

Extended Switching Regression
Models: Allowing for Multiple
Latent State Variables

**Arie Preminger, Uri Ben-Zion and
David Wettstein**

Discussion Paper No. 03-08

July 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: Allowing for Multiple Latent State Variables

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 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 provide conditions for the consistency and asymptotic normality of the derived estimates. Extensive simulation studies are carried out to assess the performance of the EM estimation method. Finally, we use the ESR model to combine forecasts of foreign exchange rates volatility. The resulting forecasts dominate those generated by traditional combining procedures.

JEL classification: C51, C53, C61

Keywords: EM algorithm, Forecasting, Extended Switching Regressions Models

(*)The authors thank Ezra Einy for several insightful comments and suggestions and seminar participants in the FFM 2003 conference and the Technion – Israel Institute of Technology for their comments. 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

Large parts of economic and financial time series data are characterized by dramatic changes over time, which can be attributed to various reasons, such as: wars, changes in economic policy and business cycles. When the aim of the analysis is prediction or simulation, it is important to model such behavior explicitly, and to take into account the factors that might cause these changes.

One of the main approaches to explain the behavior of time series is the switching regression approach. Quandt (1958, 1960) first proposed these models to determine whether subsets of the sample observations were generated by different probability distributions. Kiefer (1978) showed that there exists a bounded local maximum for the log likelihood function of a linear switching regression model that yields consistent and asymptotically normal estimates for the model's parameters. Recently, Li and Wong (2000, 2001) generalized the idea of mixture distributions to handle the case of nonlinear time series. Related classes of models are the hidden Markov model regressions (Hamilton 1994, Ch. 22), which differ from the switching regression models in that the unobserved state variable follows a latent Markov structure.

The main feature that characterizes these models is that the unobserved state variables are not real physical variables that happen to be missing. Instead, these variables represent underlying factors without precise physical definition, but which often, and desirably so, turn 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 a subset of the model parameters to be mutually dependent on the state variable. Such models have been used to describe many economic time series; such as: GNP (Hamilton 1990), stock price volatility (Friedman (1994), and Exchange rate fluctuations (Engel and Hamilton 1990).

We modify these models by introducing several state variables, which independently influence the model selection procedure, through the picking of a partial and disjoint group of the model parameters. This approach relaxes the assumption that the model parameters are mutually dependent on one state variable, i.e. it assumes that the data generation process is influenced by multiple independent causes or factors. The advantage of formulating state variables in such a way is that interesting qualitative information may result from the nature of the variables. For

instance, a model for risky assets can be independently influenced by latent changes in the individuals' risk aversion and government policy, which are not observed by the analyst. 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.

The layout of this paper is as follows: In Section 2 we introduce the idea of an extended switching regression (ESR) model. In Section 3 we address the large sample properties of a correctly specified ESR model and establish the consistency and asymptotic normality of the maximum likelihood estimates. In Section 4 we present the special case of a linear ESR model. In Section 5 we develop the relevant version of an EM algorithm for estimation. In Section 6 we discuss the use of the ESR model for prediction. In Section 7 we report the results of a simulation study designed to evaluate the performance of the EM estimation method. An empirical application is provided in Section 8, where we use the ESR model to combine forecasts and compare it to other combination methods. Finally, Section 9 summarizes the main findings and outlines several extensions.

2. A Regression Model with Several State Variables

In this work we extend the switching regression models. These 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 typical model:

$$(1) \quad y_t = \mu_t(x_t, \psi_0) + \varepsilon_t$$

where $\mu_t : X \times \Psi \rightarrow R$ 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 zero-mean white noise and $x_t \in X$ is a vector of explanatory variables. While these models have been popular in describing some "behavioral law", for example see the linear regression model, they are not flexible enough to account for situations where the researcher believes the parameters are not constant over time.

Switching regressions were 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 defines 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; $s_t = 2$, when the process distributes according to the second distribution, and so on, i.e. $s_t \in \{1, \dots, k\}$. The state variable is unobserved and the distribution of s_t is multinomial, therefore in the switching regression model we have:

$$(2) \quad y_t = \mu_t(x_t, \psi(s_t)) + \varepsilon_t$$

This model has the important property that $\psi(s_t)$, the parameters governing the data generation process, change over the set $\{\psi_1, \psi_2, \dots, \psi_k\}$ according to the values of the state variables. For example in linear switching regression models $\mu_t = x_t' \beta(s_t)$ and the parameter vector $\beta(s_t)$ change over the set $\{\beta_1, \beta_2, \dots, \beta_k\}$ according to the random values of the state variable. It is important to note that the explanatory variables do not have to be identical across states. We let x_t contain all the explanatory variables and put a-priori constraints on the model parameters. Likewise, we assume that in the switching regression described above, all the model parameters switch over time, but we could also relax this assumption and with out loss of generality assume that there is a subset of the model parameter which is deterministic as we do in the following examples.

Now, we look first at a linear switching regression:

$$(3) \quad y_t = \alpha + \begin{cases} \beta_{11}x_{1t} + \beta_{21}x_{2t} + \varepsilon_t & \text{if } s_t = 1 \\ \beta_{12}x_{1t} + \beta_{22}x_{2t} + \varepsilon_t & \text{if } s_t = 2 \end{cases}$$

The ε_t 's are independent identically distributed errors with mean zero and finite variance. This model can be cast in an equivalent way, which will prove useful later on. In order to do that we define the following vectors $\beta_1 = [\beta_{11}, \beta_{21}]$; $\beta_2 = [\beta_{12}, \beta_{22}]$; let S_t be a random two-dimensional vector, with the j -element equal to one if $s_t = j$, and equal to zero otherwise. Therefore, if $s_t = 1$, the vector S_t is equal to the first column of the identity matrix I_2 (the 2 x 2 identity matrix); when $s_t = 2$ the vector S_t is the second column of I_2 , and so on:

$$(4) \quad S_t = \begin{cases} (1,0)' & \text{if } s_t = 1 \\ (0,1)' & \text{if } s_t = 2 \end{cases}$$

Therefore, equation (4) can be rewritten as:

$$(5) \quad y_t = \alpha + \beta_1 S_t x_{1t} + \beta_2 S_t x_{2t} + \varepsilon_t$$

We will now define S_t^1, S_t^2 to be two-dimensional random vectors, so that at any point in time only one element in each vector equals one while the other equals zero. Equation (5) becomes:

$$(6) \quad y_t = \alpha + \beta_1 S_t^1 x_{1t} + \beta_2 S_t^2 x_{2t} + \varepsilon_t$$

One can also see that these models, i.e. linear switching regression models, characterized by a single latent variable s_t , are equivalent to models containing two state variables in which the correlation between the state variables is perfect. Therefore, it is likely that we will achieve better results if we put fewer constraints on the probability structure of the latent variables, and let the data characterize this structure. We assume independence across state variables and over time.

More formally, let us assume there exists a partition $\{d_i\}_{i=1}^p$ of the parameter set $\Psi \subset R^d$ such that $d = \sum_{i=1}^p d_i$ and for each of the p sets of parameters we define an independent state variable s_t^i , which picks randomly one of k_i elements from the i -th set, where $\psi_i(s_t^i)$ is a subset of the model parameters which can assume one of k_i values according to the realization of its state variable. 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 extended switching regression model could be described as follow:

$$(7) \quad y_t = \mu_t(x_t, \psi_1(s_t^1), \dots, \psi_p(s_t^p)) + \varepsilon_t$$

Now, suppose $p = d = 2$, building on (6), the dynamics of y_t can be described as:

$$(8) \quad y_t = \alpha + \begin{cases} \beta_{11}x_{1t} + \beta_{21}x_{2t} & \text{if } s_t^1 = 1, s_t^2 = 1 \\ \beta_{11}x_{1t} + \beta_{22}x_{2t} & \text{if } s_t^1 = 1, s_t^2 = 2 \\ \beta_{12}x_{1t} + \beta_{21}x_{2t} & \text{if } s_t^1 = 2, s_t^2 = 1 \\ \beta_{12}x_{1t} + \beta_{22}x_{2t} & \text{if } s_t^1 = 2, s_t^2 = 2 \end{cases}$$

We notice that if the state variables are perfectly correlated equation (8) is equivalent to equation (3). We define the process described by (8) as the Extended Switching Regression (ESR) model, unlike the Switching Regression (SR) model described in equation (3).

The model described in equation (8) can be estimated by a simple switching regression (SR) model with four states. The estimation of this model is inefficient due to the loss in degrees of freedom. This loss increases exponentially in the number of partitions and the number of states. 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 . Furthermore, the advantage of this structuring of the latent variables is that interesting qualitative information may result from the nature of the state variables. Therefore, 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 encompasses it with the cost of having to estimate an exponentially increasing number of parameters.

