

**EXTENDED SWITCHING REGRESSION
MODELS WITH TIME VARYING
PROBABILITIES FOR COMBINING
FORECASTS**

Arie Preminger, Uri Ben-Zion and
David Wettstein

Discussion Paper No. 03-13

November 2003

Monaster Center for Economic Research
Ben-Gurion University of the Negev
P.O. Box 653
Beer Sheva, Israel

Fax: 972-8-6472941
Tel: 972-8-6472286

Extended Switching Regression Models with Time Varying Probabilities for Combining Forecasts

by

Arie Preminger, Uri Ben-Zion, David Wettstein*

Department of Economics, Monaster Center for Economic Research

Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

Abstract

In this paper we introduce a new methodology which extends the well known switching regression model. The extension is via the introduction of several latent state variables, each one of which influencing a disjoint set of the model parameters. Furthermore, the probability distribution of the state variables is allowed to vary over time. We also develop an EM algorithm for estimation and provide mild conditions for consistency and asymptotic normality of the model parameters. This model is called the time varying extended switching regression (TV-ESR) model. We use it to combine volatility forecasts of several currencies (JPY/USD, GBP/USD, and CHF/USD). We perform a detailed comparison of the forecasts generated by the TV-ESR approach to those of traditional linear combining procedures and other methods for combining forecasts derived from the switching regression model. On the basis of out-of-sample forecast encompassing tests as well as other measures for forecasting accuracy, our results indicate that the use of this new method yields overall better forecasts than those generated by competing models.

JEL classification: C51, C53, C61

Keywords: EM algorithm, Forecast combining, TV-ESR Models

(*) The authors thank Ezra Einy for several insightful comments and suggestions and the seminar participants of the FFM 2003 conference and the Technion – Israel institute of Technology for their comments, which improve this paper. We are grateful to the Federal Reserve Bank of St. Louis and BIS for providing access to their databases. Preminger gratefully acknowledges research support from the Kreitman Foundation.

1. Introduction

Many economic and financial time series are characterized by periods in which the behavior of the series seems to change quite dramatically. Such apparent changes are often captured through the use of models with time varying parameters. One notable class of models is given by the switching regression models where the whole set of parameters moves over a finite number of value sets. The switches between the sets of values are controlled by an unobserved state variable. Important methodological contributions include the work of Quandt (1958, 1960) and the more recent work of Le et al. (1996) and Li and Wong (2000), which extend the switching regression model to the case of dependent data or more specifically an autoregression. A related class of models are the hidden Markov model regressions (see Hamilton 1994), which differ from the switching regression models in that the unobserved state variable follows a latent Markov structure

The switching regression models assume that there is only one state variable in each period which controls the switches in the model parameters. 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 a subset of the model parameters to be mutually dependent on the state variable.

Preminger et al. (2003) introduce the Extended Switching Regression (ESR) model, which is characterized by several state variables that independently influence the model selection procedure through the picking of a partial and disjoint group of the model parameters. The advantage of formulating state variables in such a way is that interesting qualitative information may result from the nature of the state variables. Furthermore, the assumption of independence among the state variables allows us to provide a parsimonious parameterization of the model, while expanding the possible number of states the model can assume.

In the ESR model, we assume that the state probabilities are constant, whereas economic as well as financial considerations suggest the desirability of allowing the state probabilities to vary over time. In this paper, we introduce the time varying extended switching regression, TV-ESR, model in which the probability distribution changes over time. The ESR models are nested in the TV-ESR models; this implies

that the usage of the TV-ESR models may allow for better description and prediction for the variables of interest. A related class of models is the switching regression model (i.e. models with one state variable) with time varying probabilities, proposed by Li and Wong (2001) and which can be traced back to Goldfeld and Quandt (1972).

We apply the TV-ESR models to combine forecasts of exchange rate volatility and investigate the incremental value of going from the traditional linear forecast combining methods and other methods for combining forecasts, derived from switching regression models, to the class of TV-ESR models. The motivation for using such models is that when one performs 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 are associated with the realization of several independent latent state variables.

A detailed comparison of the forecast combination method on the basis of the root mean squared error (RMSE), root mean absolute error (RMAE), Theil-U statistic, the correct direction change prediction and the forecast encompassing tests, is performed. We show that the TV-ESR modeling procedure performs at least as well and often better than forecasts based on rival combining methods

The layout of this paper is as follows: In Section 2 we describe our model with relation to switching regression models. In Section 3 we address the large sample properties of the model and establish the consistency and asymptotic normality of the maximum likelihood estimates. In Section 4 we develop the relevant version of an EM algorithm for estimation. In Section 5 we report the results of a simulation study. In Section 6 we present our forecast combining models, one of which is the TV-ESR model. In section 7 we describe the data set and discuss our empirical findings and Section 8 concludes.

2. The model

Switching regression models were developed as a way of allowing data to arise from a combination of two or more distinct data generation processes, which depend on the realization of s_t , an unobserved random discrete variable, which will be called a state variable. Therefore, in the switching regression model we have:

$$(1) \quad y_t = \mu_t(w_t, \psi(s_t)) + \varepsilon_t$$

where $\mu_t : W \times \Psi \rightarrow R$ are known functions measurable on W for each $\psi(s_t)$ in Ψ , a compact parameter set in R^d and continuous on Ψ almost surely for all t , and the error term ε_t is a zero-mean white noise, s_t is a state variable that can assume one of k integer values $\{1, 2, \dots, k\}$ and $w_t \in W$ is a vector of explanatory variables. The function $\psi(s_t) \in \Psi$ associates with each realization of the state variable, a parameter vector which is chosen from the set $\{\psi_1, \psi_2, \dots, \psi_k\}$.

On the other hand, in the extended switching regression (ESR) model which was proposed by Preminger et al. (2003), we assume the existence of p discrete switches in disjoint groups of the model parameters. The changes in the i -th group depend only on the realization of s_t^i , an unobserved i.i.d state variable which can assume one of k integer values $\{1, 2, \dots, k\}$. The ESR model can be described as follows:

$$(2) \quad y_t = \mu_t(w_t, \psi_1(s_t^1), \dots, \psi_p(s_t^p)) + \varepsilon_t$$

where $\mu_t : W \times \Psi_1 \cdots \times \Psi_p \rightarrow R$ are known functions measurable on X for each $\psi_i(s_t^i)$ in Ψ_i a compact subset and continuous on $\Psi = \times_{i=1}^p \Psi_i$ almost surely for all t . In other words, the ESR is characterized by p independent selections from the disjoint parameter sets. From each set we select one element among k possible ones. The selection is random and dependent on the realization of the latent state variables. A concise comparison between the latent structures of the ESR and the switching regression (SR) models is given in figure 1.

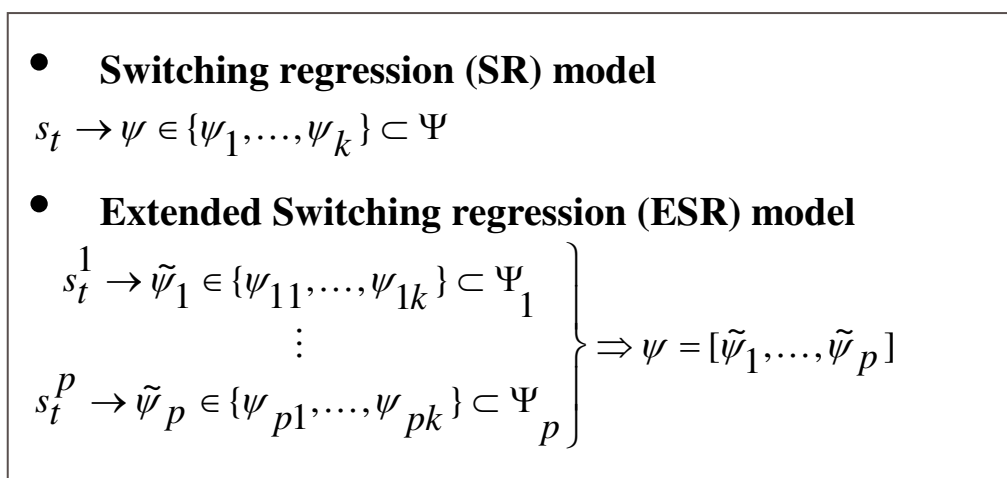


Figure 1: The latent structure of the ESR model and the SR model

From figure 1 we can see that the ESR models have the advantage that they take into account situations where disjoint groups of the model parameters change independently over time. This assumption allows us to provide a parsimonious parameterization of the model while allowing us to consider a variety of k^p states which our model can assume, and hence describe more efficiently structural changes in the data. However the probabilities governing these changes are constant over time. The time varying ESR (TV-ESR) model relaxes this assumption.

In addition, we assume for simplicity, and without loss of generality, that $k = 2$. The distribution of the state variable is assumed to be time varying, and given by logistic functions:

$$(3) \quad \Pr(s_t^i = 1 | \tilde{z}_{it}; \gamma_i) = \frac{\exp(\tilde{z}_{it}' \gamma_i)}{1 + \exp(\tilde{z}_{it}' \gamma_i)} \quad i = 1, \dots, p$$

where the $(l_i \times 1)$ conditioning vector \tilde{z}_{it} contains explanatory variables that affect the state probabilities and $\gamma_i \in \Gamma_i$ is an unknown set of logistic function parameters. It will be convenient to stack the parameters governing the probabilities of the state variables into one vector, $\gamma = [\gamma'_1, \dots, \gamma'_p]'$ $\in \times_{i=1}^p \Gamma_i$, and let \tilde{z}_t be a vector containing all the information in the \tilde{z}_{it} 's. It is obvious, but worth noting, that when the last $(l_i - 1)$ terms of the $(l_i \times 1)$ state probability vector γ_i , are set to zero, the probabilities of the state vector are time-invariant and this model reduces to the ESR model.

The logistic modeling we add allows us to incorporate non-constant probabilities and model the effects \tilde{z}_{it} may have on the probability of the state variables, which determine a subset of the model parameters. This is an important and useful extension to the ESR model. Adding the flexibility of time varying probabilities should improve the model's ability to describe the data. The TV-ESR model we consider is given by:

$$(4) \quad y_t = \sum_{i=1}^d \beta_{it} w_{it} + \varepsilon_t$$

where $\varepsilon_t \sim i.i.N(0, \sigma)$, $\beta_{it} \in \{\beta_{i1}, \beta_{i2}\}$ and w_{it} are explanatory variables which could include lagged variables of y_t . The model's slopes are determined by d latent state variables which are distributed according to equation (3). The conditional distribution of y_t may be obtained as follows:

(5)

$$f(y_t | w_t, \tilde{z}_t; \theta) = \sum_{j_1=1}^2 \cdots \sum_{j_d=1}^2 \left(\prod_{i=1}^d \Pr(s_t^i = j_i | \tilde{z}_t; \gamma_i) \right) \cdot f(y_t | w_t, \{s_t^i = j_i\}; \{\beta_{i1}, \beta_{i2}\}, \sigma)$$

$j_i \in \{1,2\}$ and $\theta = [\beta_{11}, \beta_{12}, \dots, \beta_{d1}, \beta_{d2}, \gamma, \sigma]$ is the vector of all model parameters and under the normality assumption, the density of y_t conditional upon $x_t, \{s_t^i\}_{i=1}^d$ is given by

$$(6) \quad f(y_t | x_t, \{s_t^i = j_i\}; \{\beta_{i1}, \beta_{i2}\}, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y_t - \sum_{i=1}^d \beta_{ij_i} x_{it})^2}{2\sigma^2} \right\}$$

