Forecasting in the presence of in and out of sample breaks *

Jiawen Xu[†] **Pierre Perron**[‡]

Shanghai University of Finance and Economics **Boston University**

January 30, 2017

Abstract

We present a frequentist-based approach to forecast time series in the presence of in-sample and out-of-sample breaks in the parameters of the forecasting model. We first model the parameters as following a random level shift process, with the occurrence of a shift governed by a Bernoulli process. In order to have a structure so that changes in the parameters be forecastable, we introduce two modifications. The first models the probability of shifts according to some covariates that can be forecasted. The second incorporates a built-in mean reversion mechanism to the time path of the parameters. Similar modifications can also be made to model changes in the variance of the error process. Our full model can be cast into a conditional linear and Gaussian state space framework. To estimate it, we use the mixture Kalman filter and a Monte Carlo expectation maximization algorithm. Simulation results show that our proposed forecasting model provides improved forecasts over standard forecasting models that are robust to model misspecifications. We provide two empirical applications and compare the forecasting performance of our approach with a variety of alternative methods. These show that substantial gains in forecasting accuracy are obtained.

Keywords: instabilities; structural change; forecasting; random level shifts; mixture Kalman filter.

JEL Classification: C22, C53

*We are grateful to two referees for useful comments and to David Pettenuzzo for sharing the code used in Pesaran et al. (2006).

[†]School of Economics, Shanghai University of Finance and Economics, 777 Guoding Rd., Shanghai China 200433 (xu.jiawen@mail.shufe.edu.cn).

[‡]Department of Economics, Boston University, 270 Bay State Rd., Boston MA 02215 (perron@bu.edu).

1 Introduction

Forecasting is obviously of paramount importance in time series analyses. The theory of constructing and evaluating forecasting models is well established in the case of stable relationships. However, there is growing evidence that forecasting models are subject to instabilities, leading to imprecise and unreliable forecasts. This is so in a variety of fields including macroeconomics and finance. Indeed, Stock and Watson (1996) documented widespread prevalence of instabilities in macroeconomic time series relationships. A prominent example is forecasting inflation; see, e.g., Stock and Watson (2007). This problem is also prevalent in finance. Pastor and Stambaugh (2001) document structural breaks in the conditional mean of the equity premium using long time return series. Paye and Timmermann (2006) examined model instability in the coefficients of ex post predictable components of stock returns. See also Pesaran and Timmermann (2002), Rapach and Wohar (2006) and Pettenuzzo and Timmermann (2011).

There is a vast literature on testing for and estimating structural changes within a given sample of data; see, e.g., Andrews (1993), Bai and Perron (1998, 2003) and Perron (2006) for a survey. Much of the literature does not model the breaks as being stochastic. Hence, the scope for improving forecasts is limited. There can be improvements by relying on the estimates of the last regime (or at least putting more weights on them) but even then such improvements are possible if there are no out-of-sample breaks. In the presence of out-of-sample breaks the limitation imposed by treating the breaks as deterministic mitigates the forecasting ability of models corrected for in-sample breaks. This renders forecasting in the presence of structural breaks quite a challenge; see, e.g., Clements and Hendry (2006).

Some Bayesian models have been proposed to address this problem; see, e.g., Pesaran et al. (2006), Koop and Porter (2007), Maheu and Gordon (2008), Maheu and McCurdy (2009) and Hauwe et al. (2011). The advantage of the Bayesian approach steams from the fact that it treats the parameters as random and by imposing a prior (or meta-prior) distribution one can model the breaks and allow them to occur out-of-sample with some probability. Such methods can, however, be sensitive to the exact prior distributions used.

We propose a frequentist-type approach with a forecasting model in which the changes in the parameters have a probabilistic structure so that the estimates can help forecast future out-of-sample breaks. Our approach is best suited to the case for which breaks occur both in and out-of-sample, which in particular avoids the problematic use of a trimming window assumed to have a stable structure. The method will work best indeed if there are many

in-sample breaks, so that a long span of data is beneficial. This is unavoidable since good out-of-sample forecasts of breaks require in-sample information about the process generating such breaks, the more so the more efficient the forecasts will be. The same applies to previously proposed Bayesian methods, though the use of tight priors can partially substitute for the lack of precise in-sample information. Having said that, our method still yields considerable improvements even if relatively few breaks are present in-sample.