3. Asymptotic Theory

In this section we discuss large sample properties of the maximum likelihood estimators of the ESR models. Let $\{Z_t\}_{t \in \mathbb{N}}$ be an $\nu \times 1$ observed random variable taking its values in a measurable Euclidean space Ω . Let \mathfrak{F}_Z be the Borel σ -field on Ω . 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 $1 \times \ell$ dimensional vector of explanatory variables where $\nu = 1 + \ell$. Let $(\Omega_1, \mathfrak{F}_Y)$ and $(\Omega_2, \mathfrak{F}_X)$ be the measurable spaces associated with Y_t and X_t respectively. Let P_Z^t be the true joint distribution of Z_t . We are interested in a parametric family of conditional distributions $\{P_{Y|X}^t(\psi), \psi \in \Psi\}$ of Y_t given X_t , which exists by Jirina's theorem (see, Bauer (1972), p.319). Let P_X^t be the true marginal distribution of X_t and ν_Y be a σ -finite measure on $(\Omega_1, \mathfrak{F}_Y)$.

Assumption 1:

- (a) The random vectors $\{Z_t\}_{t \in \mathbb{N}}$ are independent with the true probability measures $\{P_Z^t\}_{t \in \mathbb{N}}$ on (Ω, \mathfrak{F}_Z)
- (b) For all ψ in Ψ and for P_X^t -almost all x the conditional distributions $\{P_{Y|X}^t(\cdot | x, \psi)\}_{t \in \mathbb{N}}$ are absolutely continuous with respect to ν_Y .

Assumption 1 allows us to use the Radon-Nikodym theorem (Royden (1988), p.276) to establish the existence of a measurable non-negative Radon-Nikodym density $f(y_t | x_t, \psi)$ for each ψ in Ψ . In the case of an absolutely continuous distribution function on the real line, the Radon-Nikodym derivative corresponds to the usual notion of a density function.

Now, let assume that there exists a latent model selection procedure which picks among the set $\{f(y_t | x_t, \psi), \psi \in \Psi\}$ of conditional densities each time. This selection process is unobserved and characterized by independent selections from the disjoint parameter sets $\{\Psi_i\}$, where $\Psi = \prod_{i=1}^p \Psi_i$. From each set Ψ_i we select one element among k possible ones. The selection is random and dependent on the realization of p unobserved state variables $s_t^i \in \{1, \dots, k\}$ as described in the previous section. A concise comparison between the latent structures of the ESR and the SR models is given in Figure 1.

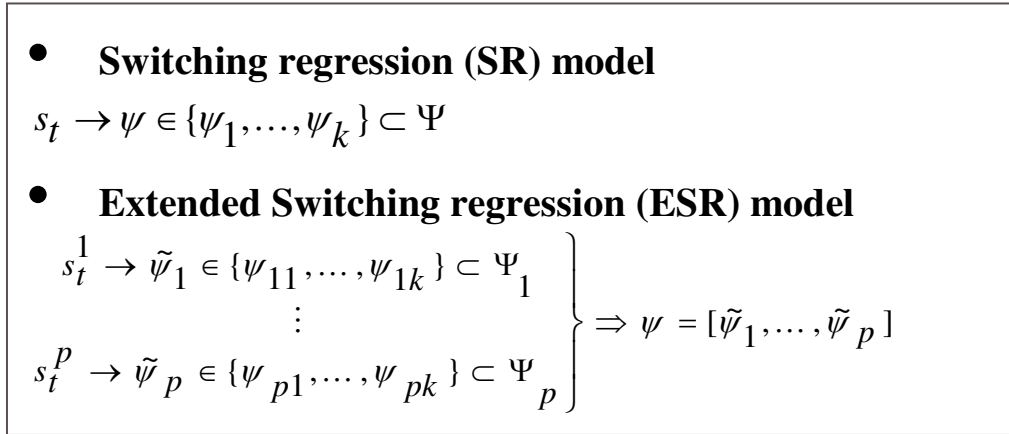


Figure 1: The latent structure of the ESR model and the SR mode.

Let $\{p_{i1}, \dots, p_{ik}\}$ denote the probabilities of the i -th selection. We assume for simplicity that these probabilities are constant over time. The results would not change were the probabilities to depend on some X_t . We define this model as ESR, and note that when $p = 1$ the process is the same as the switching regression model (SR). 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_i} and it is chosen with probability $p_{ij_i} \in [0,1]$, where $j_i \in \{1, \dots, k\}$. The conditional density of y_t is given by:

$$(9) \quad f(y_t | x_t, \theta) = \sum_{j_1=1}^k \cdots \sum_{j_p=1}^k \left[\left(\prod_{i=1}^p p_{ij_i} \right) \cdot f(y_t | x_t, \varphi_{1j_1}, \dots, \varphi_{pj_p}) \right]$$

where $\theta = (\varphi_1, \dots, \varphi_p, p_{11}, \dots, p_{1k-1}, \dots, p_{p1}, \dots, p_{pk-1}) \in \Phi$ is the vector of the model parameters. Hence, the likelihood function of the sample is

$$(10) \quad L_T(\theta) = \frac{1}{T} \sum_{t=1}^T \log f(y_t | x_t; \theta)$$

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

Assumption 2: The conditional density $f(y_t | x_t, \psi)$ is continuous on Ψ , a compact subset of \mathbb{R}^L , for $(P'_Z \cdot)$ almost all (y_t, x_t) , for all t .

Let \mathfrak{F}_Z^T be the T-product of \mathfrak{F}_Z . Under assumptions 1 and 2 we show in theorem 1 that the maximum likelihood estimator of the ESR model is a \mathfrak{F}_Z^T -measurable function of the data. Notice that since, the likelihood function for each observation is a convex combination of the 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 ESR model, as in the SR model the log-likelihood function attains its maximum at several different choices of θ obtained from the true parameter 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. For example, let the observations be generated by one of two distributions, where θ_1 denotes the parameters related to one distribution and θ_2

denotes the parameters related to the other distribution. In the case where these distributions come from the same parametric family, the likelihood function will give the same value for $\theta(1) = [\theta_1, \theta_2]$ or $\theta(2) = [\theta_1, \theta_2]$. Let

$$(11) \quad \theta^* \in C = \{\theta \in \Phi \mid f(y_t \mid x_t, \theta) = f(y_t \mid x_t, \theta_0)\}$$

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 ESR model, if we have p state variables where each variable can assume k values, and if the true parameters are θ_0 , 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) For all $\theta \in \Theta \mid \log f(y_t \mid x_t, \theta) \leq m(y_t, x_t)$, where $E\left(\left|m(y_t, x_t)\right|^{1+\lambda}\right) < \Delta < \infty$ for some $\lambda > 0$, for all t.

b) There exists $n_t(y_t, x_t)$ such that $\frac{1}{T} \sum_{t=1}^T E(n_t) = O(1)$ and a function

$h : [0, \infty) \rightarrow [0, \infty)$ with $h(0) = 0$ and $h(\cdot)$ is continuous at zero, such that

$\left| \log f(y_t \mid x_t, \tilde{\theta}) - \log f(y_t \mid x_t, \theta) \right| < n_t(y_t, x_t) h(d(\tilde{\theta}, \theta))$, for all $\tilde{\theta}, \theta \in \Theta$ where $d(\cdot, \cdot)$ is some metric on Θ .

c) For each $\theta_i \in C$ and all $\varepsilon > 0$, $\limsup_{T \rightarrow \infty} [\max_{\theta \in \bar{\eta}(\varepsilon)} E(L_T(\theta)) - E(L_T(\theta_i))] < 0$,

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

Assumption 3(a)-(b) ensures the uniform convergence of the sample likelihood function as a result of corollary 3.1 of Newey (1991). These conditions can be verified if the likelihood function is continuously differentiable on an open, convex set containing the parameter set. Assumption 3(c) is an identification requirement which guarantees that the likelihood function does not become increasingly flat in the neighborhood of $\theta^* \in C$. This implies that the number of states, we assume before estimation, should not exceed the correct number. Since, any model nesting the true

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

$$p \lim(\hat{\theta}_T) \rightarrow \theta^* \in C$$

Proof: See Appendix A.