$\beta_{ij_i} \in \{\beta_{i1}, \beta_{i2}\}$. The log-likelihood function of the sample is:

$$(7) \quad L_T(\theta) = \frac{1}{T} \sum_{t=1}^T \log f(y_t | w_t, \tilde{z}_t; \theta)$$

The maximum likelihood estimates (MLE) are obtained by maximizing equation (7) with respect to the unknown model parameters. We will obtain under mild regularity assumptions the consistency and the asymptotic normality of the estimates in the next section. The prediction of y_t the conditional on $\{w_t, \tilde{z}_t, \{s_t^i\}_{i=1}^d\}$ is straightforward from the conditional density in equation (6). The prediction is given by:

$$(8) \quad E(y_t | \{s_t^i = j_i\}_{i=1}^d, x_t, \tilde{z}_t; \theta) = \sum_{i=1}^d \beta_{j_i} w_{it} \quad j_i \in \{1,2\}, \quad i = 1, \dots, d$$

Since the state variables are not observable, there are 2^d conditional predictions associated with 2^d possible states. Therefore, the unconditional prediction based on the data, is calculated as follows: (9)

$$\begin{aligned} E(y_t | w_t, \tilde{z}_t; \theta) &= \int y_t \cdot f(y_t | w_t, \tilde{z}_t; \theta) dy_t = \int y_t \cdot \left(\sum_{j_1=1}^2 \cdots \sum_{j_d=1}^2 \Pr(y_t, \{s_t^i = j_i\}_{i=1}^d | w_t, \tilde{z}_t; \theta) \right) dy_t \\ &= \int y_t \cdot \left(\sum_{j_1=1}^2 \cdots \sum_{j_d=1}^2 f(y_t | \{s_t^i = j_i\}_{i=1}^d, w_t, \tilde{z}_t; \theta) \cdot \prod_{i=1}^d \Pr(s_t^i = j_i | w_t, \tilde{z}_t; \theta) \right) dy_t \\ &= \sum_{j_1=1}^2 \cdots \sum_{j_d=1}^2 \left(\prod_{i=1}^d \Pr(s_t^i = j_i | w_t, \tilde{z}_t; \theta) \right) \cdot \int y_t \cdot \left(f(y_t | \{s_t^i = j_i\}_{i=1}^d, w_t, \tilde{z}_t; \theta) \right) dy_t \\ &= \sum_{j_1=1}^2 \cdots \sum_{j_d=1}^2 \left\{ \left(\prod_{i=1}^d \Pr(s_t^i = j_i | w_t, \tilde{z}_t; \theta) \right) \cdot E(y_t | \{s_t^i = j_i\}_{i=1}^d, w_t, \tilde{z}_t; \theta) \right\} \end{aligned}$$

Thus, the prediction for y_t is a weighted average of the predictions given in equation (8). In addition it is a nonlinear function of the observations although our model is linear, since the state probabilities depend nonlinearly on the data. In the simple case where the state variables are perfectly correlated our model reduces to the time varying switching regression model (Li and Wong (2001)).

3. Asymptotic Theory

Typically in most applied work estimation and inference are conducted conditional upon some information set, and usually the analyst tries to find the conditional density of the “dependent” data (see, e.g. Voung (1983)). In our analysis we assume the conditional density is specified as a general TV- ESR model. The observed data are a realization of a stochastic process $\{Z_t : \Omega \rightarrow R^v, t=1,2,\dots\}$ on a complete probability space $(\Omega, \mathfrak{F}, P_0)$ where $\Omega = \times_{t=1}^{\infty} R^v$ and \mathfrak{F} is the Borel σ -field generated

by measurable finite dimensional product cylinders and P_0 is the probability measure governing the behavior of the data. Also let \mathfrak{F}_t be the σ -field generated by current and past Z_t , i.e. $\mathfrak{F}_t = \sigma(\dots, Z_{t-1}, Z_t)$, where $\mathfrak{F}_{t-1} \subset \mathfrak{F}_t \dots \subset \mathfrak{F}$. The vector Z_t can be partitioned into $Z_t = (Y_t, X_t)$ where Y_t is the dependent variable and X_t is the $1 \times \ell$ dimensional vector of explanatory ("exogenous") variables where $v = 1 + \ell$. We are interested in a parametric family of conditional probability distributions

$\{P_t^\psi(y_t | \bar{\mathfrak{F}}_{t-1}; \psi) : \psi \in \Psi, \bar{\mathfrak{F}}_{t-1} \subset \mathfrak{F}_t\}$, which exists, by Jirina's theorem (see, Bauer (1972), p.319), where $\bar{\mathfrak{F}}_{t-1} \equiv \sigma(Z_{t-\tau}, \dots, Z_{t-1}, X_t) = \sigma(w_t)$, $\tau < \infty$, which imply that our model explicitly includes only a finite numbers of lags $Z_t = (Y_t, X_t)$.

Assumption 1:

- a) The random vectors $\{Z_t\}_{t \in \mathbb{N}}$ are strictly stationary and ergodic.
- b) The family of conditional probability distributions have Radon-Nikodym densities $f(y_t | w_t, \psi) \equiv dP_t^\psi(y_t | w_t; \psi) / d\nu$ which are measurable-
 $\sigma(y_t, w_t) \subset \bar{\mathfrak{F}}_t$ for every ψ in Ψ and for all t.

Note that w_t are the variables that the analyst has chosen for the purpose of explaining or forecasting y_t , which might include X_t and lagged values of the dependent variables. Now, the ESR model is characterized by several unobserved selection processes which are characterized by p independent selections from the parameter sets $\{\Psi_i\}$. From each set the selection is random and dependent on the realization of the state variables $\{s_t^i\}$, with the probability distribution $\left\{\Pr(s_t^i = j \mid \tilde{z}_{it}, \gamma_i)\right\}_{j=1}^k$. The state probabilities are not constant where the vector \tilde{z}_{it} contains explanatory variables that are made up of known measurable functions of w_t and $\gamma_i \in \Gamma_i$ is the parameter vector.

Let $\varphi_i \subset \Psi_i$ be a set of k distinct values chosen by s_t^i , where each element of this set is denoted by φ_{ij} with probability $\Pr(s_t^i = j_i \mid \tilde{z}_{it}, \gamma_i)$ where $j_i \in \{1, \dots, k\}$.

The conditional density of y_t can be described by:

$$(10) \quad f(y_t \mid w_t, \theta) = \sum_{j_1=1}^k \cdots \sum_{j_p=1}^k \left[\left(\prod_{i=1}^p \Pr(s_t^i = j_i \mid \tilde{z}_{it}, \gamma_i) \right) \cdot f(y_t \mid w_t, \varphi_{1j_1}, \dots, \varphi_{pj_p}) \right]$$

$\theta = (\varphi_{11}, \dots, \varphi_{1k}, \dots, \varphi_{p1}, \dots, \varphi_{pk}, \gamma_1, \gamma_2, \dots, \gamma_p) \in \Phi \subset R^L$ is the vector of the model parameters. Hence, the likelihood function of the sample is

$$(11) \quad L_T(\theta) = \frac{1}{T} \sum_{t=1}^T \log f(y_t \mid w_t; \theta)$$

We define the maximum likelihood estimator (MLE) as a parameter vector $\hat{\theta}_T$ which maximizes the likelihood function.

Assumption 2:

(a) The conditional density $f(y_t \mid w_t, \psi)$ is continuous on Ψ a compact subset of R^L for each (y_t, w_t) a.s. P_0 .

(b) For all $i \in \{1, \dots, p\}, j = \{1, \dots, k-1\}$, the functions $\left\{\Pr(s_t^i = j \mid \tilde{z}_{it}, \gamma_i)\right\}$ are measurable- $\sigma(\tilde{z}_{it})$ for each γ_i in Γ_i a compact set and are continuous on Γ_i , for each (\tilde{z}_{it}) a.s. P_0 .

Under assumptions 1 and 2 we prove in theorem 1 that the maximum likelihood estimator of the ESR model above is a \mathfrak{F}_T -measurable function of the data. Notice

that since, the likelihood function for each observation is a convex combination of likelihood functions given the state variables, our assumptions are sufficient in order for the sample likelihood function to satisfy the standard continuity and measurability conditions, which are needed to establish the measurability of the MLE (see White (1994)). This result establishes that $\hat{\theta}_T$ is a random variable, and therefore has stochastic properties (consistency, asymptotic distribution) that will be proven in the sequel of this section.

Theorem 1: Given assumptions 1-2 there exists a measurable MLE $\hat{\theta}_T$.

Proof: See Appendix A.

We should also note that in the TV-ESR model, as in the ESR model the log-likelihood function attains its maximum at several different choices of θ obtained from the true parameters θ_0 by “label switching” (see, Render (1981), Render and Walker (1984)). Therefore the global identifiability condition, necessary for the consistency of the estimator is violated.

$$(12) \quad \theta^* \in C = \{\theta \in \Phi \mid f(y_t \mid w_t, \theta) = f(y_t \mid w_t, \theta_0)\}$$

We will prove that when the MLE is "close" enough to θ^* , we can obtain the convergence of the estimator to one of the elements in C. In the TV-ESR model, if we have p state variables where each variable can assume k values, and if the true parameters are θ_0 and the distribution of the data under each state came from the same parametric family there exist $(k!)^p$ distinct values, which give the same likelihood. In order to show that $\hat{\theta}_T$ is the MLE for θ^* , we impose the following additional condition.

Assumption 3:

- a) $|\log f(y_t \mid w_t, \theta)| \leq m(y_t, w_t)$ for all $\theta \in \Theta$ and for each (y_t, w_t) a.s. P_0 ,
and $E(m(y_t, w_t)) < \Delta < \infty$.

b) For each $\theta_i \in C$ and all $\varepsilon > 0$

$$[\max_{\theta \in \bar{\eta}(\varepsilon)} E(\log f(y_t | w_t, \theta) - E(\log f(y_t | w_t, \theta_i))] < 0, \text{ where}$$

$$\bar{\eta}(\varepsilon) = \left(\bigcup_{i=1}^{\#C} \eta_i(\varepsilon) \cap \Theta \right) \text{ and } \eta_i(\varepsilon) \text{ be an open sphere centered at } \theta_i.$$

Assumption 3(a) ensures the uniform convergence of the sample log likelihood function as a result of corollary 3.1 of Newey (1991). These conditions can be verified if the log-likelihood function is continuously differentiable on an open, convex set containing the parameter set. Assumption 3(b) is an identification requirement which guarantees that the log-likelihood function does not become increasingly flat in the neighborhood of θ^* . This implies that the number of states, we assume before estimation, should not exceed the correct number. Since, any model nesting the true TV-ESR model might violate this identification requirement. 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, then its corresponding probability parameters are not identified.

Theorem 2: Let $\hat{\theta}_T$ be the maximum likelihood estimator, under assumptions 1-3

$\hat{\theta}_T \rightarrow \theta^* \in C$ almost surely.

Proof: See Appendix A.

Next, we introduce the conditions, which ensure the asymptotic normality of our MLE. This property is established by Taylor's expansion of the first order conditions around the true parameter vector and using sufficient conditions in order to apply the central limit theorem for the score function and the uniform law of large numbers for the information matrix. In the TV-ESR model, the idea is to show that if the MLE is "close" to one of the elements in C then $\sqrt{T}(\hat{\theta}_T - \theta^*)$ is distributed asymptotically normal.