Our approach is similar in spirit to unobserved components models in which the parameters are modeled as random walk processes. There are, however, important departures. Most importantly, a shift need not occur every period. It does so with some probability dictated by a Bernoulli process for the occurrence of shifts and a normal random variable for its magnitude. This leads to a specification in which the parameters evolve according to a random level shift process. Some or all of the parameters of the model can be allowed to change and the latent variables that dictate the changes can be common or different for each parameters. Also, the variance of the errors may change in a similar manner.

The basic random level shift model has been used previously to model changes in the mean of a time series, whether stationary or long-memory, in particular to try to assess whether a seemingly long-memory model is actually a random level shift process or a genuine long-memory one; see Ray and Tsay (2002), Perron and Qu (2010), Lu and Perron (2010), Qu and Perron (2013), Xu and Perron (2014), Li et al. (2016) and Varneskov and Perron (2016). It has been shown to provide improved forecasts over commonly used short or long-memory models. Our basic framework is a generalization in which any or all parameters of a forecasting model are modeled as random level shift processes.

To improve the forecasting performance we augment the basic model in two directions. First, we model the probability of shifts as a function of some covariates which can be forecasted. Second, we allow a mean-reversion mechanism such that the parameters tend to revert back to the pre-forecast average. This last feature is especially influential in providing improvements in forecasting performance at long horizons. Functional forms for these two modifications are suggested for which the parameters can be estimated and incorporated in the forecast scheme to model the future path of the parameters.

Modeling parameters as random level shifts has been suggested previously but, to our knowledge, only in a Bayesian framework. McCulloch and Tsay (1993) considered an autoregression in which the intercept is subject to random level shifts, though the autoregressive parameters are held fixed. They also allow the probability of shifts to depend on some covariates and changes in the variance of the errors (though using a different specification than

ours). Gerlach, Carter and Kohn (2000) consider a class of conditionally linear Gaussian state-space models with a vector of latent variables indicating the occurrence of changes in the coefficients that follow a Markov process. Pesaran, Pettenuzzo and Timmerman (2006) extend the Markovian structure of Chib (1998) with a fixed number of regimes by adopting a hierarchical prior with a constant transition probability matrix out of sample, thereby allowing breaks to occur at each date in the post-sample period. Koop and Potter (2007) consider models with a random number of regimes with the transitions from one regime to another being dictated by a Markov process and the durations of the regimes following a Poisson distribution. Giordani and Kohn (2008) extend their analysis, and that of Gerlach, Carter and Kohn (2000) to allow an arbitrary number of shifts occurring independently for the coefficients and error variance using a random level shift process with constant probability of shifts. Giordani, Kohn and van Dijk (2007) consider a class of conditionally linear and Gaussian state-space models which allows nonlinearity, structural change and outliers that can accommodate a fixed number of regimes with Markov transitions probabilities or random level shift processes, though in the applications they restrict the magnitudes of change and impose restrictive structures on the latent variables indicating the occurrence of changes. Groen, Paap and Ravazzolo (2013) use a model with random level shifts in the coefficients and error variance with constant probabilities to model and forecast inflation. Smith (2012) consider a Markov breaks regression model akin to a random coefficient model with all parameters changing at the same time and the probability of shifts being Markovian. As noted in some of the applications, the results can be quite sensitive to the prior used. Our approach is closest to that of McCulloch and Tsay (1993) except that we consider a general forecasting linear model with the same type of changes in coefficients and variance of the errors, allowing the probabilities of shifts to depend on some covariate. We also incorporate a mean-reversion mechanism. More importantly, we do not adopt a Bayesian approach and thereby bypass the need to specify priors and have the results influenced by them. Also, our focus is explicitly on providing improved forecasts. As stated in the previous review of the literature, the basic ingredient of the structure adopted has been considered previously, though not advanced as a widely applicable forecasting framework. Our aim is to generalized it and provide a 'general purpose' forecasting model that performs well for diverse scenarios with or without breaks. We believe this will be useful for empirical work related to forecasting.

Our model can be cast into a non-linear non-Gaussian state space framework for which standard Kalman filter type algorithms cannot be used. The state space representation of our model is actually a linear dynamic mixture model in the sense that it is linear and Gaussian

conditional on some latent random variables. Chen and Liu (2000) propose a special sequential Monte Carlo method, the mixture Kalman filter, which uses a random mixture of Gaussian distributions to approximate a target distribution. Giordani et al. (2007) discuss the advantages of the class of conditionally linear and Gaussian state space models. The EM (Expectation Maximization) algorithm is used to obtain the maximum likelihood estimates of the parameters. This allows treating the latent state variables as missing data (see Bilmes, 1998) and using a complete or data-augmented likelihood function which is easier to evaluate than the original likelihood. Since the missing information is random, the complete-data likelihood function is a random variable and we end up maximizing the expectation of the complete-data log-likelihood with respect to the missing data. Wei and Tanner (1990) introduced the Monte Carlo EM algorithm where the evaluation step is executed by Monte Carlo methods. Random samples from the conditional distribution of the missing data (state variables) can be obtained via a particle smoothing algorithm. The forecasting procedure is then relatively simple and can be carried out in a straightforward fashion once the model has been estimated.

