C\left(\theta^\ell \middle| \{y_t, x_t\}_{t=1}^T; \theta^{\ell-1}\right)\right) = \frac{1}{2\sigma^2}\sum_{t=1}^{T}\Big\{y_t^2 - 2y_t(B\odot\tilde{X}_t)'\hat{S}_t(\theta^{\ell-1}) + (B\odot\tilde{X}_t)'\Lambda_t(\theta^{\ell-1})(B\odot\tilde{X}_t)\Big\} + \sum_{t=1}^{T}\sum_{i=1}^{d}\sum_{j=1}^{k}\hat{S}_t^{ij}(\theta^{\ell-1})\log p_{ij} - 0.5T\log(2\pi\sigma^2) \tag{12}$$

Differentiating with respect to B yields

$$B^\ell = \left(\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t'\odot\Lambda_t(\theta^{\ell-1})\right)^{-1}\sum_{t=1}^{T}y_t\tilde{X}_t\odot\hat{S}_t(\theta^{\ell-1}) \tag{13}$$

Differentiating with respect to σ and p_{ij} is seen to yield

$$\sigma^\ell = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\Big[y_t^2 - 2y_t(B^\ell\odot\tilde{X}_t)'\hat{S}_t(\theta^{\ell-1}) + (B^\ell\odot\tilde{X}_t)'\Lambda_t(\theta^{\ell-1})(B^\ell\odot\tilde{X}_t)\Big]} \tag{14}$$

$$p_{ij}^\ell = \frac{1}{T}\sum_{t=1}^{T}\hat{S}_t^{ij}(\theta^{\ell-1}) \tag{15}$$

The E-step

Calculation of the expectation of the log-likelihood function of the complete data is done with respect to the distribution of the latent state variables given the data, and the parameters estimated in the previous iteration. When calculating the expectation, one should sum across all the possible permutations of the discrete variables. Let $\hat{S}_t^{ij}$ be the conditional expectation of S_t^{ij} (the j-element of S_t^i). The elements of $\Lambda_t(\theta^{\ell-1})$ and $\hat{S}_t(\theta^{\ell-1})$, can be deduced from the following calculations:

$$\begin{aligned}\hat{S}_t^{ij} = \Pr(S_t^{ij}=1|y_t, x_t; \theta^{\ell-1}) &= \frac{\Pr(y_t, S_t^{ij}=1|x_t;\theta^{\ell-1})}{\Pr(y_t|x_t;\theta^{\ell-1})} = \frac{p_{ij}\Pr(y_t|x_t, S_t^{ij}=1;\theta^{\ell-1})}{\sum_{\{S_t^{ij}\}}\Pr(y_t, S_t^{ij}|x_t;\theta^{\ell-1})}\\ &= \frac{p_{ij}\sum_{\{S_t^{rj}\}_{r\neq 1}}\Pr(y_t, \{S_t^{rj}\}_{r\neq i}|x_t, S_t^{ij}=1;\theta^{\ell-1})}{\sum_{\{S_t^{ij}\}}\Pr(y_t, S_t^{ij}|x_t;\theta^{\ell-1})}\end{aligned} \tag{16}$$

$$\begin{aligned}\Lambda_{mnt}(\theta^{\ell-1}) = E(S_t^{mj}S_j^{nj}|y_t, x_t;\theta^{\ell-1})_{m\neq n} &= \Pr(S_t^{mj}=1, S_t^{nj}=1|y_t, x_t;\theta^{\ell-1})_{m\neq n}\\ &= \frac{\Pr(S_t^{mj}=1, S_t^{nj}=1, y_t|x_t;\theta^{\ell-1})}{\Pr(y_t|x_t;\theta^{\ell-1})}\\ &= \frac{p_{mj}p_{nj}\Pr(y_t|x_t, S_t^{mj}=1, S_t^{nj}=1;\theta^{\ell-1})}{\Pr(y_t|x_t;\theta^{\ell-1})}\\ &= \frac{p_{mj}p_{nj}\sum_{\{S_t^{rj}\}_{r\neq m,n}}\Pr(y_t, \{S_t^{rj}\}_{r\neq m,n}|x_t, S_t^{mj}=1, S_t^{nj}=1;\theta^{\ell-1})}{\Pr(y_t|x_t;\theta^{\ell-1})}\end{aligned} \tag{17}$$

DOI: 10.1002/for

Given an initial guess of the parameters, we repeat the procedure until $\|\theta^{\ell-1} - \theta^{\ell-1}\|$ and $L_T(\theta^{\ell}) - L_T(\theta^{\ell-1})$ are smaller than some pre-specified tolerance level. Several different starting values should be used, and the maximum likelihood estimate of the model parameters will correspond to that associated with the largest value of the log-likelihood function that was obtained from the different starting values.