Let $D_T(\theta)$ and $H_T(\theta)$ denote the gradient and the Hessian of the log likelihood function

i.e. $D_T(\theta) = \frac{1}{T} \sum_{t=1}^T \partial \log f(y_t | w_t, \theta) / \partial \theta$, $H_T(\theta) = \frac{1}{T} \sum_{t=1}^T \partial^2 \log f(y_t | w_t, \theta) / \partial \theta \cdot \partial \theta'$ and

let $H(\theta) = E \left[\partial^2 \log f(y_t | w_t, \theta) / \partial \theta \cdot \partial \theta' \right]$ and $Q(\theta) = -H(\theta)^{-1}$.

Assumption 4:

a) The elements of $|\partial \log f(y_t | w_t, \theta) / \partial \theta \cdot \partial \log f(y_t | w_t, \theta) / \partial \theta|$ and the elements of $|\partial^2 \log f(y_t | w_t, \theta) / \partial \theta \cdot \partial \theta'|$ are dominated by P_0 -integrable functions independent of the parameter set.

b) $\theta^* \in C$ is in the interior of Θ , and the matrix $H(\theta)$ exists and is nonsingular and positive definite in some open neighborhood of θ^* in C .

c) The elements of $|\partial f(y_t | w_t, \theta) / \partial \theta|$ and $|\partial^2 f(y_t | w_t, \theta) / \partial \theta \cdot \partial \theta'|$ are dominated by P_0 -integrable functions independent of the parameters.

Theorem 3: Let $\hat{\theta}_T$ be a consistent sequence of MLE of $\theta^* \in C$ then under assumptions 1-4 we have that

$$\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} N(\underline{0}, Q(\theta^*))$$

where $-H_T(\hat{\theta}_T)^{-1} \rightarrow Q(\theta^*)$ almost surely.

Proof: See Appendix A.

Assumption 4 imposes mild moment and smoothness conditions which guarantee the asymptotic normality of each element in $\sqrt{T}(\hat{\theta}_T - \theta^*)$. These conditions are satisfied if the sample log-likelihood function is three times continuously differentiable with respect to the parameter set, with derivatives that can be uniformly bounded by integrable functions (see Andrews (1987)). We also assume that V_T converges to an invertible matrix where we should note that, this assumption will be violated when the estimated model is a degenerate TV-ESR version of the true model as we mention above.

4. Estimation

In order to estimate the model parameters we use the EM algorithm introduced by Dempster, Laird, and Rubin (1977), see also Wu (1983) and McLachlan and Krishnan (1997). 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. There are pathological constructions in which the EM estimates may converge to a critical point other than a local maximum see Wu (1983), but such aberrations are usually overcome by changing the starting values of the algorithm.

Suppose that the observations $\{y_t, w_t', \tilde{z}_t\}$ are generated as in equation (4). Let $S_t = \{S_t^1, \dots, S_t^d\}'$ where S_t^i is a two-dimensional unobserved vector $(S_t^{i1}, 1 - S_t^{i1})$, with S_t^{i1} equal one if $s_t^i = 1$ and zero if $s_t^i = 2$ and let $B = (\beta_{11}, \beta_{12}, \dots, \beta_{d1}, \beta_{d2})'$ and $W_t = (w_{1t}, w_{1t}, w_{2t}, w_{2t}, \dots, w_{dt}, w_{dt})'$. We can write (4) more succinctly as:

$$(13) \quad y_t = (B * W_t)' \cdot S_t + \varepsilon_t$$

The symbol "*" denotes the Hadamard product, which means element-by-element multiplication. The log-likelihood function of the complete data thus, assuming that the values of the state variables are known, is given by: (14)

$$Q_T(\theta) = \sum_{t=1}^T \left(\sum_{i=1}^d [S_t^{i1} \log p_{i1}(\tilde{z}_{it}, \gamma_i) + (1 - S_t^{i1}) \log(1 - p_{i1}(\tilde{z}_{it}, \gamma_i))] - \sum_{i=1}^d \sum_{j=1}^2 \frac{(y_t - (B * W_t)' \cdot S_t^j)^2}{2\sigma^2} - \log(2\pi\sigma^2) \right)$$

where $p_{i1}(\tilde{z}_{it}, \gamma_i) = \Pr(s_t^i = 1 | \tilde{z}_{it}, \gamma_i)$. The iterative EM procedure for estimating the model parameters consists of an E-step in which the expectation of equation (14) is taken with respect to the distribution of the state variables given the data and the parameters estimated in the previous iteration, and an M-step where a new set of parameters is generated through the maximization of the expectation. In the E-step we get:

$$(15) \quad E(Q_T(\theta^\ell | \{y_t, w_t, \tilde{z}_t\}_{t=1}^T, \theta^{\ell-1})) = \frac{-1}{2\sigma^2} \sum_{t=1}^T \{y_t^2 - 2y_t \cdot (B * W_t)' \cdot \hat{S}_t(\theta^{\ell-1}) + (B * W_t)' \cdot \Lambda_t(\theta^{\ell-1}) \cdot (B * W_t)\} + \sum_{t=1}^T \sum_{i=1}^d [\hat{S}_t^{i1} \log p_i(\tilde{z}_{it}, \gamma_i) + (1 - \hat{S}_t^{i1}) \cdot \log(1 - p(\tilde{z}_{it}, \gamma_i))] - T \log(2\pi\sigma^2)$$

where $\theta^\ell = [B^\ell, \sigma^\ell, \gamma_1^\ell, \dots, \gamma_d^\ell]$ denotes the parameters estimated in the ℓ -th iteration and $\Lambda_t(\theta^{\ell-1}) = E(S_t \cdot S_t' | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})$, $\hat{S}_t(\theta^{\ell-1}) = E(S_t | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})$ and \hat{S}_t^{ij} be the conditional expectation of S_t^{ij} (the j -element of the vector S_t^i). The elements of $\Lambda_t(\theta^{\ell-1})$, $\hat{S}_t(\theta^{\ell-1})$, can be deduced from the following calculations:

$$(16) \quad \hat{S}_t^{i1} = \Pr(S_t^{i1} = 1 | y_t, \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1}) = \frac{\Pr(y_t, S_t^{i1} = 1 | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})}{\Pr(y_t | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})} = \frac{p_{i1}(\cdot) \Pr(y_t | S_t^{i1} = 1, \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})}{\Pr(y_t | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})} =$$

$$= \frac{p_{i1}(\cdot) \sum_{\{S_t^{rj}\}_{r \neq i}} \Pr(y_t, \{S_t^{rj}\}_{r \neq i} | S_t^{i1} = 1, \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})}{\Pr(y_t | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})}$$

$$(17) \quad E(S_t^{mj} S_t^{nj} | y_t, \bar{\mathfrak{S}}_{t-1})_{m \neq n} = \Pr(S_t^{mj} = 1, S_t^{nj} = 1 | y_t, \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})_{m \neq n} =$$

$$= \frac{\Pr(S_t^{mj} = 1, S_t^{nj} = 1, y_t | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})}{\Pr(y_t | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})} = \frac{p_{mj}(\cdot) \cdot p_{nj}(\cdot) \cdot \Pr(y_t | \bar{\mathfrak{S}}_{t-1}, S_t^{mj} = 1, S_t^{nj} = 1; \theta^{\ell-1})}{\Pr(y_t | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})}$$

$$= \frac{p_{mj}(\cdot) \cdot p_{nj}(\cdot) \sum_{\{S_t^{rj}\}_{r \neq m, n}} \Pr(y_t, \{S_t^{rj}\}_{r \neq m, n} | \bar{\mathfrak{S}}_{t-1}, S_t^{mj} = 1, S_t^{nj} = 1; \theta^{\ell-1})}{\Pr(y_t | \bar{\mathfrak{S}}_{t-1}; \theta^{\ell-1})}$$

Next, we perform the M-step, in which the estimates of the parameters are obtained by maximizing (15) above. Solution of the first order conditions yields parameter estimates for B and σ given by:

$$(18) \quad B^\ell = \left(\sum_{t=1}^T W_t \cdot W_t' * \Lambda_t(\theta^{\ell-1}) \right)^{-1} \sum_{t=1}^T y_t W_t * \hat{S}_t(\theta^{\ell-1})$$

$$(19) \quad \sigma^\ell = \sqrt{\frac{1}{T} \sum_{t=1}^T \left[y_t^2 - 2y_t (B^\ell * W_t)' \hat{S}_t(\theta^{\ell-1}) + (B^\ell * W_t)' \Lambda_t(\theta^{\ell-1}) (B^\ell * W_t) \right]}.$$

By differentiating equation (15) with respect to γ_i we get

$$(20) \quad \sum_{t=1}^T (\hat{S}_t^{i1} - p_{i1}(\tilde{z}_{it}, \gamma_i)) \tilde{z}_{it} = \underline{0}$$

In order to solve the nonlinear equation (20) for γ_i , we will use the Newton-Raphson method. That is, given initial guess γ_i^0 , the values of γ_i in the subsequent iteration are given by

$$(21) \quad \gamma_i^{h+1} = \gamma_i^h + \left(\sum_{t=1}^T p_{i1}(\tilde{z}_{it}, \gamma_i^h) (1 - p_{i1}(\tilde{z}_{it}, \gamma_i^h)) \cdot \tilde{z}_{it} \cdot \tilde{z}_{it} \right)^{-1} \sum_{t=1}^T (\hat{S}_t^{i1} - p_{i1}(\tilde{z}_{it}, \gamma_i^h)) \tilde{z}_{it}$$

For $h = 1$, we set $\gamma_i^h = \gamma_i^{\ell-1}$ (our initial guess equals the estimated parameters in the previous iteration) and update the values of γ_i according to equation (21) until convergence and for this h , we set $\gamma_i^h = \gamma_i^\ell$. Starting from an initial guess of the model parameters, we repeat the procedure until $\|\theta^\ell - \theta^{\ell-1}\|$ 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 estimates will correspond to those associated with the largest value of the likelihood function that was obtained from the different starting values.

5. A Simulation Study

We now report the results of a simulation study designed to evaluate the EM estimation methods. In this simulation experiment, 1000 independent sample paths were generated from the TV-ESR model described in equation (4) where we set $d = 2$ and $x_{1t}, x_{2t}, \varepsilon_t$ were generated independently from standard normal distribution. The parameters of the model are given as follows: $(\alpha, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \sigma) = (1.15, 2, 3, 4, 5, 1)$. The logistic function in the form of (3) contains only slope and intercept parameters with \tilde{z}_{it} as an explanatory variable, which was also drawn from a standard normal distribution, thus

$$(22) \quad \Pr(s_t^i = 1 | \tilde{z}_{it}) = \frac{\exp(\gamma_{0i} + \gamma_{1i} \tilde{z}_{it})}{1 + \exp(\gamma_{0i} + \gamma_{1i} \tilde{z}_{it})} \quad i = 1, 2$$

where $(\gamma_{0i}, \gamma_{1i}) = (1, 2)$ for each i . For each observation we generate the Bernoulli random variables, s_t^i , with means given by the probabilities in equation (22). If $s_t^i = 1$ we set $\beta_{it} = \beta_{i1}$ otherwise we set $\beta_{it} = \beta_{i2}$ and draw y_t from the proper distribution.

Table 1: Results of the simulation study for the TV-ESR model

Parameter	True value	Average	Empirical SE	Theoretical SE
α	1.15	1.15	0.05	0.05
β_{11}	2.00	1.97	0.13	0.09
β_{12}	3.00	3.04	0.21	0.14
β_{21}	4.00	3.97	0.08	0.09
β_{22}	5.00	5.03	0.13	0.14
σ	1.00	0.97	0.04	0.04
γ_{01}	1.00	0.93	0.55	0.54
γ_{11}	2.00	1.89	0.63	0.54
γ_{02}	1.00	0.91	0.53	0.54
γ_{12}	2.00	1.87	0.63	0.48

For each sample, we estimate the parameters using the EM algorithm. Each simulation produces not only the model parameters, but also generates estimates of their standard errors. The standard errors for each simulation are averaged, and the resulting means are included in table 1 under the title of “Theoretical SE”. A second method for estimating the standard errors involves calculating the standard deviation of the sample of model parameters, obtained through simulations. These standard errors are denoted in table 1 as “Empirical SE”.