Simulations show that the estimation method provides very reliable results in finite samples. The parameters are estimated precisely and the filtered estimates of the time path of the parameters follow closely the true process. To show the robustness of our forecasting model, we design simulations comparing the forecasting performances of various popular models (various form of the RLS models, historical average, rolling average, ARMA, ARIMA, Markov Switching, Time Varying Parameters) when the Data Generating Process (DGP) is one of the forecasting models considered. The results show that our random level shift model with built-in mean reversion always performs nearly as well as the model corresponding to the true DGP, and can even be better (e.g., when the true DGP is ARIMA or Markov Switching). All other forecasting methods perform very poorly in one or more of the cases considered. Hence, our method provides reliable results that are robust to a wide range of processes.

We apply our forecasting model to two series which have been the object of considerable attention from a forecasting point of view. The emphasis is on the equity premium. We compare the forecast accuracy of our model relative to the most important forecasting methods applicable for this variable. We also consider different forecasting sub-samples or periods. The results show clear gains in forecasting accuracy, sometimes by a very wide margin; e.g., over 90% reduction in mean squared forecast error relative to popular contenders. For this particular series, it turns out that the Time Varying Parameter Model performs quite well

being a close second best. To show the robustness of our forecasting model, we also consider the Treasury bill rate. Our method continues to provide the best forecasts overall, while the Time Varying Parameter Model lead to very poor forecasts in most samples considered Other applications can be found in the working paper version and in Xu (2017).

Finally, note that given the availability of the proper code for estimation and forecasting, the method is very flexible and easy to implement. For a given forecasting model, all that is required by the users are: 1) which parameters (including the variance of the errors if desired) are subject to change; 2) whether the same or different latent Bernoulli processes dictates the timing of the changes in each parameters; 3) which covariates are potential explanatory variables to model the probability of shifts.

The rest of the paper is organized as follows. Section 2 describes the basic model with random level shifts in the parameters. Section 3 discusses the modifications introduced to improve forecasting: the modeling of the probability of shifts and the allowance for a mean-reverting mechanism. Section 4 presents the estimation methodology: the mixture Kalman filtering algorithm in Section 4.1, the particle smoothing algorithm in Section 4.2, the Monte Carlo Expectation Maximization method to evaluate the likelihood function in Section 4.3. Section 5 introduces the construction of in-sample confidence bands and out-of-sample forecast bands. Section 6 provides forecasting simulations of various models to show the reliability and robustness of our proposed method. Section 7 contains the applications and comparisons with other forecasting methods. Section 8 offers brief concluding remarks. Detailed estimation algorithms are included in an appendix.

2 Model setup

We consider a basic forecasting model specified by

$$y_t = X_t\beta_t + e_t \tag{1}$$

where y_t is a scalar variable to be forecasted, X_t is a k-vector of covariates and, in the base case, $e_t \sim i.i.d.\ N(0, \sigma_e^2)$. It is assumed that some or all of the parameters are time-varying and exhibit structural changes at some unknown time. The specification adopted for the time-variation in the parameters is the following:

$$\beta_t = \beta_{t-1} + K_t^{\beta}\eta_t$$

where $K_t^{\beta} = diag(K_{1,t}^{\beta}, \ldots, K_{k,t}^{\beta})$ and $\eta_t = (\eta_{1,t}, \ldots, \eta_{k,t})' \sim i.i.d.\ N(0, \Sigma)$. The latent variables $K_{j,t}^{\beta} \sim Ber(p^{(j)})$ and are independent across j. Hence, each parameter evolves

according to a Random Level Shift (RLS) process such that the shifts are dictated by the outcomes of the Bernoulli random variables $K_{j,t}^{\beta}$. When $K_{j,t}^{\beta} = 1$, a shift $\eta_{j,t}$ occurs drawn from a $N(0, \sigma_{\eta,j}^2)$ distribution, otherwise when $K_{j,t}^{\beta} = 0$ the parameter does not change. The shifts can be rare (small values of $p^{(j)}$) or frequent (larger values of $p^{(j)}$).