Next, we introduce the conditions, which ensure the asymptotic normality of our MLE. This property is established by taking the mean-value 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 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

thus $D_T(\theta) = \frac{1}{T} \sum_{t=1}^T \partial \log f(y_t | x_t, \theta) / \partial \theta$, $H_T(\theta) = \frac{1}{T} \sum \partial^2 \log f(y_t | x_t, \theta) / \partial \theta \cdot \partial \theta'$ and let

$$H(\theta) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E \left[\partial^2 \log f(y_t | x_t, \theta) / \partial \theta \cdot \partial \theta' \right] \text{ and } Q(\theta) = -H(\theta)^{-1}.$$

Assumption 4:

- a) For some $\lambda > 0$, the elements of $|\partial \log f(y_t | x_t, \theta) / \partial \theta|$ and $|\partial^2 \log f(y_t | x_t, \theta) / \partial \theta^2|$ are respectively dominated by $2 + \lambda$, and $1 + \lambda$, P'_Z -integrable functions independent of θ , for all t .
- b) The elements of $H_T(\theta) - H(\theta)$ are asymptotically stochastically equicontinuous functions on the parameters.
- c) $\theta^* \in C$ is in the interior of Θ , and the matrix $H(\theta)$ exists and is nonsingular and positive definite in some open neighborhood of θ^* .
- d) The elements of $|\partial f(y_t | x_t, \theta) / \partial \theta|$ and $|\partial^2 f(y_t | x_t, \theta) / \partial \theta \cdot \partial \theta'|$ are dominated by P_0 -integrable functions independent of θ for all t .

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

$$\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} N(\underline{0}, Q(\theta^*)) \quad \text{where } -H_T(\hat{\theta}_T)^{-1} \xrightarrow{p} Q(\theta^*)$$

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 ESR version of the true model as we mention above. The assumptions above can be modified to handle cases, where the data exhibits some dependence. One such case is that of, strong mixing, analyzed by White and Domowitz (1982, 1984), also see Davidson (1994) and Gallant and White (1988) for a comprehensive survey. These processes exhibit both time-dependence and heterogeneity. Strong mixing essentially requires that events occurring now are almost independent of the events, which occurred in the distant past, although there might be considerable dependence on recent events. Common examples of such processes are m-dependent, as well as finite order Gaussian ARMA processes. By strengthening the moment conditions in assumptions 3-4 we can obtain similar results.

4. A linear ESR model

The ESR models we consider in this work are linear where the dependent variable is y_t and x_t is a $d \times 1$ vector of explanatory variables. The model is given by:

$$(12) \quad y_t = \alpha + \sum_{i=1}^d \beta_{it} x_{it} + \varepsilon_t$$

where $\varepsilon_t \sim i.i.N(0, \sigma)$ and $\beta_{it} \in \{\beta_{i1}, \dots, \beta_{ik}\}$ for $1 \leq i \leq d$, the normality assumption of the error, is common (see, e.g. Kiefer (1978), Li and Wong (2000)) and will simplify the estimation of the MLE as we see in the next section. The parameter β_{it} is

a random variable. Its distribution is discrete and is determined by a group of state variables $\{s_t^i\}_{i=1}^d$ as described in the previous section. An equivalent description of (12) is given by:

$$(13) \quad y_t = \alpha + \sum_{i=1}^d \beta_i S_t^i x_{it} + \varepsilon_t$$

where $\beta_i = [\beta_{i1}, \dots, \beta_{ik}]$, 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, and so on.

Therefore:

$$(14) \quad S_t^i = \begin{cases} [1, \dots, 0] & \text{if } s_t^i = 1 \\ \vdots & \vdots \\ [0, \dots, 1] & \text{if } s_t^i = k \end{cases}$$

Note that in this case we assume that $p = d$. Also note that this reduces to a switching regression model when the correlation between the state variables equals one. Such an assumption imposes a severe constraint on the relationship between the state variables. Imposing some degree of dependence across the state variables implies that the density function defined over $\{S_t^i\}_{i=1}^d$ contains k^d probabilities for k^d possible combinations of the model. In such a case, we have to estimate $k^d - 1$ probabilities, as well as $k \times d + 2$ other parameters. This creates an identification and estimation problem as we increase the number of explanatory variables. To address this problem we assume independence across the state variables. The intermediate case of limited dependence is equivalent to a switching regression model with some equality constraints, imposed on the parameters; therefore this case will not be addressed in this work. In the case of independence it is easy to see that we need to estimate $(k - 1) \times d$ probabilities and $k \times d + 2$ additional parameters.

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. Furthermore, we note that despite the fact independence is assumed, there might well be dependence across the state variables, given additional information from the observations. So, although we use the independence assumption as a basis for our estimation in the rest of the study, 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.

5. Estimation

In order to estimate the model parameters, we modify the EM algorithm, popularized by Dempster, Laird, and Rubin (1977) – 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 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 (e.g. Newton-Raphson based methods) will fail. 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.

Now, let $S_t = (1, S_t^1, S_t^2, \dots, S_t^{d_t})'$, $B = (\alpha, \beta_{11}, \dots, \beta_{1k}, \dots, \beta_{d_1}, \dots, \beta_{dk})'$,

$\tilde{X}_t = (1, x_{1t}, \dots, x_{1t}, x_{2t}, \dots, x_{d-1,t}, x_{dt}, \dots, x_{dt})'$. We first write the ESR model in a more compact way that would simplify calculations later on:

$$(15) \quad y_t = (B * \tilde{X}_t)' S_t + \varepsilon_t$$

The symbol "*" denotes the Hadamard product, which means element-by-element multiplication. Also $\theta^\ell = [B^\ell, \sigma^\ell, p_{11}^\ell, \dots, p_{1k}^\ell, \dots, p_{d_1}^\ell, \dots, p_{dk}^\ell]$ denotes the parameters estimated in the ℓ -th iteration and let $\Lambda_t(\theta^{\ell-1}) = E(S_t S_t' | y_t, x_t; \theta^{\ell-1})$,

$$\hat{S}_t(\theta^{\ell-1}) = E(S_t | y_t, x_t; \theta^{\ell-1}).$$

5.1 The M-step

The likelihood function of the complete data thus, assuming that the values of the state variables are known, is given by:

$$(16) L(\theta^\ell | \{y_t, x_t\}_{t=1}^T; \theta^{\ell-1}) = \sum_{t=1}^T \sum_{i=1}^d \sum_{j=1}^k S_t^{ij} \log(p_{ij}) - \sum_{t=1}^T \frac{(y_t - (B * \tilde{X}_t)' S_t)^2}{2\sigma^2} - T \log(2\pi\sigma^2)$$

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 get (17):

$$Q(\theta) = E(L(\theta^\ell | \{y_t, x_t\}_{t=1}^T; \theta^{\ell-1})) = \frac{1}{2\sigma^2} \sum_{t=1}^T \{y_t^2 - 2y_t (B * \tilde{X}_t)' \hat{S}_t(\theta^{\ell-1}) + (B * \tilde{X}_t)' \Lambda_t(\theta^{\ell-1}) \cdot (B * \tilde{X}_t)\} + \sum_{t=1}^T \sum_{i=1}^d \sum_{j=1}^k \hat{S}_t^{ij}(\theta^{\ell-1}) \cdot \log p_{ij} - T \log(2\pi\sigma^2)$$

Differentiating with respect to B yields:

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

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

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

$$(20) \quad p_{ij}^\ell = \frac{1}{T} \sum_{t=1}^T \hat{S}_t^{ij}(\theta^{\ell-1})$$

5.2 The E-step

The calculation of the expectation of the likelihood function is done with respect to the distribution of the latent state variables given the data, and the parameters estimated during 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} which is the j -element of S_t^i .

The elements of $\Lambda_t(\theta^{\ell-1})$, $\hat{S}_t(\theta^{\ell-1})$ can be deduced from the following calculations:

$$(21) \quad \hat{S}_t^{ij} = \Pr(S_t^{ij} = 1 | y_t, x_t; \theta^{\ell-1}) = \frac{\Pr(y_t, S_t^{ij} = 1 | x_t; \theta^{\ell-1})}{\Pr(y_t | x_t; \theta^{\ell-1})} = \frac{p_{ij} \Pr(y_t | x_t, S_t^{ij} = 1; \theta^{\ell-1})}{\sum_{\{S_t^{ij}\}} \Pr(y_t, S_t^{ij} | x_t; \theta^{\ell-1})} =$$

$$= \frac{p_{ij} \sum_{\{S_t^{rj}\}_{r \neq i}} \Pr(y_t, \{S_t^{rj}\}_{r \neq i} | x_t, S_t^{ij} = 1; \theta^{\ell-1})}{\sum_{\{S_t^{ij}\}} \Pr(y_t, S_t^{ij} | x_t; \theta^{\ell-1})}$$

$$(22) \Lambda_{m+1, n+1}(\theta^{\ell-1}) = E(S_t^{mj} S_t^{nj} | y_t, x_t)_{m \neq n} = \Pr(S_t^{mj} = 1, S_t^{nj} = 1 | y_t, x_t; \theta^{\ell-1})_{m \neq n} =$$

$$\frac{\Pr(S_t^{mj} = 1, S_t^{nj} = 1, y_t | x_t; \theta^{\ell-1})}{\Pr(y_t | x_t; \theta^{\ell-1})} = \frac{p_{mj} p_{nj} \Pr(y_t | x_t, S_t^{mj} = 1, S_t^{nj} = 1; \theta^{\ell-1})}{\Pr(y_t | x_t; \theta^{\ell-1})}$$

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

Given an initial guess of the 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 prespecified tolerance level. Several different starting values should be used, and the maximum likelihood estimates will correspond to that associated with the largest value of likelihood function that was obtained from the different starting values.