The results presented in table 1 show that the EM estimation method has performed relatively well in general with reasonable standard errors, but the bias in estimating the parameters of the logistic function is relatively high. These results indicate that adding the possibility of time varying probabilities, might result in some

bias (relatively to the other model parameters) in the parameters related to these probabilities.

However, this bias should become negligible the larger sample we use. The theoretical standard errors are very close to the empirical standard errors, but we can observe a small downward bias, which is expected to disappear in large sample situations.

6. The TV-ESR model for combining conditional volatility forecasts

Forecasting the exchange rate volatility has been a challenging area of research ever since the collapse of the Bretton Woods system of fixed parties. Volatility is important to policy makers and financial market participants since it can be used as a measure of risk. From the theoretical perspective, volatility plays a central role in pricing of derivative securities. Furthermore, for the purpose of forecasting, confidence intervals may be time varying so that more accurate intervals can be obtained by modeling volatility of returns.

There is a vast literature on forecasting volatility, and many econometric models which are likely to be misspecified have been used. However, no single model was found to be superior as was noted by Hu and Tsoukalas (1999). This raises the question: How should volatility forecasts, if at all, be combined into a single forecast with better forecasting performance. The rationale behind forecast combining, as was pointed out by Diebold (2001), is that in practice, forecasting models are intentional abstractions of a much more complex reality. By combining individual forecasts based on different specification and/or information sets, we can improve our forecasts. A common approach for combining forecasts is the simple average of the individual forecasts, which according to Clemen (1986) 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 see 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 in using Bayesian techniques rather than linear regression.

In this work, we will illustrate the use of the TV-ESR models as a forecast combining tool for exchange rates volatility forecasts and compare their performances to other common combining methods. We employ five alternative models of combining forecasts given by: the simple average of the individual forecasts (AVERAGE), the linear regression where the coefficients are estimated by Ordinary Least Squares (OLS) the switching regression (SR) model with constant and time varying probabilities, the extended switching regression (ESR) model and the time varying extended switching regression (TV-ESR) model.

We expect the ESR model to perform better relative to the traditional combining methods because the ESR model would account for situations where the "best" model switches over time, which implies that one should change the weighting scheme for each of the individual forecasts over time. When the probabilities of such switches are time varying, the TV-ESR model might be more flexible than the ESR and perform better as a combining method.

We consider two common models for forecasting the conditional volatility. The GARCH (1,1) and the moving average variance (MAV) model. The GARCH (1,1) assumes that the current conditional volatility of the currency's return depends on the lagged squared error term of the return, and the conditional volatility in the previous period. That is $\sigma_t^2 = \omega + \lambda\sigma_{t-1}^2 + \delta\varepsilon_{t-1}^2$; where σ_t^2 and ε_t^2 are the conditional variance and the squared error term of the return at time t respectively. The model (GARCH) parameters are estimated jointly by maximum likelihood methods assuming conditional normality. Previous studies show that lower order GARCH models in general and the GARCH (1,1) in particular provide a parsimonious representation of the temporal dependencies in the conditional volatility see e.g. Bollerslev et al. (1992). The second model we use is the MAV model (for details, see Pagan and Schwert, 1990) where the volatility is modeled as a simple average of the lagged squared error terms: $\sigma_t^2 = \frac{1}{H} \sum_{h=1}^H \varepsilon_{t-h}^2$, where H , the number of lags, is chosen to minimize the Schwarz Criterion (1978).

In order to combine our individual volatility forecasts we estimate the following TV-ESR model:

$$(23) \quad \sigma_t^2 = \alpha + \beta_{1t} \hat{\sigma}_{1t}^2 + \beta_{2t} \hat{\sigma}_{2t}^2 + \varepsilon_t$$

where ε_t is a Gaussian white noise, $\beta_{it} \in \{\beta_{i1}, \beta_{i2}\}$ for $i = 1, 2$ and σ_t^2 is the actual volatility for time t and $\hat{\sigma}_{1t}^2, \hat{\sigma}_{2t}^2$ are the individual forecasts for the GARCH and MAV models respectively. The weights which are given to each individual forecast are determined by the realization of its latent state variable, s_t^i , where the probabilities of the state variables are evolving according to equation (3). Using the estimates of the TV-ESR model, described in equation (23), the two individual forecasts are combined through equations (8)-(9). The EM algorithm, which is discussed above, is used for estimation. The application of TV-ESR models for combining forecasts requires modeling the dynamics which characterize the probabilities of the state variables. In the absence of a specific theory for modeling, we consider two simple alternative models for the dynamics.

In the first model we assume that the probability distribution of the state variables is a function of the sign of the lagged return of the exchange rate, thus the probabilities which correspond to each of the state variable, respond asymmetrically to changes in past returns. This variable was shown to be important in modeling conditional mean and the conditional variance in exchange rate markets (e.g. see Laopodis, 2001) and in stock markets (e.g. see Koutos, 1998). Therefore, it seems reasonable to assume that this variable influences unobserved state variables of the model. We denote this model as TV-ESR1 and the state probabilities are given as follows:

$$(24) \quad \Pr(s_t^i = 1 | r_{t-1,i}; \eta_{0i}, \eta_{1i}) = \frac{\exp(\eta_{0i} + \eta_{1i} \cdot \text{sign}(r_{t-1}))}{1 + \exp(\eta_{0i} + \eta_{1i} \cdot \text{sign}(r_{t-1}))} \quad i = 1, 2$$

In the second model we assume that the probability distribution of the state variables depends on the performance of our individual forecasts relative to each other. The performance is measured by the absolute forecast error of the forecasting models. Such a model could be useful since information about the forecast errors could aid considerably in determining the states. We denote this model as TV-ESR2. Let e_{1t}, e_{2t} denote the forecast errors of the first (GARCH) and second (MAV) forecasting models respectively, the state variable probabilities are then given by

$$(25) \quad \Pr(s_t^i = 1 | I_{t-1}; \gamma_{0i}, \gamma_{1i}) = \frac{\exp(\gamma_{0i} + \gamma_{1i} \cdot I_{t-1})}{1 + \exp(\gamma_{0i} + \gamma_{1i} \cdot I_{t-1})} \quad i = 1, 2$$

$$I_t = 1 \text{ when } \left| \frac{e_{t,1}}{e_{t,2}} \right| > 1 \quad I_t = 0, \text{ otherwise.}$$

Now, when the combination of forecasts is derived from the usage of switching regression models, with time varying probabilities, we will estimate the parameters of equations (23-24) under the restriction that $s_t^1 = s_t^2$. Thus, there exists only one state variable which influences simultaneously the weights given to each of the individual forecast. This model is denoted as TV-SR1. In the second model we consider under this restriction, the case where the probability distribution of the state variable changes according to equation (25); we call this model TV-SR2. The estimation results for all our combining methods are presented in Appendix B.

7. The Results

7.1 The data and the estimation procedure

The exchange rate data, used in the empirical analysis, consist of noon bid rates on the Japanese yen (JPY), the British pound (GBP) and the Swiss franc (CHF). All rates are against the US dollar (USD). Let E_t be the exchange rate at time t where the returns are measured as $r_t = \ln(E_t / E_{t-1})$ and are calculated for each exchange rate on a weekly frequency. Our exchange rate data is from the first week of January 1980 to the last week of December 1996. These rates have been taken from BIS database, giving us 889 observations per exchange rate. Each r_t series is plotted in figure 2 below. In order to calculate the volatility we follow the standard approach suggested by Pagan and Schwert (1990) and Day and Lewis (1992). A proxy for the true volatility is given by $(r_t - \bar{r})^2$, where \bar{r} is the average return over the sample period.

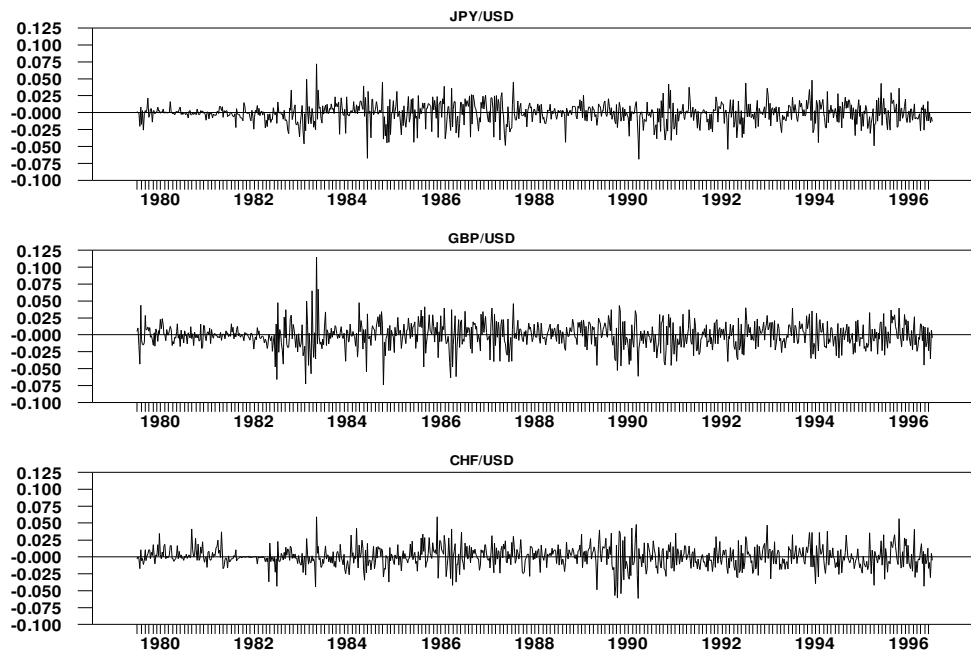


Figure 2: The rate of depreciation (or appreciation) of the currencies against the dollar

The exchange rates are modeled as a random walk, in line with many extensive empirical studies (see among others, Meese and Rogoff (1983); Diebold and Nerlove (1989), West and Cho (1995)) and an analysis of the data. 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.

Table 2: Summary Statistics for the weekly data in the period 1980-1996

Statistic	JPY/USD	GBP/USD	CHF/USD
Mean	-0.0010	-0.0007	0.0003
Std. Dev.	0.0155	0.0187	0.0155
Skewness	-0.2748	-0.0237	0.0161
Kurtosis	4.8740	5.5668	4.5129
BJ_test	141.2800	244.1200	84.8200
Maximum	0.0719	0.1148	0.0594
Q3	0.0075	0.0108	0.0092
Median	-0.0001	-0.0007	-0.0006
Q1	-0.0085	-0.0110	-0.0082
Minimum	-0.0687	-0.0741	-0.0614
r1	-0.0083 (0.335)	-0.0620 (0.335)	-0.0012 (0.340)
r2	0.1112 (0.335)	0.0628 (0.337)	0.0024 (0.338)
r3	0.0088 (0.335)	0.0460 (0.335)	0.0616 (0.335)
LB1(24)	31.3429	24.1347	13.8146
LB2(12)	39.5171	101.8860	62.9393

(*) Q1 and Q3 are the first and third quartile respectively and BJ test is the Bera and Jarque test (1982) joint test of normality that is based on skewness and kurtosis and follows chi-square distribution with two degrees of freedom. r1, r2 and r3 are the first three autocorrelations along with their standard errors in parentheses. LB1(24) is the Ljung and Box (1978) test for the 24th serial correlation. LB2(12) is the same test estimated for the 12th serial correlation for the squared returns of our data.