This specification is ideally suited to model changes in the parameters occurring at unknown dates. Many specifications are possible depending on the assumptions imposed on K_t^{β} and Σ. First, when $K_{1,t}^{\beta} = \ldots = K_{k,t}^{\beta}$, we can interpret the model as one in which all parameters are subject to change at the same times, akin to the pure structural change model of Bai and Perron (1998). A partial structural model, can be obtained by setting $p^{(j)} = 0$ for the parameters not allowed to change, or equivalently by setting the corresponding rows and columns of Σ to 0. The case with $K_{1,t}^{\beta} = \ldots = K_{k,t}^{\beta}$ is arguably the most interesting for a variety of applications. However, it is also possible not to impose equality for the different $K_{j,t}^{\beta}$. This allows the timing of the changes in the different parameters to be governed by different independent latent processes. This may be desirable in some cases. For instance, it is reasonable to expect changes in the constant to be related to low frequency variations of the random level shifts type, while changes in the coefficients associated with random regressors to be related to business-cycle type variations. In such cases, it would therefore be desirable to allow the timing of the changes to be different for the constant and the other parameters. Of course, many different specifications are possible, and the exact structure needs to be tailored to the specific application under study.

The assumption that the latent Bernoulli processes $K_{j,t}^{\beta}$ are independent across j may seem strong. It implies that the timing of the changes are independent across parameters. As stated above, this can be relaxed by imposing a perfect correlation, i.e., setting some latent variables to be the same. Ideally, one may wish to have a more flexible structure that would allow imperfect though non-zero correlation. This generalization is not feasible in our framework. In many cases, it may also be sensible to impose that Σ is a diagonal matrix. This implies that the magnitudes of the changes in the various parameters are independent. In our applications, we follow this approach as it appears the most relevant case in practice and also considerably reduces the complexity of the estimation algorithm to be discussed in Section 4. Hence, for the j^{th} parameter β_j $(j = 1, \ldots, k)$, we have

$$\beta_{j,t} = \beta_{j,t-1} + K_{j,t}^{\beta} \eta_{j,t} \tag{2}$$

where $\eta_{j,t} \sim N(0, \sigma_{\eta,j}^2)$ and $K_{j,t}^{\beta} \sim Ber(p^{(j)})$.

In some cases, it may also be of interest to allow for changes in the variance of the errors.

The specification for the distribution is then $e_t = \sigma_{\epsilon,t}\epsilon_t$ with

$$ln\sigma^2_{\epsilon,t} = \phi ln\sigma^2_{\epsilon,t-1} + K^{\sigma}_t v_{\epsilon,t} \tag{3}$$

where $\epsilon_t \sim N(0,1)$, $K^{\sigma}_t \sim Ber(p^{\sigma})$ and $v_{\epsilon,t} \sim N(0,\sigma^2_v)$.

Remark 1 *When $p^{(j)} = p^{\sigma} = 0$ for all j, the model reduces to the classic regression model with time invariant parameters. When $p^{(j)} = 1$ for all j and $p^{\sigma} = 0$, it becomes the standard time varying parameter model; e.g., Rosenberg (1973), Chow (1984), Nicholls and Pagan (1985) and Harvey (2006).*

Remark 2 *In equation (3), if $\phi = 1$, we have a random level shift model for volatility. And if we add that $K^{\sigma}_t = 1$, we have the stochastic volatility modeled as a random walk. If $|\phi| < 1$ and $K^{\sigma}_t = 1$, we have the commonly used stochastic volatility as an approximation to the stochastic volatility diffusion of Hull and White (1987). Stock and Watson (2007) used a similar unobserved component stochastic volatility (UC-SV) model to forecast inflation, in which the stochastic volatility equation is specified with $\phi = 1$ and $K^{\sigma}_t = 1$.*

3 Modifications useful for forecasting improvements

The framework laid out in the previous section is well tailored to model in-sample breaks in the parameters. However, as such it does not allow future breaks to play a role in forecasting. In order to be able to do so, we incorporate some modifications. Two features that are likely to improve the fit and the forecasting performance is to allow for changes in the probability of shifts and model explicitly a mean-reverting mechanism for the level shift component. In the first step, we specify the jump probability to be

$$p^{(j)}_t = f(\gamma, w_t)$$

where γ is a m-vector of parameters, w_t are m covariates that would allow to better predict the probability of shifts and f is a function that ensures $p_t \in [0,1]$. Note that w_t needs to be in the information set at time t in order for the model to be useful for forecasting. We shall adopt a linear specification with the standard normal cumulative distribution function $\Phi(\cdot)$, so that $K^{\beta}_{j,t} \sim Ber(p^{(j)}_t)$ with $p^{(j)}_t = \Phi(r_0 + r_1' w_t)$, where r_0 is a scalar and r_1 and m-vector of parameters. As similar specification can be made for the probability of the Bernoulli random variable K^{σ}_t affecting the shifts in the variance of the errors.