COMBINING CONDITIONAL VOLATILITY FORECASTS USING AN ESR MODEL: AN APPLICATION TO EXCHANGE RATE DATA

We examine the problem of using a set of forecasts to generate a single forecast. The motivation for using a combination of forecasts stems from the low forecasting ability of models in general. This is not surprising because, in practice, forecasting models are intentional abstractions of a much more complex reality. By combining individual forecasts based on different specifications and/or information sets, we can improve our forecasts. Furthermore, as was pointed out by Diebold and Lopez (1996), one could refine and improve the model as more and more information becomes available. This approach might be the correct one to take in the long run. However, in the short run, the cost of getting more information is very high; hence we focus on the forecasts of the models rather than on the models themselves. In economics, one can find numerous examples for the use of a combination of forecasts; see Clemen (1989) for a review.

A common approach for combining forecasts is the simple average of the individual forecasts which according to Clemen (1989) tends to outperform more complicated combining methods. Another method of combining forecasts suggested by Granger and Ramanathan (1984) is a linear regression on a set of forecasts, where the dependent variable is the true value. Other combination methods, such as the Bayesian time varying weight methods Min and Zellner (1993), have been proposed as well. However, for the forecasting horizon we investigate in this work, the same authors have demonstrated that there is no substantial benefit to using Bayesian techniques rather than linear regression.

In this work we consider the use of two alternative methods: the combination of forecasts using a switching regression (SR) model and the linear ESR model (ESR). We compare these methods with two common approaches to combining forecasts: (a) average of the individual forecasts (AVERAGE); and (b) a linear regression where the coefficients are estimated by ordinary least squares (OLS).

We expect the ESR model to perform better relative to the traditional combining methods since this model would account for situations where the 'best' forecasting model switches over time, which implies that one should change the weighting scheme for each of the individual forecasts over time. However, if we assume that all the change-points are the same, the ESR model reduces to the traditional switching regression model in which all the combining weights shift at the same point in time. Thus the switching regression model incorporates a potentially binding constraint because all the individual combining weights are mutually dependent on one state variable. For example, the ESR combining method might take into account situations where the performance of forecasters employing simple dynamic models might deteriorate in times where the economy is undergoing drastic changes, whereas forecasts based on more 'fundamental' models might fare better under such circumstances and vice versa in other circumstances.

The forecasts of the daily volatility in several exchange rates are considered as the basis for an application of our model. The exchange rate data consist of noon (New York time) buying rates for the Japanese yen (JPY), the British pound (GBP) and the Swiss franc (CHF). All rates are against the US dollar (USD). The data are for the time period 1 January 1989 to 30 December 1995 (1736