The descriptive statistics, presented in table 2, clearly indicate that all the series have excessive kurtosis and asymmetry and the Jarque-Bera test (1982) strongly rejects the normality hypothesis for all series. The first three autocorrelations (r1, r2, r3) along with their standard errors (in parentheses), calculated for each exchange rate, indicate white noise for each series. For the joint test of autocorrelation, we compute the Ljung-Box statistic (LB1) up to the 24-th order serial correlation. Under the null hypothesis of no autocorrelation, such statistics are distributed chi-square asymptotically with twenty-four degrees of freedom. The test does not reject the white noise hypothesis for all currencies at 10% significance level. Since our currencies have a mean very close to zero, we can use the squared returns as a measure of their variance and the absolute return as a measure for a standard deviation. The squared

returns are clearly not uncorrelated over time, as reflected by the highly significant Ljung-Box (LB2) test for 12-th serial correlation, implying heteroskedasticity.

Next, the 17-year study period is split up into three sub-samples; the first sub-sample contains 574 observations from the years 1980 to 1990 which we use to estimate the MAV and the GARCH model parameters and then produce a one-step ahead volatility forecast. Next, the data is updated by adding the first week of 1991 and dropping the last week of 1980, and the model parameters are re-estimated again in order to produce a one step ahead forecast from our individual forecasts for the second week of 1991. This procedure is repeated until we get volatility forecasts for each week of the period 1991-1996. The second sub-sample, from the first week of 1991 to the last week of 1994, is used to estimate the parameters of our combining models. The third sub-sample, from the period 1995-1996 which contains 106 observations, serves as the out-of-sample forecasting period. Given our estimation results for each of our combining models, we calculated the one-step ahead combined volatility forecasts for this period.

7.2 Forecasting accuracy

We use four measures to compare the one-step-ahead volatility forecasts obtained from our models. The statistics are the Root Mean Square Error (RMSE), the Root Mean Absolute Error (RMAE) and the Theil-U statistic (Theil-U), and we also compute the "correct directional change" (CDC) statistic which measures the ability of our models to correctly predict the actual change which has subsequently occurred in the volatility. These statistics are given as follows:

$$(26) \quad RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\sigma_i^2 - \hat{\sigma}_i^2)^2}$$

$$(27) \quad RMAE = \sqrt{\frac{1}{N} \sum_{i=1}^N |\sigma_i^2 - \hat{\sigma}_i^2|}$$

$$(28) \quad Theil - U = \frac{\sum_{i=1}^N (\hat{\sigma}_i^2 - \sigma_{i-1}^2)^2}{\sum_{i=1}^N (\sigma_i^2 - \sigma_{i-1}^2)^2}$$

$$(29) \quad CDC = \frac{100}{T} \sum_{i=1}^N D_i \text{ where } D_i = \begin{cases} 1 & \text{if } (\sigma_i^2 - \sigma_{i-1}^2) \cdot (\hat{\sigma}_i^2 - \sigma_{i-1}^2) > 0 \\ 0 & \text{if } (\sigma_i^2 - \sigma_{i-1}^2) \cdot (\hat{\sigma}_i^2 - \sigma_{i-1}^2) \leq 0 \end{cases}$$

The RMSE and the RMAE are two popular measures to test the forecasting power of a model. However, these measures are not invariant to scale transformations. In the Theil-U statistic, the forecasting error is standardized by the error from a random walk. For the random walk, the Theil-U statistic equals one. Of course, the random walk is not a naïve rival, particularly in many financial and economic series; therefore a value of the Theil-U statistic close to one is not necessarily an indication of bad forecasting performance. The advantage of the Theil-U statistic is that it is independent of the scale of the variables.

When comparing different models, it can also be useful to measure the number of times a given model correctly predicts the direction of change of the actual values being forecast. Furthermore, as was noted by Dunis and Huang (2002), the RMSE, RMAE and Theil-U statistics are important measures for forecasting accuracy of the model concerned, but ignore the profitability point of view. There are, however, cases when the forecaster is less interested in the forecasting accuracy of his volatility models and cares more about the ability of these models to forecast the direction of the change right, which is an important issue in trading strategy that relies on the direction of a forecast rather than its level. The CDC statistic as a measure of the direction, addresses this issue.

Table 3: Forecasting performance of competing forecast combining methods for the period 1995-1996

		GARCH	MAV	AVERAGE	OLS	ESR	TV-ESR1	TV-ESR2	SR	TV-SR1	TV-SR2
JPY/USD	RMSE	4.43E-04	4.58E-04	4.45E-04	4.50E-04	4.42E-04	4.44E-04	4.43E-04	4.56E-04	4.56E-04	4.56E-04
	RMAE	1.66E-02	1.68E-02	1.66E-02	1.62E-02	1.64E-02	1.65E-02	1.65E-02	1.61E-02	1.60E-02	1.60E-02
	Theil-U	0.545	0.580	0.550	0.561	0.541	0.547	0.543	0.576	0.575	0.576
	CDC	65.71%	76.19%	74.29%	67.62%	70.48%	72.38%	71.43%	69.52%	68.57%	69.52%
GBP/USD	RMSE	6.83E-04	7.08E-04	6.89E-04	6.79E-04	6.85E-04	6.73E-04	6.84E-04	6.89E-04	6.85E-04	6.89E-04
	RMAE	1.89E-02	1.88E-02	1.87E-02	1.95E-02	1.94E-02	1.94E-02	1.96E-02	1.97E-02	1.96E-02	1.99E-02
	Theil-U	0.481	0.517	0.490	0.477	0.484	0.467	0.483	0.490	0.485	0.491
	CDC	83.95%	83.95%	85.08%	81.06%	83.38%	82.23%	82.23%	80.47%	81.06%	81.65%
CHF/USD	RMSE	1.72E-04	1.70E-04	1.67E-04	2.37E-04	1.87E-04	1.90E-04	1.90E-04	1.96E-04	1.97E-04	1.97E-04
	RMAE	1.17E-02	1.06E-02	1.11E-02	1.49E-02	1.27E-02	1.28E-02	1.28E-02	1.32E-02	1.32E-02	1.31E-02
	Theil-U	0.597	0.581	0.565	1.128	0.702	0.729	0.724	0.778	0.783	0.780
	CDC	79.88%	81.06%	82.23%	74.96%	78.07%	78.07%	78.68%	77.46%	77.46%	78.68%

Table 3 reports the RMSE, RMAE, Theil-U and the CDC statistics for each of the individual forecasts and our combining methods. In terms of the RMSE the ESR and TV-ESR1/TV-ESR2 models as a group show superior out-of-sample forecasting performance compared to other models, for the JPY/USD and the GBP/USD volatility. Conversely, on the basis of the RMAE criterion MAV, GARCH and the AVERAGE perform as well as, or better than the other models for the GBP/USD volatility. The TV-SR1 and TV-SR2 combination methods provide the most accurate forecasts for the JPY/USD volatility, while the SR model is ranked second. For the CHF/USD volatility, the GARCH, MAV and the AVERAGE combining methods outperform the other models under all forecast accuracy measures and they also provide the best combining models of the directional change for all the exchange rates.

Examining the Theil-U statistic shows only one model performs worse than the random walk model. This is the OLS combined forecast for the CHF/USD volatility for which the Theil-U is greater than one. The ESR and the TV-ESR models as a group, provide the best forecasting performance for the JPY/USD and GBP/USD volatility.

It is hard to identify the best combining model overall since the simple average, the GARCH and the MAV models, perform quite well under any statistic for the CHF/USD volatility. There is also no clear preference between the ESR model and the TV-ESR models, where comparisons across different measures and currencies, yield different conclusions. However, on the basis of RMSE, Theil-U and the CDC statistics, the ESR and TV-ESR emerge as the best forecast combining method in comparison to the OLS and SR models.

7.3 The encompassing test

Although useful, the forecasting evaluation measures, discussed in the previous section, 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 Kamasra (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 $(\sigma_{jt}^2 - \hat{\sigma}_{jt}^2)$ and $(\sigma_{kt}^2 - \hat{\sigma}_{kt}^2)$ 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:

$$(30) \quad (\sigma_{jt}^2 - \hat{\sigma}_{jt}^2) = \lambda_1 + \delta_{jk} \hat{\sigma}_{kt}^2 + \eta_t$$

$$(31) \quad (\sigma_{kt}^2 - \hat{\sigma}_{kt}^2) = \lambda_2 + \pi_{kj} \hat{\sigma}_{jt}^2 + \nu_t$$

in which η_t and ν_t 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 significant, we accept the null hypothesis that neither model encompasses the other.

The encompassing tests results are presented in tables 4, 5 and 6 for the JPY/USD, GBP/USD and CHF/USD exchange rates, respectively. The name of the dependent variable, the model forecasting error, is listed down the left side of the tables, while the independent variable which is the model forecast, is listed at the top of the tables. The entries of the tables contain p-values associated with the heteroskedasticity robust t-statistics (White 1980) on δ_{jk} and π_{kj} . P-values less than 0.10 indicate that the forecast from the model listed along the top of each table explains, with 10% significance, the

forecast error from the model listed down the left side of the table and thus the model listed on the left side cannot encompass the model listed on the top.

Table 4: Encompassing results (JPY/USD) - marginal significance level

Forecast error $\sigma_{jt}^2 - \hat{\sigma}_{jt}^2$ from ↓	Forecast σ_{kt}^2 from ↓									
	GARCH	MAV	AVERAGE	OLS	ESR	TV-ESR1	TV-ESR2	SR	TV-SR1	TV-SR2
GARCH	NA	0.8172	0.8989	0.5681	0.8902	0.9456	0.9413	0.3764	0.3884	0.4438
MAV	0.0102	NA	0.0065	0.0832	0.0065	0.0067	0.0074	0.7471	0.6651	0.7109
AVERAGE	0.1738	0.2054	NA	0.2415	0.1901	0.1804	0.1864	0.5448	0.5148	0.5677
OLS	0.3781	0.1529	0.2011	NA	0.1953	0.2329	0.2351	0.2679	0.3036	0.3577
ESR	0.6859	0.7667	0.7378	0.6443	NA	0.6899	0.7188	0.7046	0.7004	0.7559
TV-ESR1	0.656	0.7109	0.6895	0.6552	0.6916	NA	0.6639	0.7575	0.7083	0.807
TV-ESR2	0.7627	0.8812	0.8420	0.6554	0.8426	0.7801	NA	0.6442	0.6444	0.8048
SR	0.0822	0.0244	0.0337	0.5369	0.0325	0.0459	0.0434	NA	0.5142	0.578
TV-SR1	0.0891	0.0269	0.0371	0.5547	0.0358	0.0524	0.0476	0.4423	NA	0.5782
TV-SR2	0.0903	0.0273	0.0376	0.5592	0.0363	0.0506	0.0427	0.4392	0.5049	NA

The table reports robust p-values on δ_{jk} from the OLS regression: $(\sigma_{jt}^2 - \hat{\sigma}_{jt}^2) = \lambda_1 + \delta_{jk} \hat{\sigma}_{kt}^2 + \eta_t$ where $\hat{\sigma}_{kt}^2$ is model k 's one step ahead forecast of the variance and $(\sigma_{jt}^2 - \hat{\sigma}_{jt}^2)$ is the out-of-sample forecasting error of model j for the JPY/USD exchange rate on weekly data in the period 1995-1996 for our forecast combination methods, mentioned above.