6. Prediction

The problem we consider is the forecasting of y_t , given a set of covariates x_t . Denote the optimal predictor by \hat{y}_t . As is common in the literature, we assume that our loss function is symmetric as defined by Granger (1969)¹ and therefore the optimal predictor is the conditional expectation, i.e. $\hat{y}_t = E(y_t | x_t)$. For our linear ESR model, the optimal predictor is:

$$(23) \quad \hat{y}_t = \alpha + \sum_{i=1}^d E(\beta_i) x_{it}.$$

We let $x_t = [1, x_{1t}, \dots, x_{dt}]$ and $\bar{\beta}_0 = [\alpha, E(\beta_{1t}), \dots, E(\beta_{dt})]$, rewrite equation (23) as

$$\hat{y}_t = x_t \bar{\beta}_0' \text{ and note that } y_t = x_t \bar{\beta}_0' + \varepsilon_t^* = x_t \bar{\beta}_0' + (\varepsilon_t + x_t' \beta_t) = \hat{y}_t + (\varepsilon_t + x_t' \beta_t)$$

where $\beta_t = [0, \beta_{1t} - E(\beta_{1t}), \dots, \beta_{dt} - E(\beta_{dt})]$ and $E(\beta_t) = \underline{0}$. Therefore, it is possible to use the OLS estimates, which we denote β_{OLS} , from a linear regression of y_t on x_t to obtain a consistent estimate of this predictor under the following assumptions:

¹ The loss function defined by Granger (1969) satisfies $L(z) = L(-z)$, $L(0) = 0$ and has an increasing first derivative.

Assumption 5:

$$y_t = \alpha + \sum_{i=1}^d \beta_{it} x_{it} + \varepsilon_t \quad t = 1, 2, \dots$$

$\beta_{it} \in \{\beta_{i1}, \dots, \beta_{ik}\}$ and the parameters $\{\beta_{it}\}$ are random variables which are dependent on the realization of i.i.d unobservable discrete state variables (a linear ESR model).

Assumption 6:

- (a) The vector (x_t, ε_t) is independent over time.
- (b) $E(\varepsilon_t x_{it}) = 0, i = 1, \dots, d, t = 1, 2, \dots$
- (c) $E|\varepsilon_t x_{it}|^{1+\lambda} < \Delta < \infty$ for some $\lambda > 0, i = 1, \dots, d, t = 1, 2, \dots$

Assumption 7:

- (a) $E(|x_{it} x_{jt}|^{2+\lambda}) < \Delta < \infty$ for some $\lambda > 0$ and $i, j \in \{1, \dots, d\}, t = 1, 2, \dots$
- (b) $Q = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E(x_t x_t')$ is nonsingular matrix.

Theorem 4: Given assumptions 5-7, $\beta_{OLS} \xrightarrow{a.s.} \bar{\beta}_0$.

Proof: See Appendix A.

We should note that this result apply to linear ESR model describe earlier in which we assume the error term is i.i.d. Gaussian noise. The MLE can be used to obtain predictions using Slutsky's theorem. We can also use the OLS (which is computationally straightforward) estimator for this purpose. But there are some advantages to using the MLE. These estimators are more efficient then the OLS estimator (Cramer-Rao bound), while the error in the linear regression is higher due to a misspecification error added to it.

Furthermore, the information provided in the ML estimators concerning parameters and their distribution² provides the researcher with better opportunities to incorporate additional information regarding the parameters of the model distribution, for example calculating the probability of the state variable given the predicted data, and so on.

In this work we consider an ESR model where the probabilities are constant, when the probabilities are allowed to change, running a linear regression would not be useful in prediction³. Also, the ESR model is non-linear, and forecasting a few steps ahead would require the exact specification of the functional form of the ESR model, hence using a linear model would not be sufficient. See Granger and Terasvitra (1993 pp. 130-135) for nonlinear forecasting methods and Le, Martin and Raftery (1996), Li and Wong (2000,2001), for a description of forecasting methods in a mixture autoregressive model, and Kwok et al. (1998) on an application of a mixture of ARMA and mixture AR models to forecast financial time series.

7. Simulations

In order to assess the performance of the EM algorithm and the convergence of the MLE to the true parameters, we perform simulations. The simulations correspond to a linear ESR model with two i.i.d. explanatory variables which are sampled from standard normal distribution, that is

$$(24) \quad y_t = \alpha + \beta_{1t}x_{1t} + \beta_{2t}x_{2t} + \sigma \cdot \varepsilon_t$$

where $\beta_{1t} \in \{\beta_{11}, \beta_{12}\}$, $\beta_{2t} \in \{\beta_{21}, \beta_{22}\}$, $\varepsilon_t \sim i.i.N(0,1)$.

The parameter β_{it} is determined according to the realization of its latent state variable (s_t^i). We sample each of the state variables, from a binomial distribution. The error terms ε_t were also sampled from a normal distribution with mean zero and unit variance. There are 1000 sample paths, each with 250 data points, generated by the ESR model. The parameters of the model are chosen as follows:

$$(\alpha, \sigma) = (1.15, 1.0), (\beta_{11}, \beta_{12}) = (2.0, 3.0), (\beta_{21}, \beta_{22}) = (4.0, 5.0), (p_{11}, p_{22}) = (0.5, 0.5).$$

² OLS estimates provide just the mean of the estimated parameters.

³ Our results are limited to the linear case, while in other cases, the usage of linear regression might not be sufficient

Table 1: Averages and Standard Deviation for the EM Parameter Estimate from the simulated data

Parameters	True value	Average	Empirical SE	Theoretical SE
α	1.15	1.153	0.052	0.054
β_{11}	2	1.975	0.181	0.178
β_{12}	3	3.029	0.178	0.182
β_{21}	4	3.978	0.174	0.167
β_{22}	5	5.028	0.178	0.17
σ	1	0.986	0.051	0.051
p_1	0.5	0.500	0.149	0.14
p_2	0.5	0.503	0.147	0.131

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, which are obtained using the large sample results from the previous section. 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 indicate that the estimation method works well. The sample means are very close to the true ones, and the standard deviations are small. Also note that the theoretical and empirical standard errors are close, with a small downward bias observed in the theoretical standard errors.

8. 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 and in economics in particular. This is not surprising, since it is reasonable to assume that most models are a simplified version of a complex real environment. One should refine and improve its 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.

The common approach in these methods is that of linear regression, first presented by Bates & Granger (1969), and expanded by Granger & Ramanathan (1984). More elaborate procedures such as the Bayesian time varying weight method (see e.g. Min and Zellner (1993)) have been proposed as well. However, for the forecasting horizon we investigate in our work, the same authors have demonstrated that there is no substantial benefit in 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; and an extended switching regression model. 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 the Ordinary Least Squares (OLS).

We expect the use of the ESR model to be superior to the common methods of (AVERAGE) and (OLS). This is due to the fact that performances of forecasters might vary over time. For example, the ESR model 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, forecasters using more “fundamental” models might fare better exactly under such circumstances and vice versa in other circumstances.

The forecasts of the volatility in daily exchange rates are considered as the basis for an application to 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 is for the time period from January 1, 1989 to December 30, 1995. Let E_t denote the logarithm of a spot exchange rate at time t . We concentrate on the exchange rate change $R_t = E_t - E_{t+1}$, so that R_t is the depreciation (or appreciation) of the domestic currency over time.

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 simple approach and assume that $R_t = \mu + \sigma_t \cdot \varepsilon_t$ where ε_t is an error term with zero mean and unit variance, μ and σ_t are the conditional mean and variance of R_t respectively. We estimate the variance according to the square of the error (see Amemiya (1985)).