DOI: 10.1002/for

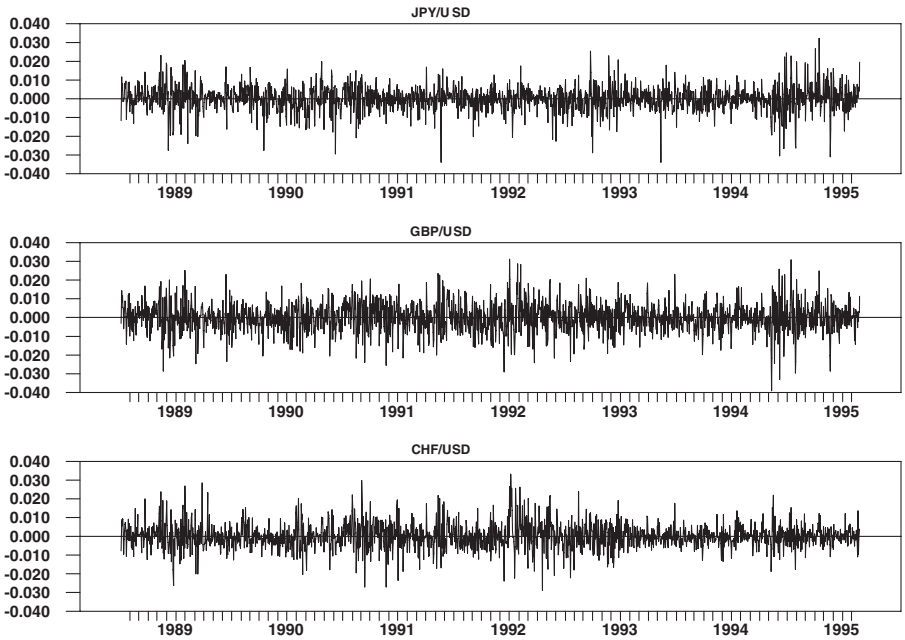

Figure 2. The return of the currencies against the dollar

observations). Let le_t denote the logarithm of a spot exchange rate at time t. We concentrate on the exchange rate change $r_t = le_t - le_{t-1}$, so that r_t is the rate of depreciation (or appreciation) of the domestic currency over time; the r_t series are plotted in Figure 2.

The exchange rates are modeled as a random walk. Meese and Rogoff (1983) and MacDonald and Taylor (1992) stress the empirical superiority of the random walk model over structural models of exchange rates determination, particularly in the short run. We follow this approach and assume that $r_t = \mu + \varepsilon_t$ where ε_t is the error term and μ is the mean return. The conditional variance of the error term, σ_t, is approximated by the square of the error term.

Next, we generate volatility forecasts by two common models. The first model is the GARCH (1,1), in which the current conditional variance of the currency's return depends on the lagged squared error term of the return, and the conditional variance in the previous period; thus $\sigma_t^2 = a_0 + a_1\sigma_{t-1}^2 + a_2\varepsilon_{t-1}^2$. The model (GARCH) parameters $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$, $\hat{\mu}$ are estimated jointly by maximum likelihood methods assuming conditional normality. The second model is the MAV model, which defines the conditional volatility as a simple average of lagged squared of the error terms:

$$\sigma_t^2 = \frac{1}{H}\sum_{h=1}^{H}\hat{\varepsilon}_{t-h}^2$$

with H chosen to minimize the Schwarz criterion (1978), where $\hat{\varepsilon}_{t-h}^2 = (r_{t-h} - \hat{\mu})^2$ are the estimated lagged squared residuals. Our choice is motivated by the fact that these models are widely employed in the financial econometric literature. Hansen and Lunde (2005) compare the out-of-sample performances of 330 different volatility models to the GARCH (1,1) model, concluding: 'Interestingly, the best models do not provide a significantly better forecast than the GARCH (1, 1) model.' See

DOI: 10.1002/for

also Donaldson and Kamstra (1997) and Pagan and Schwert (1990) for similar results regarding the simple GARCH and MAV models, respectively.

The individual volatility forecasts produced by the GARCH and MAV models were combined through the following equation:

$$F_t = \alpha + \beta_1(s_t^1) f_{1t} + \beta_2(s_t^2) f_{2t} \tag{18}$$

where $s_t^i \in \{1, \ldots, k_i\}$, $1 \leq k_i \leq \bar{k}$, $i = 1, 2$, F_t is the combined volatility forecast and f_{1t}, f_{2t} are the individual forecasts representing the GARCH and MAV models, respectively, for time t. $\beta_1(s_t^1)$, $\beta_2(s_t^2)$ are the weights which are given to each forecast and depend on the realization of the latent state variables s_t^i, which can assume one of k_i values with probability p_{ij}. Note that equation (18) provides the simple average (AVERAGE) as a forecast combining method when $k_1 = k_2 = 1$, $\alpha = 0$, $\beta_{11} = \beta_{21} = 0.5$, and when $k_1 = k_2 = 1$ the combining method is based on the ordinary least squares (OLS) estimates of α, β_{11}, β_{21}. In order to obtain a combination of forecasts using parameters derived from the SR model we will have to impose $s_t^1 = s_t^2$ on the equation above, while when we use the ESR model as a combining procedure we impose no such restrictions. Note that when $k_i > 1$ performing the prediction requires us to substitute $\beta_1(s_t^1)$, $\beta_2(s_t^2)$ by their means, because we usually would not know the realization of the state variables.