Table 5: Encompassing results (GBP/USD) - marginal significance level

Forecast error $\sigma_{jt}^2 - \hat{\sigma}_{jt}^2$ from ↓	Forecast σ_{kt}^2 from ↓									
	GARCH	MAV	AVERAGE	OLS	ESR	TV-ESR1	TV-ESR2	SR	TV-SR1	TV-SR2
GARCH	NA	0.6333	0.5153	0.9264	0.6263	0.3327	0.0479	0.3249	0.6735	0.8357
MAV	0.0021	NA	0.0017	0.0049	0.0021	0.5421	0.0548	0.3190	0.4054	0.7421
AVERAGE	0.0394	0.0715	NA	0.1411	0.0704	0.8680	0.2675	0.9804	0.5248	0.9471
OLS	0.9458	0.8323	0.8644	NA	0.8341	0.0295	0.8279	0.7065	0.5070	0.9241
ESR	0.1766	0.2319	0.2054	0.3236	NA	0.5559	0.4792	0.9634	0.4888	0.8953
TV-ESR1	0.8102	0.5593	0.6261	0.4504	0.5630	NA	0.6139	0.3549	0.8932	0.9917
TV-ESR2	0.6341	0.4238	0.4756	0.3480	0.4266	0.0584	NA	0.3423	0.5781	0.5417
SR	0.3152	0.1339	0.1731	0.0918	0.1356	0.0233	0.298	NA	0.7002	0.8087
TV-SR1	0.3431	0.1587	0.1966	0.1134	0.1605	0.0790	0.3081	0.1516	NA	0.8775
TV-SR2	0.3462	0.1609	0.1992	0.1152	0.1627	0.0241	0.5852	0.1530	0.6484	NA

The table reports robust p-values on δ_{jk} from the OLS regression: $(\sigma_{jt}^2 - \hat{\sigma}_{jt}^2) = \lambda_1 + \delta_{jk} \hat{\sigma}_{kt}^2 + \eta_t$ where $\hat{\sigma}_{kt}^2$ is model k 's one step ahead forecast of the variance and $(\sigma_{jt}^2 - \hat{\sigma}_{jt}^2)$ is the out-of-sample forecasting error of model j for the GBP/USD exchange rate on weekly data in the period 1995-1996 for our forecast combination methods, mentioned above.

Table 6: Encompassing results (CHF/USD) - marginal significance level

Forecast error $\sigma_{jt}^2 - \hat{\sigma}_{jt}^2$ from ↓	Forecast σ_{kt}^2 from ↓									
	GARCH	MAV	AVERAGE	OLS	ESR	TV-ESR1	TV-ESR2	SR	TV-SR1	TV-SR2
GARCH	NA	0.4838	0.3071	0.8383	0.4153	0.0298	0.6564	0.5641	0.7371	0.7917
MAV	0.0067	NA	0.0019	0.0035	0.0018	0.0019	0.0552	0.0121	0.0028	0.0077
AVERAGE	0.0337	0.0538	NA	0.1146	0.0466	0.0366	0.2304	0.0639	0.0933	0.1384
OLS	0.9223	0.9757	0.9872	NA	0.9883	0.7527	0.8182	0.9626	0.9531	0.9945
ESR	0.1483	0.1471	0.1317	0.2057	NA	0.1032	0.4143	0.1574	0.1888	0.2501
TV-ESR1	0.0144	0.1444	0.1287	0.2037	0.1369	NA	0.3402	0.1549	0.1936	0.2147
TV-ESR2	0.3328	0.4396	0.3805	0.567	0.4164	0.2483	NA	0.4672	0.5391	0.1459
SR	0.1493	0.0813	0.0888	0.0941	0.0823	0.0629	0.3348	NA	0.0905	0.1357
TV-SR1	0.1038	0.0194	0.0303	0.0162	0.0202	0.0202	0.1951	0.0176	NA	0.0343
TV-SR2	0.1586	0.0431	0.0607	0.0375	0.0475	0.0297	0.0255	0.0424	0.0394	NA

The table reports robust p-values on δ_{jk} from the OLS regression: $(\sigma_{jt}^2 - \hat{\sigma}_{jt}^2) = \lambda_1 + \delta_{jk} \hat{\sigma}_{kt}^2 + \eta_t$ where $\hat{\sigma}_{kt}^2$ is model k 's one step ahead forecast of the variance and $(\sigma_{jt}^2 - \hat{\sigma}_{jt}^2)$ is the out-of-sample forecasting error of model j for the CHF/USD exchange rate on weekly data in the period 1995-1996 for our forecast combination methods mentioned above.

Consider first the results for the JPY/USD data as reported in table 4. The absence of any p-values less than 0.1 in the GARCH, AVERAGE, OLS, ESR and TV-ESR1/TV-ESR2 rows reveals that none of these models' forecast error can be explained by other models' forecasts and therefore these models are not encompassed by other models. Conversely, these models (except the OLS) can explain at 10% significant level the forecast error for the MAV, and the switching regression models. Therefore, we conclude that the GARCH, AVERAGE, ESR and the TV-ESR1/TV-ESR2 encompass the MAV, SR, TV-SR1, and TV-SR2 models.

Table 5 reports the results for the GBP/USD data. Pair-wise comparisons shows that the ESR, TV-ESR1 and the TV-ESR2 forecast error is not explained by any other model at 10% significant level therefore they are not encompassed while these models forecasts explain significantly the forecast errors from other models. Finally, table 6 reports the results for the CHF/USD data. The GARCH, OLS, ESR and the TV-ESR models are superior to other models without being encompassed.

The results from the encompassing tests reported in tables 4-6 imply that the ESR and the TV-ESR models often encompass rival models in terms of out-of-sample forecasting ability. But these models are not encompassed by rival models while every other model is encompassed at least once. Therefore, we conclude the TV-ESR which nests the ESR is significantly better than other forecast combining techniques on the basis of the encompassing tests.

8. Summary and conclusions

This paper presents a new class of models which generalize the concept of switching regression models. These models, denoted by ESR and TV-ESR, can capture occasional but recurrent independent switches in disjoint subsets of the model parameters, which are determined by latent state variables. This approach increases the number of states of the model with parsimonious parameterization. In the ESR model, the probability distribution underlying the parameters switches is constant over time, while in the TV-ESR model it is allowed to vary over time as a function of relevant explanatory variables. We show that under general conditions the maximum likelihood estimates of these models are

consistent and asymptotically normal and develop an EM algorithm in order to estimate the parameters of a linear TV-ESR model.

The new methodology is used to combine forecasts of exchange rate volatility. For simplicity, the individual forecasts used are those given by GARCH and MAV forecasting models. Alternative methods of combining forecasts that are considered are the linear regression (OLS), the simple average (AVERAGE) and the switching regression (SR) models. The forecasts obtained are compared on the basis of forecast accuracy measures and encompassing tests. The results presented suggest the ESR and TV-ESR models are overall preferred to a variety of competing models.

We should note that although the TV-ESR model nests the ESR model and the traditional linear combining methods as special cases, there is no guarantee that the TV-ESR model would dominate out-of-sample, especially if it over-fits the in-sample data. The empirical findings which suggest that the TV-ESR model performs overall better than the rival models indicate that this model should be preferred to more restrictive forecast combining methods such as the OLS or the SR models.

Further empirical work should apply these models to the study of other financial time series. This approach can also be used to estimate the relationships between financial variables, e.g. the transmission of volatility across financial markets. In addition, the approach adopted in this paper can be extended to allow for more than two latent state variables as well as other types of probability distributions for the state variables such as the Probit function. These extensions leave several interesting and challenging areas for future research.

Appendix A:

Proof of Theorem 1: Assumption 1-2 imply that $f(y_t | w_t, \theta)$ satisfies standard measurability and continuity requirements as defined in Wooldridge (1994), p. 2726. The theorem follows immediately by theorem 2.12 of White (1994), p.16. \square

Proof of Theorem 2: Under assumptions 1-3 we can establish the weak uniform law of large numbers by using theorem A.2.2 of White (1994, p.351). The likelihood function converges almost surely, uniformly for all $\theta \in \Phi$ to $L(\theta) = E(\log f(y_t | w_t; \theta))$, a continuous function on a compact set, which attains its maximum on the set C . So, for $\delta > 0$ there exists a \hat{T} such that for $T > \hat{T}$, $P(\sup_{\theta \in \Theta} |L_T(\theta) - L(\theta)| > \delta) = 0$. Now, for each point in C we select an open set N_i that contains it and does not contain any other point in the set C . We let $N = \bigcup_{i=1}^{\#C} N_i$, then $\bar{N} \cap \Phi$ where \bar{N} the complement of N is compact.

Therefore $\max_{\theta \in \bar{N} \cap \Phi} L(\theta)$ exists. Let $e = L(\theta^*) - \max_{\theta \in \bar{N} \cap \Phi} L(\theta) > 0$ for $\theta^* \in C$, by assumption 3(b). We now define the event: $E_T = \{\theta \in \Theta \mid L_T(\theta) - L(\theta) < e/2\}$. Then $E_T \Rightarrow L(\hat{\theta}_T) > L_T(\hat{\theta}_T) - e/2$ and $E_T \Rightarrow L_T(\theta^*) > L(\theta^*) - e/2$ for all $\theta^* \in C$. Since, $L_T(\hat{\theta}_T) \geq L_T(\theta^*)$ for $\theta^* \in C$ by the definition of $\hat{\theta}_T$, we have that $E_T \Rightarrow L(\hat{\theta}_T) > L_T(\theta^*) - e/2$ for $\theta^* \in C$. After adding both sides of the above-mentioned two inequalities we get: $E_T \Rightarrow L(\hat{\theta}_T) > L(\theta^*) - e \Rightarrow L(\hat{\theta}_T) > \max_{\theta \in \bar{N} \cap \Phi} L(\theta)$. Therefore we conclude that $E_T \Rightarrow \hat{\theta}_T \in \bigcup_i N_i$ and so $\Pr(E_T) \leq \Pr(\hat{\theta}_T \in \bigcup_i N_i) \leq 1$. Set $\delta = e/2$, for $T > \hat{T}$, $\Pr(E_T) = 1$, hence assumptions 1-3 imply $\hat{\theta}_T \rightarrow \theta^* \in C$ almost surely. \square

Proof of Theorem 3: The mean value expansion allows us to write

$$(A.1) \quad D_T(\hat{\theta}_T) = D_T(\theta^*) + H_T(\bar{\theta})(\hat{\theta}_T - \theta^*)$$

where $\bar{\theta}$ lies on the chord between θ_T and θ^* for $\theta^* \in C$ recalling that $D_T(\hat{\theta}_T) = \underline{0}$ we see

$$(A.2) \quad D_T(\hat{\theta}_T) = 0 = D_T(\theta^*) + [H_T(\theta^*) + H_T(\bar{\theta}) - H_T(\theta^*)]^{-1} \cdot (\hat{\theta}_T - \theta^*)$$

We want to show that the inverse of the square brackets converges almost surely to $-H(\theta^*)^{-1}$. Hence, we need to show that for any given $\delta > 0$ there exists a value \bar{T} such that for $T > \bar{T}$, $P(\|H_T(\bar{\theta}) - H_T(\theta^*)\| > \delta) = 0$.