Next, we generate volatility forecasts by two common models. The first model is the GARCH (1,1), in which volatility is estimated in the following way: $\sigma_t^2 = a_0 + a_1\sigma_{t-1}^2 + a_2\varepsilon_{t-1}^2$, where the parameters $\hat{a}_0, \hat{a}_1, \hat{a}_2, \mu$ are estimated using the ML estimation method. The second model is the MAV model (Pagan and Schwert, 1990), which defines volatility as a simple average of lagged squared residuals:

$$\sigma_t^2 = \frac{1}{n} \sum_{i=1}^n \varepsilon_{t-i}^2$$

where the number of lags n is chosen to minimize the Schwartz

Criterion. The GARCH model we use is widely employed for the purpose of forecasting volatility in financial markets, see Bollerslev (1987, 1992), Pagan and Schwert, (1990) and numerous references cited therein. The individual volatility forecasts produced by the GARCH and MAV models were combined through the following formula:

$$(25) \quad F_t = \alpha + \beta_1(s_t^1)f_{1t} + \beta_2(s_t^2)f_{2t}$$

where $s_t^i \in \{1, \dots, k_i\}$, $1 \leq k_i \leq \bar{k}$, $i = 1, 2$, 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 (17) provides the simple average (AVERAGE) as a combination 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 $\alpha, \beta_{11}, \beta_{21}$. In order to obtain combination of forecasts using parameters derived from switching regression (SR) model we will have to impose $s_t^1 = s_t^2$ on the formula above, while when we use the extended switching regression (ESR) model as a combining procedure, we impose no 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 up into three sub-samples; the first sub-sample contains observations from January 1, 1989, to December 30, 1991, which are used to estimate the parameters of the conditional volatility models (GARCH, MAV). Then each model 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 in 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 process continues until we get volatility forecasts from the first of January 1992 to December 30, 1995. The second sub sample is from January 1, 1992 to the December 30, 1994 and it is used to estimate the parameters of the OLS forecasts combining model as well as the parameters of the ESR and SR models. Given our model specification we obtain one-step ahead forecasts for the period from January 1, 1995 to the December 30, 1995 by weighting the individual forecasts of GARCH and MAV for the combining methods we described above.

8.1 Data analysis

In order to assess the distributional properties of the data, various descriptive statistics are reported in table 2, including mean, standard deviation, skewness, kurtosis and other statistics. In particular, the hypothesis of normality is rejected for each exchange rate, using the Bera and Jarque (1982) joint test (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 always very close to zero, while the kurtosis is very large. The rate of depreciation (or appreciation) of the currencies against the dollar, in our sample, are shown in Figure 2(a), 2(b) and 2(c) below. Visual inspection of each series revealed no evidence of serial correlation, although there seems to be persistence in the conditional variances.

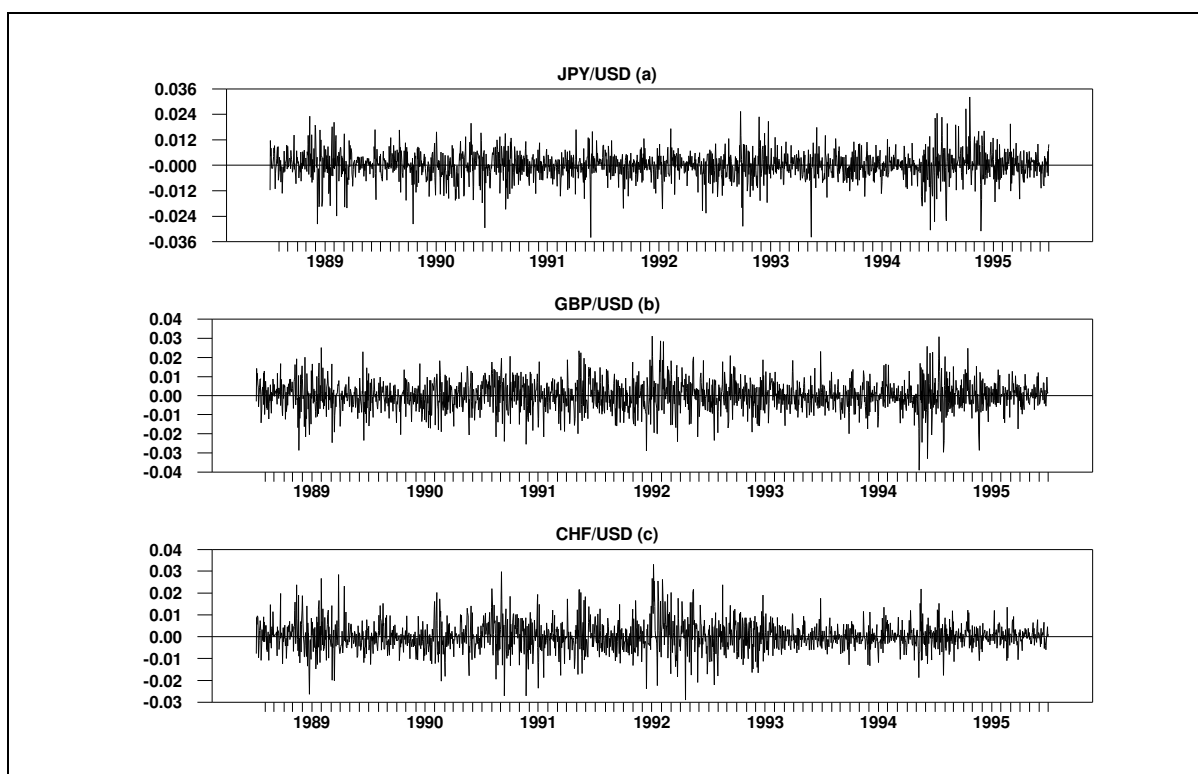


Figure 2: The rate of depreciation (or appreciation) of the currencies against the dollar

Table 2 also provides information about the autocorrelation structure of the data. It presents the first three-autocorrelation coefficients (r_1, r_2, r_3) along with their standard errors, which reveal no evidence of serial correlation. The Ljung-Box (1978) statistic (LB(12)), for 12-th serial correlation of the squared return of the exchange rate imply significant relationship. Bollerslev (1987) interprets the high autocorrelation of the squared data as a sign of conditional heteroscedasticity.

Table 2: Summary Statistics for the daily return from January 1, 1989 to the December 30, 1995

Statistic	JPY/USD	BP/USD	CHF/USD
Mean	-0.0002	-0.0001	0.0001
Std. Dev.	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
r_1	0.0365 (0.0260)	0.0379 (0.0260)	0.0841 (0.0260)
r_2	-0.0103 (0.0260)	-0.0334 (0.0260)	-0.0213 (0.0261)
r_3	-0.0227 (0.0260)	-0.0129 (0.0260)	0.0003 (0.0262)
LB(12)	42.53	43.99	94.58

(*) 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. r_1 , r_2 and r_3 are the first three autocorrelations along with their standard errors in parentheses. LB(24) is the Ljung and Box (1978) test estimated for the 12-th serial correlation for the squared returns of our data.

8.2 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 generalized 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 state probabilities are on the boundary of the parameter space and the parameters are not identifiable under the null model. A common approach in the literature to determine the appropriate number of states is to use a class of information criteria. See, Le, Martin and Raftery (1996) and Li and Wong (2000,2001) for using the Bayes Information Criterion (BIC) in order to

choose the number of states, in a mixture autoregressive model. We use the same criteria, which is defined as follows:

$$(26) \quad \text{BIC} = 2\log(\text{maximum likelihood}) - \text{MP}\log(T)$$

where MP is the number of the model parameters. Clearly, the more parameters we use, the better we are able to fit the data. Therefore this measure, first derived by Schwarz (1978), imposes a “penalty”, related to the number of parameters estimated, on the likelihood function. In order to identify the number of states in the SR and in the ESR models, we estimate each model, using the data in second sub sample, assuming that each state variable can take two to six states and choose the model with the highest BIC. 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 3: Values of the log-likelihood and the BIC statistic for the chosen SR and ESR models for January 1, 1992-December 30, 1994 (747 daily observations)

Exchange rate	Model	W	Log-likelihood	BIC	No. of states
JPY/USD	SR	11	6336	12599	3
	ESR	16	6621	13137	2, 5
GBP/USD	SR	14	6181	12270	4
	ESR	18	6265	12411	4, 4
CHF/USD	SR	11	6191	12310	3
	ESR	12	6355	12631	3, 2

Table 3 above presents the number of states, the value of log likelihood and the BIC for the model we chose. 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 these criteria can compare rival, non-nested models, we also see that in our sample the ESR model is strongly preferred to the SR model. The estimation results are presented in Appendix B. The EM algorithm, developed in the previous section is used for estimation.