The data is split into three subsamples; the first subsample contains observations from 1 January 1989 to 30 December 1991, which are used to estimate the parameters of the conditional volatility models (GARCH, MAV). Each model then produces a one-step-ahead volatility forecast for the first trading day of 1992. The first observation from 1989 is dropped and the first observation from 1992 is added to the estimation set and the parameters of the conditional volatility models are again estimated and used to give a one-step-ahead volatility forecast for the second trading day. This procedure continues until we get volatility forecasts from 1 January 1992 to 30 December 1995. The second subsample is from 1 January 1992 to 30 December 1994 and it is used to estimate the parameters of the OLS forecast combining model as well as the parameters of the ESR and SR models. Given our estimated models, we obtain one-step-ahead forecasts for the period 1 January 1995 to 30 December 1995 by combining the individual forecasts of GARCH and MAV using the combining methods described above.

Data analysis

In order to assess the distributional properties of the data, various descriptive statistics are reported in Table I, including mean, standard deviation, skewness, kurtosis and other statistics. Visual inspection of the series, which are presented in Figure 2, reveals no evidence of serial correlation, although the conditional variances are characterized by typical 'volatility clustering', that is, periods of high volatility followed by periods of low volatility. In particular, the hypothesis of normality is rejected for each exchange rate, using the Bera and Jarque (1982) joint test of normality (BJ test). Further evidence on the nature of deviations from normality may be derived from the sample skewness and kurtosis measures. The skewness of each series is very close to zero, while the kurtosis is very large.

Table I also provides information about the autocorrelation structure of the data. It presents the first three autocorrelation coefficients (ρ_1, ρ_2, ρ_3) along with their standard errors, which reveal no evidence of serial correlation. The Ljung–Box (1978) statistic (LB(12)), for 12th serial correlation of the squared return of the exchange rate, implies a significant relationship. Bollerslev (1987) interprets the high autocorrelation of the squared data as a sign of conditional heteroskedasticity.

DOI: 10.1002/for

Table I. Summary statistics for the daily return data in the period 1989–1995

Statistic	JPY/USD	BP/USD	CHF/USD
Mean	–0.0002	–0.0001	0.0001
SD	0.0063	0.0077	0.0071
Skewness	–0.4803	–0.0131	0.3156
Kurtosis	5.6797	3.7781	5.0411
BJ test	501.07	37.48	282.23
Maximum	0.0254	0.0311	0.0331
Q3	0.0036	0.0046	0.0038
Median	0.0000	0.0001	0.0000
Q1	–0.0034	–0.0048	–0.0039
Minimum	–0.0339	–0.0290	–0.0289
ρ_1	0.0365	0.0379	0.0841
	(0.0260)	(0.0260)	(0.0260)
ρ_2	–0.0103	–0.0334	–0.0213
	(0.0260)	(0.0260)	(0.0261)
ρ_3	–0.0227	–0.0129	0.0003
	(0.0260)	(0.0260)	(0.0262)
LB(12)	42.53	43.99	94.58

Note: The data are from 1 January 1989 to 30 December 1995 (1736 observations). Q1 and Q3 are the first and third quartile, respectively, and BJ test is the Bera and Jarque (1982) joint test of normality, which is based on skewness and kurtosis and follows chi-square distribution with two degrees of freedom. ρ_1, ρ_2 and ρ_3 are the coefficients of the autocorrelation with 1, 2 and 3 lags along with their standard errors in parentheses. LB(12) is the Ljung and Box (1978) test estimated for the 12th serial correlation for the squared returns of the data.

Model selection

When applying the SR and ESR models to real data, the actual number of states for each state variable is unknown. Unfortunately, the standard likelihood ratio statistic for testing the null hypothesis of N states against the alternative hypothesis of $N + 1$ states for each state variable is not distributed chi-square asymptotically, since under the null hypothesis the parameters that describe the $N + 1$-th state are unidentifiable. A common approach in the literature to determine the appropriate number of states is to use a class of information criteria. We consider the Bayes information criterion (BIC), which was shown to be useful in other cases of switching regression models (see Le *et al.*, 1996; Li and Wong, 2000).

In order to identify the number of states in the SR and ESR models, we estimate each model, using the data in the second subsample, assuming that each state variable can assume any number of states between two and six, and choose the model with the highest BIC value. Note that in the SR model a state is 'common' to all the forecasters, whereas in the ESR model the states are for each forecaster.