Where $\|\cdot\|$ is the Euclidean matrix norm. Theorem 3.35 of White (2001, p.44) and assumption 1 imply that $D_T(\theta), H_T(\theta)$ are also stationary and ergodic. Therefore assumptions 1, 2, 4(a) allow us to apply theorem A.2.2 of White (1994, p.351) to conclude that $H_T(\theta)$ converges to $H(\theta)$ almost surely uniformly on Θ and $H(\theta)$ is uniformly continuous function on Θ .

$$(A.3) \quad P(\|H_T(\bar{\theta}) - H_T(\theta^*)\| > \delta) \leq \\ P(\|H_T(\bar{\theta}) - H(\bar{\theta})\| > \delta/3) + P(\|H(\bar{\theta}) - H(\theta^*)\| > \delta/3) + P(\|H(\theta^*) - H_T(\theta^*)\| > \delta/3) \leq \\ 2 \cdot P(\sup_{\theta \in \Theta} \|H_T(\theta) - H(\theta)\| > \delta/3) + P(\|H(\bar{\theta}) - H(\theta^*)\| > \delta/3)$$

For the last inequality, the uniform convergence implies there exist T_1 such that $P(\sup_{\theta \in \Theta} \|H_T(\theta) - H(\theta)\| > \delta/3) = 0$ for $T > T_1$. Now, the continuity of $H(\theta)$ implies that for all $\delta/3 > 0$ there exists $\eta(\delta/3) > 0$ such that if $d(\bar{\theta}, \theta^*) < \eta(\delta/3)$ then

$\|H(\bar{\theta}) - H(\theta^*)\| < \delta/3$ and from theorem 2 there exists a T_2 such that for $T > T_2$, $P(d(\bar{\theta}, \theta^*) < \eta(\delta/3)) = 1$ (where $d(\cdot, \cdot)$ is some metric on Θ). Therefore $P(\|H(\bar{\theta}) - H(\theta^*)\| > \delta/3) = 0$ and for $T_3 > T = \max\{T_1, T_2\}$, we see that

$$(A.4) \quad P(\|H_T(\bar{\theta}) - H_T(\theta^*)\| > \delta) = 0$$

Assumptions 1, 4(a), and the ergodic theorem imply that $[H_T(\theta^*) + o(1)] \rightarrow H(\theta^*)$ almost surely, where the convergence is element-wise. By Assumption 4(b), $H(\theta^*)$ is nonsingular so that $H_T(\theta^*)$ is nonsingular almost surely for T sufficiently large. Since the elements of the inverse matrix are continuous functions of the original matrix elements, they are Borel measurable and applying theorem 18.8 of Davidson (1994, p.286) we get almost surely that:

$$(A.5) \quad [H_T(\theta^*) + o(1)]^{-1} \rightarrow H(\theta^*)^{-1}.$$

Multiplying by \sqrt{T} and rearranging the expression above yields:

$$(A.6) \quad \sqrt{T}(\hat{\theta}_T - \theta^*) = H(\theta^*)^{-1} \sqrt{T} D_T(\theta^*)$$

Next we will show that the random vectors¹ $\{\partial \log f(y_t | w_t, \theta^*) / \partial \theta\}$ are a martingale difference sequence adapted to $\{\mathfrak{F}_t\}$

$$(A.7) \quad E\left[\partial \log f(y_t | w_t, \theta^*) / \partial \theta | \mathfrak{F}_{t-1}\right] = \int (\partial \log f(y_t | w_t, \theta^*) / \partial \theta) \cdot f(y_t | w_t, \theta) dy_t \\ = \int (\partial f(y_t | w_t, \theta^*) / \partial \theta) dy_t = \partial \left(\int f(y_t | w_t, \theta^*) dy_t \right) / \partial \theta = \underline{0}$$

By differentiating under the integral sign again we obtain the information matrix equality

$$(A.8) \quad E\left(\left(\partial \log f(y_t | w_t, \theta^*) / \partial \theta\right) \cdot \left(\partial \log f(y_t | w_t, \theta^*) / \partial \theta\right)\right) = -E\left(\partial^2 \log f(y_t | w_t, \theta^*) / \partial \theta \cdot \partial \theta\right)$$

These equalities follow since assumptions 1(b) and 4(c) allow the application of theorem 12.13 (Bartle (2001)), permitting us to interchange the differentiation and integration procedures. These equalities will be used later on.

Now, to apply the central limit theorem for the multivariate case, we use the Cramer-Wold device and the central limit theorem for martingale difference sequence; see Davidson (1994, pp.383-385). Because the data are stationary and ergodic, it is sufficient to verify that the score has a finite variance that can be estimated consistently.

Let $V = E\left(\left(\partial \log f(y_t | w_t, \theta) / \partial \theta\right) \cdot \left(\partial \log f(y_t | w_t, \theta) / \partial \theta\right)\right)$, since the data is stationary martingale difference sequence,

$$\text{var}(\sqrt{T} D_T) = \frac{1}{T} \sum_{t=1}^T E\left(\left(\partial \log f(y_t | w_t, \theta) / \partial \theta\right) \cdot \left(\partial \log f(y_t | w_t, \theta) / \partial \theta\right)\right) = V.$$

Let $\ddot{Z}_t \equiv \zeta' V^{-1/2} \cdot \partial \log f(y_t | x_t, \theta) / \partial \theta$, where $\zeta' \zeta = 1$, \ddot{Z}_t is measurable with respect to \mathfrak{F}_t given assumption 1(b) and theorem 3.26 of Davidson (1994, p.52). It follows from the linearity of the conditional expectation and that the score function is a martingale difference that $E(\ddot{Z}_t | \mathfrak{F}_{t-1}) = 0$. Hence, $\{\ddot{Z}_t, \mathfrak{F}_t\}$ is a martingale difference sequence. As a consequence of stationarity, $\text{var}(\ddot{Z}_t) = \zeta' V^{-1/2} V V^{-1/2} \zeta = 1$, for all t.

Since, $\{\zeta' V^{-1/2} \cdot \partial \log f(y_t | x_t, \theta) / \partial \theta \cdot \partial \log f(y_t | x_t, \theta) / \partial \theta \cdot V^{-1/2} \zeta\}$ is a stationary and ergodic sequence with finite expected absolute values given assumption 4(a),

¹ Note that the measurability of the derivatives follows from assumption 1(b) by using the fact that the derivatives are defined as the (measurable) limit of a sequence of (measurable) difference quotients.

Minkowski's inequality, the ergodic theorem and theorem 18.5 (Davidson 1994, p.284)

imply that

$$p \lim \frac{1}{T} \sum_{t=1}^T \zeta' V^{-1/2} \cdot \partial \log f(y_t | w_t, \theta) / \partial \theta \cdot \partial \log f(y_t | w_t, \theta)' / \partial \theta \cdot V^{-1/2} \zeta = \zeta' V^{-1/2} V V^{-1/2} \zeta = 1$$

Hence, for all ζ , $\zeta' \zeta = 1$, by the central limit theorem for martingale difference

$$(A.9) \quad T^{-1/2} \sum_{t=1}^T \ddot{Z}_t = T^{-1/2} \sum_{t=1}^T \zeta' V^{-1/2} \cdot \partial \log f(y_t | w_t, \theta) / \partial \theta \xrightarrow{D} N(0,1)$$

Using the Cramer-Wold device and the information matrix equality, proven above, we see that for each $\theta^* \in C$

$$(A.10) \quad \sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} N(\underline{0}, Q(\theta^*))$$

where $Q(\theta^*) = -H(\theta^*)^{-1}$.

Next, we show that $H_T(\hat{\theta}_T)^{-1}$ is a consistent estimator for $Q(\theta^*)$. From assumptions 1, 2, 4(a), we can establish the continuity of $H(\theta)$ and that

$\sup_{\theta \in \Theta} \|H_T(\theta) - H(\theta)\| \rightarrow 0$ almost surely. Hence,

$$(A.11) \quad \begin{aligned} \left\| H_T(\hat{\theta}_T) - H(\theta^*) \right\| &\leq \|H_T(\hat{\theta}_T) - H(\hat{\theta}_T)\| + \|H(\hat{\theta}_T) - H(\theta^*)\| \\ &\leq \sup_{\theta \in \Theta} \|H_T(\theta) - H(\theta)\| + \|H(\hat{\theta}_T) - H(\theta^*)\| = o(1). \end{aligned}$$

Since, the strong uniform law of large numbers implies that the first term is $o(1)$ and the continuity of $H(\theta)$, theorem 2 and Slutsky theorem imply that the second term is $o(1)$.

By the continuity of the matrix inverse and assumption 4(b) it follows that for T sufficiently large, $H_T(\hat{\theta}_T)$ is nonsingular and $-H_T(\hat{\theta}_T)^{-1} \rightarrow Q(\theta^*)$ almost surely. \square

Appendix B: Estimation results

Table B.1: Estimation Result of the linear regressions for all currencies

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	4.38E-05	0.0001	0.7408
	β_1	1.4179	0.7837	0.0704
	β_2	0.4097	0.2861	0.1522
GBP/USD	α	3.6E-04	0.0001	0.0099
	β_1	0.3816	0.5247	0.4670
	β_2	0.4052	0.2842	0.1539
CHF/USD	α	2.74E-04	0.0001	0.0026
	β_1	0.3111	0.5738	0.5877
	β_2	0.4652	0.4028	0.2481

Table B.2(a): Results on SR regression (all currencies)

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	3.81E-05	5.57E-05	0.4938
	β_{11}	10.1073	1.0336	0.0000
	β_{12}	0.5334	0.3290	0.1050
	β_{21}	5.8611	0.9774	0.0000
	β_{22}	0.1002	0.1622	0.5367
	p_1	0.0592	0.0154	0.0001
	σ	2.15E-04	1.07E-05	0.0000
GBP/USD	α	0.041711	0.0136	0.00217
	β_{11}	2.94E-04	0.0001	0.00003
	β_{12}	14.9128	1.2038	0.00000
	β_{21}	0.2660	0.3707	0.47307
	β_{22}	8.4223	0.9634	0.00000
	p_1	0.1724	0.2613	0.50944
	σ	0.0417	0.0136	0.00217
CHF/USD	α	1.57E-04	2.58E-05	0.0000
	β_{11}	0.6154	0.2579	0.0170
	β_{12}	0.1993	0.0432	0.0000
	β_{21}	11.5382	0.2053	0.0000
	β_{22}	0.3319	0.0263	0.0000
	p_1	0.0390	0.0142	0.0062
	σ	3.65E-04	7.57E-06	0.0000

Table B.2(b): Results on TV-SR1 regression (all currencies)

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	3.45E-05	1.57E-05	0.0281
	β_{11}	10.1227	0.0985	0.0000
	β_{12}	0.5512	0.0724	0.0000
	β_{21}	5.8680	0.1198	0.0000
	β_{22}	0.1002	0.0650	0.1233
	η_0	-2.8231	0.3183	0.0000
	η_1	0.1152	0.4266	0.7871
	σ	2.15E-04	3.77E-06	0.0000
GBP/USD	α	2.93E-04	7.06E-05	0.0000
	$F_1 1$	14.9297	1.1235	0.0000
	$F_2 1$	0.2498	0.3510	0.4766
	$F_1 2$	8.4343	0.8835	0.0000
	$F_2 2$	0.1577	0.2418	0.5144
	η_0	-3.8914	0.7097	0.0000
	η_1	1.1897	0.7877	0.1309
	σ	3.59E-04	1.80E-05	0.0000
CHF/USD	α	1.83E-04	4.86E-05	0.0002
	β_{11}	4.6512	0.8581	0.0000
	β_{12}	0.3635	0.2516	0.1485
	β_{21}	15.3885	0.8184	0.0000
	β_{22}	0.4047	0.1441	0.0050
	η_0	-3.0682	0.5200	0.0000
	η_1	0.0369	0.7409	0.9602
	σ	3.54E-04	1.64E-05	0.0000