8.3 Assessing the forecasting ability of the models

We compare the forecasts by using three common evaluation measures for assessing the predictive accuracy of forecasting models. The measures are the root-

mean-squared-error (RMSE), the root-mean-absolute-error (RMAE) and the encompassing test. The results in table 4 show the RMSE and the RMAE statistics for the period 1 January 1995 to the December 30, 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. While 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 all currencies.

Table 4: Root mean squared error and root mean absolute error for each exchange rate and volatility forecasting model for the period January 1,1995 – December 30 1995 (251 daily observations)

Model	JPY/USD		GBP/USD		CHF/USD				
	RMSE	RMAE	RMSE	RMAE	RMSE	RMAE	RMSE (*)	RMAE (*)	Average
GARCH	1.67E-04	9.15E-03	2.49E-04	1.12E-02	7.9E-05	6.29E-03	103.3%	99.4%	101.3%
MAV	1.68E-04	1.01E-02	2.63E-04	1.20E-02	7.86E-05	6.81E-03	105.3%	107.9%	106.6%
AVERAGE	1.63E-04	9.49E-03	2.52E-04	1.14E-02	7.66E-05	6.34E-03	101.9%	101.5%	101.7%
OLS	1.78E-04	9.50E-03	2.38E-04	1.09E-02	7.11E-05	6.41E-03	100.5%	100.3%	100.4%
SR	1.63E-04	9.30E-03	2.50E-04	1.12E-02	7.68E-05	6.83E-03	101.7%	102.6%	102.1%
ESR	1.67E-04	9.30E-03	2.36E-04	1.06E-02	7.57E-05	6.68E-03	100.0%	100.0%	100.0%

(*) RMSE and RMAE are reported as a ratio to that of the ESR model

The usage of the RMAE criterion also does not reveal any clear dominance relationship among the models for all exchange rates. The GARCH model is performing well in the JPY/USD data while the ESR model is better than other combining models in the GBP/USD data. Therefore we calculate the RMSE and RMAE relatively to a benchmark model, which is the ESR model for each currency, and average the results for each measure across the different currencies. The result indicate that according to RMSE the ESR model is better than other combining models while according to the RMAE the GARCH model is preferable but on average the ESR model dominate the rival models.

8.4 The encompassing test

Although useful, the evaluation measures used so far could not determine whether a given forecasting model is in fact “significantly” better than others. The ranking of several competing forecasting models might hence require the use of the

encompassing test (see, Chong and Hendry (1986) and Donaldson and Kamastra (1996)). The rationale behind this test is that a model j should be preferred to model k if model j can explain what model k cannot explain, while model k is unable to explain what model j cannot explain. In more specific terms, model j dominates model k , if model j 's forecasts can significantly explain model k 's forecast error, that is, model j incorporates relevant information neglected by model k . The encompassing test is implemented by running a set of OLS regressions of the forecast errors of one model on the forecasts of the other model.

More formally, let f_t^j, f_t^k be the forecasts of models j and k respectively, and e_t^j, e_t^k be the models' forecast errors respectively. The test is based on examining the significance of the parameters θ_1^j in the regression $e_t^k = \alpha_1 + \theta_1^j f_t^j + \varepsilon_{1t}$ and θ_1^k in the regression $e_t^j = \alpha_2 + \theta_1^k f_t^k + \varepsilon_{2t}$, where ε_{1t} and ε_{2t} are error terms. The null hypothesis is that neither model encompasses the other. If θ_1^j is significantly different from zero but θ_1^k is not, we reject the null hypothesis concluding that model j encompasses model k and vice versa. If both θ_1^j and θ_1^k are either significant or not significant, we do not reject the null hypothesis that neither model encompasses the other.

Our findings are reported in table 5. Columns 2 to 6 of the tables contain the p-values associated with robust consistent t-statistics based on the computation of heteroscedasticity-consistent standard errors (see White, (1980)), p-values less than 0.10 indicate that the forecasts from model j explain the error in model k , with a significance level of 10%. For example in table 5 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% 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% level. Therefore, for the case of the JPYUSD the SR model encompasses the MAV model. The bold numbers in the tables below indicate that the model in the column encompasses the model in the row, i.e. it is statistically preferred. Results from the encompassing test reported in Table 5 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 the ESR encompasses the

MAV in the JPY/USD data. The ESR also encompasses the MAV in the GBP/USD data and the OLS in the CHF/USD data. Therefore, we conclude based on the encompassing test, that the ESR is significantly preferred to the other forecast combining methods.

Table 5: Results for the out of sample encompassing test

	Dependent variable:	Independent variable: Forecast from					
Exchange Rate	Forecasting error 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	-

9. Summary

In this paper we have presented a new class of models, the ESR models. These models 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. This modeling approach enables us to increase the number of states of the model with parsimonious parameterization. These models constitute an important

addition to modeling choices as they allow the analyst to take into account multiple latent factors that might influence the data generation process.

We also developed an EM algorithm in order to estimate the parameters in a linear ESR model. We showed that the estimates are consistent and asymptotically normal under a set of general conditions. We then considered the usage of the ESR model as a method for combining forecasts and compared it to other popular methods. The results presented above suggest that ESR combined forecasts generally outperform forecasts from traditional combining methods. The practical significance of this result is evident from the out-of-sample tests employed.

The ESR model could be extended, to allow for time varying probabilities, lending a large degree of flexibility to the model. Another extension consists of allowing the state variables $\{s_t^i\}$ defined in Section 2 to follow a Markov chain rather than to be independent across time as in the present formulation. It may be possible to generalize the model further by considering time series with both discrete and continuous components. These extensions leave several interesting and challenging areas for future research.

Appendix A:

Proof of Theorem 1: Assumption 1 implies that $f(y_t | x_t, \theta)$ is measurable for each $\theta \in \Theta$ a compact set by theorem 3.25 of Davidson (1994), p.52, and Assumption 2 implies that $f(y_t | x_t, \theta)$ is continuous on Θ for P'_z -almost all (y, x) , for all t . 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 corollary 3.1 in Newey (1991). The likelihood function converges uniformly on Φ to $L(\theta) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E(\log f(y_t | x_t; \theta))$, a continuous function on a compact set, which attains its maximum on the set C . So, for $\varepsilon, \delta > 0$ there exists a \hat{T} such that, for $T > \hat{T}$, $P(\sup_{\theta \in \Theta} |L_T(\theta) - L(\theta)| > \delta) < \varepsilon$.

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(c).

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 $\theta^* \in C$

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 $\Pr(E_T) \leq \Pr(\hat{\theta}_T \in \bigcup_i N_i)$. Set $\delta = e/2$, hence for $T > \hat{T}$,

$\Pr(E_T) > 1 - \varepsilon$, because $\varepsilon > 0$ is arbitrary, assumptions 1-3 imply

$P \lim \hat{\theta}_T \rightarrow \theta^* \in C$. \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 in probability to $-H(\theta^*)^{-1}$. Hence, we need to show that $\limsup_{T \rightarrow \infty} P(\|H_T(\bar{\theta}) - H_T(\theta^*)\| > \delta) \leq \varepsilon$

where $\|\cdot\|$ is the Euclidean matrix norm. From assumption 4(b) and Theorem 2, we see that for given $\varepsilon, \delta > 0$ there exists an $\eta > 0$ such that (A.3)

$$\begin{aligned} & \limsup_{T \rightarrow \infty} P(\|H_T(\bar{\theta}) - H_T(\theta^*)\| > \delta) \leq \\ & \limsup_{T \rightarrow \infty} P(\|H_T(\bar{\theta}) - H_T(\theta^*)\| > \delta, d(\bar{\theta}, \theta^*) \leq \eta) + \limsup_{T \rightarrow \infty} P(d(\bar{\theta}, \theta^*) > \eta) \leq \\ & \limsup_{T \rightarrow \infty} P(\sup_{\theta \in \Theta, d(\theta, \theta^*) \leq \eta} \|H_T(\theta) - H_T(\theta^*)\| > \delta) < \varepsilon \end{aligned}$$

where $d(\cdot, \cdot)$ is some metric on Θ .