Table II presents the number of states, the values of the maximized log-likelihood function and the BIC for the models we selected. For example, for the JPY/USD volatility series in the best SR model according to the BIC, both volatility forecasters can assume one of three weights. In the best ESR model the weight of the first forecaster can assume one of five values, whereas the weight of the second assumes one of two values. Since this criterion can compare rival, non-nested models, we also see that in-sample the ESR model is strongly preferred to the SR model. The EM algorithm, developed in the previous section, was used for estimation.

DOI: 10.1002/for

Table II. Values of the log-likelihood and the BIC statistic for the chosen SR and ESR models for the period 1992–1994 ($T = 747$)

Exchange rate	Model	L	Log-likelihood	BIC	No. of states
JPY/USD	SR	11	6,336	12,599	3
	ESR	16	6,621	13,137	2, 5
GBP/USD	SR	14	6,181	12,270	4
	ESR	18	6,265	12,411	4, 4
CHF/USD	SR	11	6,191	12,310	3
	ESR	12	6,355	12,631	3, 2

Note: The data are from 1 January 1992 to 30 December 1994 (747 observations). For the SR model we calculated the BIC for $k \in \{1, \ldots, 6\}$ and for the ESR model this statistic is calculated for $k_1, k_2 \in \{1, \ldots, 6\}$. L is the number of parameters for the chosen model for each currency.

Table III. Root mean squared error and mean absolute error for the out-of-sample from 1 January 1995 to 30 December 1995

Model	JPY/USD		GBP/USD		CHF/USD			
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE*	MAE*
GARCH	1.67E-04	8.37E-05	2.49E-04	1.25E-04	7.9E-05	3.96E-05	103.3%	99.1%
MAV	1.68E-04	1.02E-04	2.63E-04	1.44E-04	7.86E-05	4.64E-05	105.3%	116.8%
AVERAGE	1.63E-04	9.01E-05	2.52E-04	1.30E-04	7.66E-05	4.02E-05	101.9%	103.5%
OLS	1.78E-04	9.03E-05	2.38E-04	1.19E-04	7.11E-05	4.11E-05	100.5%	100.9%
SR	1.63E-04	8.65E-05	2.50E-04	1.25E-04	7.68E-05	4.66E-05	101.7%	105.4%
ESR	1.67E-04	8.65E-05	2.36E-04	1.12E-04	7.57E-05	4.46E-05	100.0%	100.0%

Note: RMSE is the squared root of the mean squared deviation between the volatility forecast by our combining methods and the actual volatility. MAE is the mean absolute deviation of the volatility forecasted by our combining method and the actual volatility.

* RMSE and MAE are reported as a ratio to that of the ESR model.

Assessing the forecasting ability of the models

We compare the forecasts by using two common evaluation measures for assessing the predictive accuracy of forecasting models. The measures are the root mean squared error (RMSE) and the mean absolute error (MAE). The results in Table III show these statistics for the period 1 January 1995 to 30 December 1995. In terms of the RMSE, we see that in the JPY/USD data the AVERAGE and the SR model do equally well, dominating other models. Whereas in the GBP/USD data the ESR model outperforms the other models and in the CHF/USD data, the OLS has the lowest RMSE. Therefore, on the basis of the RMSE, the results do not indicate any clear preference for any currency.

Usage of the MAE criteria also does not reveal any clear dominance relationship among the models for all exchange rates. Therefore, we calculate the RMSE and MAE relative to a benchmark model, which is the ESR model for each currency, and average the results for each model across the different currencies. The results, which are presented in percentage points, indicate that according to RMSE the ESR model is better than other combining models, while according to the MAE the GARCH model is preferable.

DOI: 10.1002/for

The encompassing test

The forecasting evaluation measures discussed above cannot determine whether a given forecasting model is in fact 'significantly' better than another. In order to evaluate the statistical significance of rival models, we conduct a Chong and Hendry (1986) forecast encompassing test. Applications of this test for the out-of-sample comparison of forecasts in financial markets can be found in Donaldson and Kamstra (1996, 1997) and Darrat and Zhong (2000). To clarify the notion of forecast encompassing, note that the forecast error from a correctly specified model should be orthogonal to any additional information available to the forecaster. Thus, a model claiming to congruently represent the data-generating process must be able to account for the salient features of rival models. In more specific terms, model k encompasses model j if model k can explain what model j cannot explain, without model j being able to explain what model k cannot explain.