Table B.2(c): Results on TV-SR2 regression (all currencies)

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	3.70E-05	5.38E-05	0.4917
	β_{11}	10.1536	0.8736	0.0000
	β_{12}	0.5392	0.3220	0.0940
	β_{21}	5.9129	0.8155	0.0000
	β_{22}	0.0992	0.1609	0.5376
	ρ_0	-2.5590	0.4294	0.0000
	ρ_1	-0.3972	0.6450	0.5379
	σ	2.15E-04	1.04E-05	0.0000
GBP/USD	α	3.07E-04	7.61E-05	0.0001
	β_{11}	14.8415	1.3413	0.0000
	β_{12}	0.3471	0.3912	0.3749
	β_{21}	8.3335	1.1338	0.0000
	β_{22}	0.2375	0.2770	0.3911
	ρ_0	-3.8197	0.9042	0.0000
	ρ_1	0.9808	0.9888	0.3213
	σ	3.62E-04	1.88E-05	0.0000
CHF/USD	α	1.81E-04	4.74E-05	0.0001
	β_{11}	4.7459	0.8418	0.0000
	β_{12}	0.3543	0.2445	0.1473
	β_{21}	15.4491	0.8227	0.0000
	β_{22}	0.4004	0.1447	0.0056
	ρ_0	-3.6832	0.6654	0.0000
	ρ_1	1.0383	0.7906	0.1891
	σ	3.54E-04	1.78E-05	0.0000

Table B.3(a): Results on ESR regression (all currencies)

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	7.18E-05	4.36E-05	0.1002
	β_{11}	0.1277	0.2625	0.6265
	β_{12}	7.0861	0.4385	0.0000
	β_{21}	3.2645	0.2389	0.0000
	β_{22}	0.1081	0.1244	0.3850
	p_1	0.0287	0.4160	0.0000
	p_2	0.9132	0.2967	0.0000
	σ	1.57E-04	8.92E-06	0.0000
	GBP/USD	α	1.93E-04	7.59E-05
β_{11}		4.1118	0.3798	0.0000
β_{12}		0.2787	0.3609	0.4399
β_{21}		8.4436	0.4456	0.0000
β_{22}		0.3168	0.2332	0.1744
p_1		0.9312	0.3273	0.0000
p_2		0.9707	0.4568	0.0000
σ		2.86E-04	1.63E-05	0.0000
CHF/USD		α	1.13E-04	1.53E-05
	β_{11}	0.1322	0.0302	0.0000
	β_{12}	2.3622	0.0489	0.0000
	β_{21}	11.0393	0.1350	0.0000
	β_{22}	0.1546	0.0151	0.0000
	p_1	0.1111	0.2456	0.0000
	p_2	0.9628	0.3825	0.0000
	σ	2.03E-04	5.54E-06	0.0000

Table B.3(b): Results on TV-ESR1 regression (all currencies)

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	6.98E-05	4.40E-05	0.1130
	β_{11}	7.0919	0.4670	0.0000
	β_{12}	0.1362	0.2656	0.6082
	β_{21}	3.2604	0.2496	0.0000
	β_{22}	0.1099	0.1272	0.3879
	η_{01}	-4.6233	0.9927	0.0000
	η_{11}	1.6366	1.0859	0.1318
	η_{02}	-2.3090	0.4063	0.0000
	η_{12}	-0.0702	0.5593	0.9001
	σ	1.57E-04	8.42E-06	0.0000
	GBP/USD	α	2.96E-04	6.83E-05
β_{11}		5.7121	0.4117	0.0000
β_{12}		0.2339	0.3422	0.4942
β_{21}		0.8215	0.2364	0.0005
β_{22}		0.0022	0.2512	0.9930
η_{01}		-3.6019	0.6219	0.0000
η_{11}		1.1187	0.7143	0.1173
η_{02}		-2.1834	0.7546	0.0038
η_{12}		0.6815	0.9459	0.4712
σ		3.27E-04	2.18E-05	0.0000
CHF/USD		α	1.14E-04	2.87E-05
	β_{11}	2.3533	0.1673	0.0000
	β_{12}	0.1400	0.1485	0.3457
	β_{21}	11.0426	0.1883	0.0000
	β_{22}	0.1590	0.0862	0.0652
	η_{01}	-1.8017	0.2557	0.0000
	η_{11}	-0.6178	0.4604	0.1797
	η_{02}	-3.0082	0.5042	0.0000
	η_{12}	-0.4975	0.7371	0.4998
	σ	2.04E-04	8.16E-06	0.0000

Table B.3(c): Results on TV-ESR2 regression (all currencies)

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	7.16E-05	4.23E-05	0.0908
	β_{11}	7.0781	0.4403	0.0000
	β_{12}	0.1316	0.2561	0.6072
	β_{21}	3.2623	0.2380	0.0000
	β_{22}	0.1095	0.1231	0.3738
	ρ_{01}	-4.5747	0.9829	0.0000
	ρ_{12}	1.5413	1.0705	0.1499
	ρ_{01}	-2.2187	0.3633	0.0000
	ρ_{02}	-0.3212	0.5946	0.5890
	σ	1.58E-04	6.23E-06	0.0000
GBP/USD	α	3.30E-04	2.68E-05	0.0000
	β_{11}	5.7489	0.4345	0.0000
	β_{12}	0.3948	0.3541	0.2649
	β_{21}	0.7305	0.2365	0.0020
	β_{22}	0.1203	0.2349	0.6084
	ρ_{01}	-3.9343	0.7312	0.0000
	ρ_{12}	1.4852	0.8048	0.0650
	ρ_{01}	-1.3451	0.8225	0.1020
	ρ_{02}	-2.3784	0.0000	0.0000
	σ	3.41E-04	1.26E-05	0.0000
CHF/USD	α	1.11E-04	2.84E-05	0.0001
	β_{11}	2.3562	0.1615	0.0000
	β_{12}	0.1336	0.1424	0.3484
	β_{21}	11.0477	0.2222	0.0000
	β_{22}	0.1576	0.0812	0.0523
	ρ_{01}	-2.2405	0.3734	0.0000
	ρ_{12}	0.3217	0.4687	0.4925
	ρ_{01}	-4.4877	0.9207	0.0000
	ρ_{02}	1.7466	0.9967	0.0797
	σ	2.02E-04	1.04E-05	0.0000

References

- Bartle R.G. (2001). *A Modern Theory of Integration*, American Mathematical Society, Providence, Rhode Island.
- Bera A.K. and Jarque C.M (1982). Model specification tests: a simultaneous approach, *Journal of Econometrics*, 20, 59-82
- Bauer H. (1972). *Probability Theory and Elements of Measure Theory*, New York: Holt, Rinehart and Winston.
- Bollerslev, T., Chou, R.Y. and Kroner, K.F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence, *Journal of Econometrics*, 52, 5-59.
- Chong Y.Y. and Hendry D.F. (1986). Econometric evaluation of linear macroeconomics models, *Review of Economics Studies*, 53(4), 671-690.
- Clemen R. T. (1989). Combining forecasts: a review and annotated bibliography, *International Journal of Forecasting*, 5, 559-583.
- Darrat A.F. and Zhong M. (2000). On testing the random walk hypothesis: a model comparison approach, *The Financial Review*, 35, 105-124.
- Davidson J. (1994). *Stochastic Limit Theory*, Oxford University Press, New York.
- Day T.E. and Lewis C.M. (1992). Stock market volatility and information content of stock index options, *Journal of Econometrics*, 52, 267-287.
- Dempster A. P., Laird N. M. and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B*, 39,1-38
- Diebold F.X. and Nerlove (1989).The dynamics of exchange rate volatility: A multivariate latent factor ARCH model, *Journal of Applied Econometrics* 4(1) 1-21.
- Diebold (2001): *Elements of forecasting (second edition)*, South-Western College Publishing.
- Donaldson R.G. and Kamstra M. (1997). Artificial neural network-GARCH models for international stock return volatility, forecast combining with neural networks, *Journal of Empirical Finance* 4, 17-46.
- Dunis C.L and Huang X. (2002). Forecasting and trading currency volatility: an application of recurrent neural regression and model combination, *Journal of Forecasting*, 21, 317-354.

- Granger C.W.J. and Ramanathan R. (1984). Improved methods of forecasting, *Journal of Forecasting*, 3, 197-204.
- Hamilton (1994). *Time series Analysis*, Princeton University Press, Princeton.
- Hu M.Y. and Tsoukalas C. (1999). Combining conditional forecasts using neural networks: an application to the EMS exchange rates, *Journal of International Financial Markets, Institutions and Money*, 9, 407-422.
- Kiefer N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model, *Econometrica*, 46, 427-434.
- Koutmos G. (1998). Asymmetries in the conditional mean and the conditional variance: Evidence from nine stock markets, *Journal of Economics and Business*, 50, 277-290.
- Laopodis N.T. (2001). Time varying Behavior and asymmetry in EMS exchange rates, *International Economic Journal*, 15(4), 81-94.
- Li W.K. and Wong C.S. (2000). On a mixture autoregressive model, *Journal of the Royal Statistical Society B* 62(1), 95-115.
- Li W.K. and Wong C.S. (2001). On a logistic mixture autoregressive model, *Biometrika*, 88(3), 833-846.
- Ljung, G. and Box, G. (1978). On a measure of lack of fit in time-series models, *Biometrika*, 65, 297-303.
- McLachlan G.J. and Krishnan T. (1997). *The EM Algorithm and Extensions*, Wiley, New York.
- Meese R.A. and K. Rogoff (1983). Empirical exchange rates models of the seventies: do they fit out of sample, *Journal of International Economics*, 14, 2-24.
- Min C. and 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.
- Newey W.K. (1991). Uniform convergence in probability and stochastic equicontinuity, *Econometrica*, 59(4), 1161-1167.
- Pagan A.R. and Schwert G.W. (1990). Alternative models for conditional stock volatility, *Journal of Econometrics*, 45, 267-290.
- Preminger A., Ben-Zion U. and Wettstein D. (2003). Extended switching regression models: allowing for multiple latent state variables, Working Paper, Monaster Center for Economic Research, Ben Gurion University of the Negev.

- Quandt R.E (1958). The Estimation of parameters of linear regression systems obeying two separate regimes, *Journal of the American Statistical Association*, 53 873-880.
- Quandt R.E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes, *Journal of the American Statistical Association*, 55, 324-330.
- Render R.A. (1981). Note on the consistency of the maximum likelihood estimate for non-identifiable distributions, *The Annals of Statistics*, 9(1), 225-228.
- Render R.A. and Walker H.F. (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.
- Vuong C.H. (1983). Misspecification and conditional maximum likelihood estimation, California Institute of Technology , working paper 503.
- West K.D. and Cho D. (1995). The predictive ability of several models of exchange rate volatility, *Journal of Econometrics*, 69, 367-391.
- White H. (1980). A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, 48(4), 817-838
- White H. (1994). *Estimation, Inference and Specification Analysis*, Cambridge: Cambridge University Press.
- White H. (2001). *Asymptotic Theory for Econometricians (Revised Edition)*, New York Academic Press.
- Wooldridge J.M. (1994). Estimation and inference for dependent processes, in *Handbook of Econometrics* 4, pp. 2641-2700, edited by Engle R.F. and McFadden D.L. Elsevier Science B.V., Amsterdam.
- Wu C-F (1983). On the convergence of the EM algorithm, *The Annals of Statistics*, 11, 95-103.