The second inequality holds since $p \lim \hat{\theta}_T = \theta^*$ by Theorem 2, and $\bar{\theta}$ lies on the segment joining $\hat{\theta}_T$ and θ^* , and the last inequality uses assumption 4(b) and the definition of stochastic equicontinuity (Davidson (1994, pp.335-336)). By assumption 4(a), the law of large numbers implies that $p \lim [H_T(\theta^*) + o_p(1)] \rightarrow H(\theta^*)$, where the convergence is element-wise. By Assumption 4(c), $H(\theta^*)$ is nonsingular so that $H_T(\theta^*)$ is nonsingular in probability 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:

$$(A.4) \quad p \lim [H_T(\theta^*) + o_p(1)]^{-1} \rightarrow H(\theta^*)^{-1}.$$

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

$$(A.5) \quad \sqrt{T}(\hat{\theta}_T - \theta^*) = H(\theta^*)^{-1} \sqrt{T} D_T(\theta^*)$$

In order to analyze the distribution of $\sqrt{T} D_T$, we consider the random vectors $(\partial \log f(y_t | x_t, \theta) / \partial \theta)$, (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), which are independent but not identically distributed with:

$$(A.6) \quad E(\partial \log f(y_t | x_t, \theta) / \partial \theta) = \int (\partial \log f(y_t | x_t, \theta) / \partial \theta) \cdot f(y_t | x_t, \theta) dy_t$$

$$= \int (\partial f(y_t | x_t, \theta) / \partial \theta) dy_t = \partial \left(\int f(y_t | x_t, \theta) dy_t \right) / \partial \theta = \underline{0}$$

By differentiating under the integral sign again we obtain the information matrix equality (A.7) $E((\partial \log f(y_t | x_t, \theta) / \partial \theta) \cdot (\partial \log f(y_t | x_t, \theta) / \partial \theta)) = -E(\partial^2 \log f(y_t | x_t, \theta) / \partial \theta \cdot \partial \theta)$

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

Now, to apply the central limit theorem for the multivariate case, we use the Cramer-Wold device and the Liapounov Central limit theorem. See Davidson (1994, pp.372-

374). Let $V_T = \frac{1}{T} \sum_{t=1}^T E((\partial \log f(y_t | x_t, \theta) / \partial \theta) \cdot (\partial \log f(y_t | x_t, \theta) / \partial \theta)')$, from assumption

1(a) we can see that $\text{var}(\sqrt{T}D_T) = V_T$, also let $\ddot{Z}_{t,T} \equiv \zeta' \cdot V_T^{-1/2} \cdot \partial \log f(y_t | x_t, \theta) / \partial \theta$,

where $\zeta' \cdot \zeta = 1$ and $\ddot{Z}_T = \frac{1}{T} \sum_{t=1}^T \ddot{Z}_{t,T}$.

The summands $\ddot{Z}_{t,T}$ are independent given 1(a) with $E(\ddot{Z}_{t,T}) = 0$ given the equality above, and $E(|\ddot{Z}_{t,T}|^{2+\lambda})$ is uniformly bounded by assumption 4(a) and Minkowski's

inequality, also $\bar{\sigma}_T^2 \equiv \text{var}(\sqrt{T}\ddot{Z}_T) = \zeta' \cdot V_T^{-1/2} \text{var}(\sqrt{T}D_T) \cdot V_T^{-1/2} \zeta = \zeta' \cdot V_T^{-1/2} V_T \cdot V_T^{-1/2} \zeta = 1$.

Hence, for all ζ , $\zeta' \cdot \zeta = 1$, it follows from Liapounov Central limit theorem that

$$(A.8) \quad T^{-1/2} \sum_{t=1}^T \ddot{Z}_{t,T} = T^{-1/2} \sum_{t=1}^T \zeta' \cdot V_T^{-1/2} \cdot \partial \log f(y_t | x_t, \theta) / \partial \theta \xrightarrow{D} N(0,1)$$

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

$$(A.9) \quad \sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} N(\underline{0}, Q(\theta^*))$$

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

Next, we show that $H_T(\hat{\theta}_T)^{-1}$ is a consistent estimator for $Q(\theta^*)$. Using assumption

4(a), the strong law of large numbers implies that $\|H_T(\hat{\theta}_T) - H(\hat{\theta}_T)\| = o_p(1)$ almost

surely. Since by 4(b), $H(\theta)$ is continuous function on Θ , theorem 2 imply that

$\|H(\hat{\theta}_T) - H(\theta^*)\| = o_p(1)$, for each $\theta^* \in C$, we have:

$$(A.10) \quad \|H_T(\hat{\theta}_T) - H(\theta^*)\| < \|H_T(\hat{\theta}_T) - H(\hat{\theta}_T)\| + \|H(\hat{\theta}_T) - H(\theta^*)\| = o_p(1).$$

The continuity of the matrix inverse and assumption 4(c) it follows that $H_T(\hat{\theta}_T)$ is nonsingular for T sufficiently large and $p \lim(-H_T(\hat{\theta}_T)^{-1}) = Q(\theta^*)$. The conclusion follows by Slutsky theorem. \square

Proof of Theorem 4:

$$(A.11) \quad \beta_{OLS} = \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot x_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot y_t \right) = \bar{\beta}_0 + \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot x_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot \varepsilon_t^* \right)$$

$$= \bar{\beta}_0 + \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot x_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot \varepsilon_t + \frac{1}{T} \sum_{t=1}^T x_t \cdot (x_t' \beta_t) \right)$$

By assumption 6(a) and proposition 3.10 (White 2001, p.32) the elements of $\{x_t x_t'\}$ are independent sequence and given assumption 7(b) and Liapunov inequality we can apply the Markov's law of large numbers (see: Greene (1998) pp.296-297) to obtain:

$$(A.12) \quad \left| \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot x_t' \right) - Q \right| \xrightarrow{a.s.} 0.$$

Since Q is nonsingular by assumption 7(b) so that $\left(\frac{1}{T} \sum_{t=1}^T x_t x_t' \right)$ is nonsingular almost surely for T sufficiently large. Since the elements of the matrix inverse are continuous functions of the original matrix elements, we get:

$$(A.13) \quad \left| \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot x_t' \right)^{-1} - Q^{-1} \right| \xrightarrow{a.s.} 0$$

Next, $\{x_t \varepsilon_t\}$, $\{x_t x_t' \beta_t\}$ are independent sequences given assumption 6(a) and proposition 3.2 (White (2001), p.32) and we will show that they have finite bounded moments of order $1 + \lambda$, which enables us to apply the Markov's strong law of large numbers. Since for $\lambda > 0$, $E(|\beta_{it}^2|^{1+\lambda}) < \Delta$ given assumption 5 and also $E(|(x_{it} x_{jt})^2|^{1+\lambda}) < \Delta$ based on assumption 7(a), we derive by the Cauchy-Schwarz inequality that $E(|x_{it} x_{jt} \beta_{jt}|^{1+\lambda}) < \Delta$ for $i, j \in \{1, \dots, d\}$. By Minkowski's inequality,

$$(A.14) \quad E \left| \sum_{j=1}^d x_{it} x_{jt} \beta_{jt} \right|^{1+\lambda} \leq \left[\sum_{j=1}^d \left(E |x_{it} x_{jt} \beta_{jt}|^{1+\lambda} \right)^{1/(1+\lambda)} \right]^{1+\lambda} \leq d^{1+\lambda} \cdot \Delta$$

By Markov's law of large numbers

$$(A.15) \quad \left| \frac{1}{T} \sum_{t=1}^T x_{it} x_t' \beta_t - \frac{1}{T} \sum_{t=1}^T E(x_{it} x_t' \beta_t) \right| \xrightarrow{a.s.} 0$$

From assumptions 6(a)-(b)

$$(A.16) \quad \left| \frac{1}{T} \sum_{t=1}^T x_{it} \varepsilon_t - \frac{1}{T} \sum_{t=1}^T E(x_{it} \varepsilon_t) \right| \xrightarrow{a.s.} 0$$

Hence,

$$(A.17) \quad \left| \left(\frac{1}{T} \sum_{t=1}^T x_t x_t' \right)^{-1} \cdot \frac{1}{T} \sum_{t=1}^T x_t \varepsilon_t^* - Q^{-1} \left(\frac{1}{T} \sum_{t=1}^T (E(x_t \varepsilon_t) - E(x_t x_t' \beta_t)) \right) \right| \xrightarrow{a.s.} \underline{0}$$

By assumption 5 we see that $E(x_t \varepsilon_t) = E(x_t x_t' \beta_t) = \underline{0}$ and because Q^{-1} is bounded,

we obtain:

$$(A.18) \quad \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot x_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T x_t \cdot \varepsilon_t^* \right) \xrightarrow{a.s.} \underline{0} \Rightarrow \beta_{OLS} \xrightarrow{a.s.} \bar{\beta}_0 \square$$