The encompassing tests are therefore based on a set of linear regressions of the forecast error from one model on the forecast from the other model. Thus, with $(\hat{\varepsilon}_t^2 - \hat{\sigma}_{jt}^2)$ and $(\hat{\varepsilon}_t^2 - \hat{\sigma}_{kt}^2)$ being the forecast errors from model j and model k respectively, and $\hat{\sigma}_{jt}^2$, $\hat{\sigma}_{kt}^2$ being the forecasts of the two models, we test the significance of the δ_{jk} and π_{kj} coefficients in the following regressions:

$$(\hat{\varepsilon}_t^2 - \hat{\sigma}_{jt}^2) = \lambda_1 + \delta_{jk}\hat{\sigma}_{kt}^2 + \upsilon_{1t} \tag{19}$$

$$(\hat{\varepsilon}_t^2 - \hat{\sigma}_{jt}^2) = \lambda_2 + \pi_{kj}\hat{\sigma}_{jt}^2 + \upsilon_{2t} \tag{20}$$

in which υ_{1t} and υ_{2t} are random errors.

The null hypothesis is that neither model encompasses the other. If δ_{jk} is not significant at some predetermined level, but π_{kj} is significant, we reject the null hypothesis in favor of the alternative hypothesis that model j encompasses model k. Conversely, if δ_{jk} is significant but π_{kj} is not significant, we say that model k encompasses model j. If both δ_{jk} and π_{kj} are not significant, or if δ_{jk} and π_{kj} are both significant, we accept the null hypothesis that neither model encompasses the other.

Our findings are reported in Table IV. Columns 2–6 of the table contain the marginal p-values associated with robust consistent t-statistics based on computation of heteroskedasticity-consistent standard errors (White, 1980) where p-values less than 0.10 indicate that the forecast from the model listed along the top of the table explains, with 10% significance level, the forecast error from the model listed down the left side of the table, and thus the model listed down the side cannot encompass the model listed along the top, at the 10% significance level. For example, for the JPY/USD case the p-value of 0.0897 in the MAV row and in the SR column indicates that the SR model's forecast of the volatility in JPY/USD data explains the MAV model's forecast error at the 10% significance level. Conversely, the p-value of 0.7788 in the SR column and in the MAV row reveals that the SR forecast error could not be explained by the SR model's forecast at the 10% significance level. Therefore, for the case of the JPY\USD, the SR model encompasses the MAV model. The bold numbers in the table indicate that the model in the column encompasses the model in the row at the 10% significance level.

The results from the encompassing test show that the other models, at the 10% significance level, do not encompass the ESR model, whereas the rival models were dominated by other models at least once. For example, ESR encompasses MAV in the JPY/USD data. ESR also encompasses MAV in the GBP/USD data and OLS in the CHF/USD data. Therefore, we conclude, based on the encompassing tests, that ESR is significantly preferred to other forecast combining methods.

DOI: 10.1002/for

Table IV. Results for the out-of-sample encompassing tests

Exchange rate	Forecast error $\hat{\varepsilon}_t^2 - \hat{\sigma}_{jt}^2$	Forecast $\hat{\sigma}_{kt}^2$ from ↓					
	from ↓	GARCH	MAV	AVERAGE	OLS	SR	ESR
JPY/USD	GARCH	—	0.5119	0.4561	0.0002	0.4203	0.5782
	MAV	0.7451	—	**0.0058**	**0.0004**	**0.0897**	**0.0001**
	AVERAGE	0.8289	0.1646	—	0.0002	0.6433	**0.1005**
	OLS	0.0478	0.1234	0.0749	—	0.0506	0.0837
	SR	0.3486	0.7788	0.6130	0.0002	—	0.9269
	ESR	0.6606	0.1104	0.6086	0.0003	0.4282	—
GBP/USD	GARCH	—	0.3125	0.3492	0.1179	0.3931	**0.0201**
	MAV	0.1591	—	0.0005	0.9175	**0.0053**	0.8013
	AVERAGE	0.3239	0.0069	—	0.4124	**0.0791**	0.2236
	OLS	0.8913	0.3954	0.5471	—	0.6644	0.0198
	SR	0.9809	0.3004	0.5106	**0.0445**	—	**0.0301**
	ESR	0.51365	0.2083	0.5883	0.0261	0.9202	—
CHF/USD	GARCH	—	0.3293	0.4554	**0.0115**	0.5979	0.5358
	MAV	**0.0369**	—	**0.0217**	0.4496	**0.0418**	**0.0200**
	AVERAGE	0.1832	0.5404	—	0.1065	0.1719	0.4062
	OLS	0.5395	0.5437	0.6739	—	0.4814	**0.0486**
	SR	0.9919	0.2316	0.3124	**0.0054**	—	0.3727
	ESR	0.3504	0.9806	0.8634	0.7686	0.3206	—

Note: The table reports robust p-values on δ_{jk} from the OLS regression: $\left(\hat{\varepsilon}_t^2 - \hat{\sigma}_{jt}^2\right) = \lambda_1 + \delta_{jk}\hat{\sigma}_{kt}^2 + \upsilon_{1t}$, where $\hat{\sigma}_{kt}^2$ is model k's one-step-ahead forecast of the variance and $\left(\hat{\varepsilon}_t^2 - \hat{\sigma}_{jt}^2\right)$ is the out-of-sample forecasting error of model j. The results are for all the currencies.

SUMMARY AND EXTENSIONS

In this paper we propose a new class of models—the ESR models—which generalize the concept of switching regression models by allowing for several independent latent state variables to determine disjoint sets of the model parameters across time. These models constitute an important addition to modeling choices as they allow the analyst to parameterize large state spaces with a small number of parameters. This may be a desirable feature in modeling high-frequency data. The large-sample properties of the maximum likelihood estimates of our model are discussed. We develop an EM algorithm in order to estimate the parameters in a linear ESR model.

The ESR model is applied as a method for combining forecasts of exchange rate volatility. The individual forecasts used are those given by GARCH and MAV forecasting models. The results suggest that the ESR forecast combining method generally outperforms forecasts from competing forecast combining models. The significance of this result is evident from the forecast encompassing tests employed.

There are several interesting questions in the context of the extended switching regression framework, which warrant further research. In this work, we assumed that each of the state variables is independent over time. However, we can use our modeling approach in the context of a Markov regression model. In such models the state variable probabilities change according to the value of an unobserved Markov process. The ESR approach may also be used to model general time series, such as positive valued, censored or even time series with both discrete and continuous components,

DOI: 10.1002/for

which are typical of economic data. For example, we can use our modeling approach to extend the autoregressive conditional duration (ACD) model of Engle and Russell (1998), proposed to model the spacing between consecutive financial transactions. These extensions leave several interesting and challenging areas for future research.

ACKNOWLEDGEMENTS

The authors thank Luc Bauwens, Zvi Eckstein, Ezra Einy, Offer Lieberman, and Farshid Vahid for several insightful comments and suggestions, and seminar participants in the FFM 2004 conference, Amsterdam University, CORE-UCL in Louvain la Neuve, Ben-Gurion University and Tel Aviv University. Preminger gratefully acknowledges research support from the Kreitman Foundation. Sharon Rubin provided excellent editorial assistance.

REFERENCES

Bauer H. 1972. *Probability Theory and Elements of Measure Theory*. Holt, Rinehart & Winston: New York.

Bera AK, Jarque CM. 1982. Model specification tests: a simultaneous approach. *Journal of Econometrics* **20**: 59–82.

Billingsley P. 1961. *Statistical Inference for Markov Processes*. Holt: New York.

Bollerslev T. 1987. A conditional heteroskedastic time series model for speculation prices and rates of return. *Review of Economics and Statistics* **69**: 542–547.

Bollerslev T, Chou RY, Kroner KF. 1992. ARCH modeling in finance: a review of the theory and empirical evidence. *Journal of Econometrics* **52**: 5–59.

Chong YY, Hendry DF. 1986. Econometric evaluation of linear macro-economics models. *Review of Economic Studies* **53**: 671–690.

Clemen RT. 1989. Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* **5**: 559–583.

Darrat AF, Zhong M. 2000. On testing the random-walk hypothesis: a model-comparison approach. *Financial Review* **35**: 105–124.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**: 1–38.

Diebold FX, Lopez JA. 1996. Forecast evaluation and combination. In *Handbook of Statistics*, Maddala GS, and Rao CR (eds). North-Holland: Amsterdam; 241–268.

Domowitz I, White H. 1982. Misspecified models with dependent observations. *Journal of Econometrics* **20**: 35–58.

Domowitz I, White H. 1984. Nonlinear regression with dependent observations. *Econometrica* **52**: 143–161.