Appendix B: Estimation results

Table B.1: Estimation Result for all currencies

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	3.88E-05	3.14E-06	0.0000
	β_1	0.0238	0.0221	0.2825
	β_2	0.0653	0.036	0.6966
GBP/USD	α	6.66E-06	9.99E-0.6	0.5021
	β_1	0.7612	0.1918	0.0001
	β_2	0.0996	0.0661	0.1320
CHF/USD	α	3.95E-06	4.63E-6	0.3938
	β_1	0.7749	0.1571	0.0000
	β_2	0.07924	0.0833	0.3414

Table B. 2(a): Results on SR regression -Japanese Yen vs. US Dollar

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	1.69E-5	1.1E-5	0.1152
	β_{11}	0.0880	0.4524	0.8457
	β_{12}	7.2145	1.0479	0.0000
	β_{13}	28.5952	4.2695	0.0000
	β_{21}	0.0993	0.04664	0.03320
	β_{22}	1.1854	0.04251	0.0053
	β_{23}	0.9255	0.6145	0.1320
	p_1	0.9518	0.0076	0.0000
	p_2	0.0375	0.0036	0.0000
	σ	4.1E-5	4.57E-6	0.0000

Table B.2(b): Results on SR regression - British Pound vs. US Dollar

Exchange rate	Variable	Coefficient	Std error	Signif.
GBP/USD	α	2.63E-5	3.84E-6	0.0000
	β_{11}	12.0725	2.2475	0.0000
	β_{12}	2.0097	0.2330	0.0000
	β_{13}	4.9584	0.3638	0.0000
	β_{14}	0.1359	0.0747	0.06892
	β_{21}	0.1475	1.0811	0.8914
	β_{22}	0.4752	0.1814	0.0088
	β_{23}	0.8632	0.2169	0.0001
	β_{24}	0.1232	0.0315	0.0001
	p_1	0.105	0.0034	0.0020
	p_2	0.1110	0.0143	0.0000
	p_3	0.04544	0.0088	0.0000
	σ	3.71E-5	3.0E-6	0.0000

Table B.2(c): Results on SR regression - Swiss Franc vs. US Dollar

Exchange rate	Variable	Coefficient	Std error	Signif.
CHF/USD	α	2.30E-5	4.2E-6	0.0000
	β_{11}	12.0204	2.1906	0.0000
	β_{12}	5.0379	0.5729	0.0000
	β_{13}	0.0238	0.1055	0.8216
	β_{21}	0.9033	0.8927	0.3115
	β_{22}	0.7529	0.2466	0.0023
	β_{23}	0.0839	0.0630	0.1832
	p_1	0.0331	0.0071	0.0000
	p_2	0.0738	0.0109	0.0000
	σ	4.44E-5	4.92E-6	0.0000

Table B.3(a): Results on ESR regression - Japanese Yen vs. US Dollar

Exchange rate	Variable	Coefficient	Std error	Signif.
JPY/USD	α	2.80E-5	4.57E-06	0.0000
	β_{11}	16.9564	0.4488	0.0000
	β_{12}	0.7014	0.1986	0.0004
	β_{21}	3.9257	0.0765	0.0000
	β_{22}	22.6526	0.5560	0.0000
	β_{23}	0.0764	0.0226	0.0007
	β_{24}	9.7694	0.2174	0.0000
	β_{25}	1.3513	0.0546	0.0000
	p_{11}	0.0105	0.0033	0.0019
	p_{21}	0.0508	0.0092	0.0000
	p_{22}	0.0088	0.0033	0.0083
	p_{23}	0.8044	0.0189	0.0000
	p_{24}	0.03167	0.0064	0.0000
	σ	1.98E-05	6.407E-07	0.0000

Table B.3(b): Results on ESR regression - British Pound vs. US Dollar

Exchange rate	Variable	Coefficient	Std error	Signif.
GBP/USD	α	1.47E-5	2.98E-06	0.0000
	β_{11}	16.8942	2.0497	0.0000
	β_{12}	4.0812	0.1064	0.0000
	β_{13}	12.3321	0.5059	0.0000
	β_{14}	0.0279	0.0605	0.6449
	β_{21}	2.4076	0.0988	0.0000
	β_{22}	6.4345	0.1234	0.0000
	β_{23}	0.0833	0.0277	0.0026
	β_{24}	1.2997	0.0494	0.0000
	p_{11}	7.5E-3	8.62E-08	0.0000
	p_{12}	0.0632	6.94E-03	0.0000
	p_{13}	0.0181	4.07E-10	0.0000
	p_{21}	0.0729	0.0141	0.0000
	p_{22}	.0347	7.604E-3	0.0000
	p_{23}	0.7921	0.0195	0.0000
σ	2.6145E-05	9.207E-07	0.0000	

Table B.3(c): Results on ESR regression -Swiss Franc vs. US Dollar

Exchange rate	Variable	Coefficient	Std error	Signif.
CHF/USD	α	1.30E-5	2.15E-06	0.0000
	β_{11}	17.6640	0.4593	0.0000
	β_{12}	5.4945	0.1295	0.0000
	β_{13}	0.0046	0.0833	0.9554
	β_{21}	0.1358	0.0398	0.0006
	β_{22}	1.5484	0.0573	0.0000
	p_{11}	0.0137	0.0044	0.0022
	p_{12}	0.0688	0.0102	0.0000
	p_{21}	0.8644	0.0202	0.0000
	σ	2.6145E-05	9.707E-07	0.0000

References

Amemiya, T. (1985). *Advanced Econometrics*, Cambridge Mass: Harvard University Press.

Andrews W.K. (1987). Consistency in nonlinear econometric models: a generic uniform law of large numbers, *Econometrica*, 55, 1465-1471

Bartle R.G. (2001). *A Modern Theory of Integration*, American Mathematical Society, Providence, Rhode Island.

Bates, J.M and Granger C.W.J. (1969). The combination of forecasts, *Operational Research Quarterly*, 20, 451-468.

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. (1987). A conditional heteroskedastic time series model for speculation prices and Rates of return, *The Review of Economics and Statistics*, 69, 542-547.

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.

Davidson J. (1994). *Stochastic Limit Theory*, Oxford University Press, New York.

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

Domowitz I. and White H. (1982). Misspecified models with dependent observations, *Journal of Econometrics* 20 35-58.

Domowitz I. and White H. (1984). Nonlinear regression with dependent observations, *Econometrica*, 52(1), 143-161.

Donaldson R.G. and Kamstra M. (1996). Forecast combining with neural networks, *Journal of Forecasting*, 15, 46-61.

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

- Friedman M. (1994). A two state capital asset pricing model, Technical report, Institute for Mathematics and its Applications, University of Minnesota.
- Gallant A.R and White H. (1988). *A Unified Theory of Estimation and inference for Nonlinear Dynamic Models*, New York: Basil Blackwell.
- Granger C.W.J. (1969). Prediction with a generalized cost of error function, *Operational Research Quarterly*, 20(2), 199-207.
- Granger C.W.J., Ramanathan R. (1984). Improved methods of forecasting, *Journal of Forecasting*, 3, 197-204.
- Granger C.W.J. and Terasvirta T. (1993). *Modeling Nonlinear Economic Relationships*, Oxford University Press, New York.
- Hamilton J.D. (1990). Analysis of time series subject to changes in regime, *Journal of Econometrics*, 45, 39-70.
- Hamilton (1994). *Time series Analysis*, Princeton University Press, Princeton.
- Kiefer N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model, *Econometrica*, 46, 427-434.
- Kwok H.Y., Chen C.M. and Xu L. (1998). Comparison between mixture of ARMA and mixture of AR model with Application to Time Series Forecasting, The fifth international Conference on Neural Information Processing, 1049-1052, Japan.
- 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., 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*, John-Wiley, New York.
- MacDonald R. and M.P.Taylor (1992). Exchange rate economics: A survey, IMF staff papers, 39, 1-57.
- 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.

Quandt R.E (1958). The Estimation of parameters of linear regression systems obeying two separate regimes, *Journal of the American Statistical Associations*, 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.

Quandt R.E (1972). A new approach to estimating switching regressions, *Journal of the American Statistical Association*, 67, 306-310.

Quandt R.E. and Ramsey J.B. (1978). Estimating mixture of normal distributions and switching regression, *Journal of the American Statistical Association*, 73, 730-738

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.

Royden H.L. (1988). *Real Analysis*, Third Edition, Macmillan Publishing Company, New York.

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(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.

Wu C-F (1983). On the convergence of the EM algorithm, *The Annals of Statistics*, 11, 95-103.