Donaldson RG, Kamstra M. 1996. Forecast combining with neural networks. *Journal of Forecasting* **15**: 49–61.

Donaldson RG, Kamstra M. 1997. An artificial neural network-GARCH model for international stock returns volatility. *Journal of Empirical Finance* **4**: 17–46.

Engel C, Hamilton JD. 1990. Long swings in the dollar: are they in the data and do markets know it? *American Economic Review* **80**: 689–713.

Engle RF, Russell JR. 1998. Forecasting transaction rates: the autoregressive conditional duration model. *Econometrica* **66**: 1127–1162.

Gallant AR, White H. 1988. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell: Oxford.

Granger CWJ, Ramanathan R. 1984. Improved methods of combining forecasts. *Journal of Forecasting* **3**: 197–204.

Hamilton JD. 1990. Analysis of time series subject to changes in regime. *Journal of Econometrics* **45**: 39–70.

Hamilton JD. 1994. *Time Series Analysis*. Princeton University Press: Princeton, NJ.

DOI: 10.1002/for

Hansen PR, Lunde A. 2005. A Forecast comparison of volatility models: does anything beat a GARCH (1,1)? *Journal of Applied Econometrics* **20**: 873–889.
Jeffries ON. 1998. Logistic mixture of generalized linear model time series. PhD Dissertation, Maryland, University of Maryland at College Park.
Klaassen F. 2005. Long swings in exchange rates: are they really in the data? *Journal of Business and Economic Statistics* **23**: 87–95.
Lanne M, Saikkonen P. 2003. Modeling the US short-term interest rate by mixture autoregressive processes. *Journal of Financial Econometrics* **1**: 96–125.
Le ND, Martin RD, Raftery AE. 1996. Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association* **91**: 1504–1514.
Li WK, Wong CS. 2000. On a mixture autoregressive model. *Journal of the Royal Statistical Society B* **62**: 95–115.
Li WK, Wong CS. 2001. On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association* **96**: 982–995.
Ljung G, Box G. 1978. On a measure of lack of fit in time series models. *Biometrika* **65**: 297–303.
MacDonald R, Taylor PM. 1992. Exchange rate economics: a survey. *IMF Staff Papers* **39**: 1–57.
Maddala GS, Kim IM. 1998. *Unit Roots, Cointegration and Structural Change*. Cambridge University Press: Cambridge, UK.
McLachlan GJ, Krishnan T. 1997. *The EM Algorithm and Extensions*. Wiley: New York.
Meese RA, Rogoff K. 1983. Empirical exchange rate models of the seventies: do they fit out of sample. *Journal of International Economics* **14**: 2–24.
Min C, Zellner A. 1993. Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics* **56**: 89–118.
Pagan AR, Schwert GW. 1990. Alternative models for conditional stock volatility. *Journal of Econometrics* **45**: 267–290.
Preminger A, Ben-Zion U, Wettstein D. 2003. Extended switching regression models: allowing for multiple latent state variables. Working Paper 03–08, Monaster Center for Economic Research, Ben-Gurion University of the Negev.
Rahbek A, Shephard N. 2002. Autoregressive conditional root model. Working paper, Nuffield College, University of Oxford.
Render RA. 1981. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Annals of Statistics* **9**: 225–228.
Render RA, Walker HF. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**: 195–239.
Schwarz G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**: 461–464.
White H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**: 817–838.
White H. 1994. *Estimation, Inference and Specification Analysis*. Cambridge University Press: New York.
Wu CFJ. 1983. On the convergence properties of the EM algorithm. *Annals of Statistics* **11**: 95–103.

Authors' biographies:
Arie Preminger is a research fellow in the Center of Operations Research and Econometrics in Université Catholique de Louvain. His current research focuses on volatility models.

Uri Ben-Zion is a professor in the Department of Economics at Ben-Gurion University of the Negev, Beer-Sheva, Israel, and a visiting professor at Università Bocconi, Milan, Italy. His research interests include financial economics, forecasting and behavioral finance.

David Wettstein is a professor in the Department of Economics at Ben-Gurion University of the Negev, Beer-Sheva, Israel. His research interests include regime switching models, game theory and auctions.

Authors' addresses:
Arie Preminger, CORE Université Catholique de Louvain, Louvain-la-Neuve B-1348, Belgium.

Uri Ben-Zion and **David Wettstein**, Department of Economics, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel.

DOI: 10.1002/for