The second step involves allowing a mean reverting mechanism to the level shift model. The motivation for doing so is that we often observe evidence that parameters do not jump

arbitrarily and that large upward movements tend to be followed by a decrease. This feature can be beneficial to improve the forecasting performance if explicitly modeled. The specification we adopt is the following:

$$\begin{aligned} \eta_{j,t} &\sim N(\mu_{\eta,j,t}, \sigma^2_{\eta,j}) \\ \mu_{\eta,j,t} &= \rho(\beta_{j,t-1} - \bar{\beta}_j^{(t-1)}) \end{aligned}$$

where $\beta_{j,t-1}$ is the filtered estimate of the parameter subject to change at time $t-1$ and $\bar{\beta}_j^{(t-1)}$ is the mean of all the filtered estimates of the jump component from the beginning of the sample up to time $t-1$. This implies a mean-reverting mechanism provided $\rho < 0$. The magnitude of ρ then dictates the speed of reversion. If $\rho = 0$, there is no mean reversion. Note that the specification involves using data only up to time t in order to be useful for forecasting purposes. Also, it will have an impact on forecasts since being in a high (low) values state implies that in future periods the values will be lower (higher), and more so as the forecasting horizon increases. Hence, this specification has an effect on the forecasts of both the sign and size of future jumps in the parameters. Similar specifications can be made to p^σ and $v_{\epsilon,t}$ for the changes in the variance of the errors.

4 Estimation methodology

The model described is within the class of conditional linear Gaussian State Space models of the form

$$y_t = X_t\beta_t + e_t \tag{4}$$

$$\begin{aligned} \beta_t &= \beta_{t-1} + K_t^\beta \eta_t \\ ln\sigma^2_{\epsilon,t} &= \phi ln\sigma^2_{\epsilon,t-1} + v_{\epsilon,t} \end{aligned} \tag{5}$$

where y_t is the variable to be forecasted and $(\beta_t, K_t^\beta, ln\sigma^2_{\epsilon,t})$ is the state vector. The measurement equation is (4) and the transition equations are (5). Conditional on $(K_t^\beta, ln\sigma^2_{\epsilon,t})$, the resulting system is a linear and Gaussian state space model and $p(\beta_t|K_t^\beta, ln\sigma^2_{\epsilon,t}, Y; \Theta)$, where $Y = (y_1, ..., y_T)$ and Θ is the vector of parameters, can be evaluated by the Kalman filter. The particle filters used are due to Chen and Liu (2000) who named them the mixture Kalman filters.

Remark 3 *In equation (5), we can add random level shifts in the stochastic volatility process as in (3). See the appendix for details.*

4.1 Mixture Kalman filtering

In this section, we use the conventional notation x_t to denote the state variable, while $\lambda_t \equiv (K_t^{\beta}, ln\sigma_{\epsilon,t}^2)$ are the latent variables. Let $y^t = (y_1, ..., y_t)$, $\Lambda_t = (\lambda_1, ..., \lambda_t)$, and let λ^t be realizations of Λ_t. The filtering distribution of x_t can be written as

$$p(x_t|y^t) = \int p(x_t|y^t, \lambda^t)p(\lambda^t|y^t)d\lambda^t$$

where $p(x_t|y^t, \lambda^t) \sim N\left(\mu_t(\lambda^t), \Sigma_t(\lambda^t)\right)$, in which $\left(\mu_t(\lambda^t), \Sigma_t(\lambda^t)\right)$ can be obtained by running the Kalman filter with a given trajectory λ^t. The main idea of the mixture Kalman filter is to use a weighted sample of the indicators $S_t = \{(\lambda^{t,(1)}, w_t^{(1)}), ..., (\lambda^{t,(M)}, w_t^{(M)})\}$ to represent the distribution $p(\lambda^t|y^t)$, where $w_t^{(i)}$ are some weights to be defined below and $\lambda^{t,(i)}$ are simulated latent variables; e.g., in the basic model $\lambda^{t,(i)} \equiv (K_1^{\beta,(i)}, ..., K_t^{\beta,(i)})$, so that given a jump probability p, $\lambda^{t,(i)}$ can be generated as random draws from the Bernoulli distribution with probability p. One then uses a random mixture of Gaussian distributions

$$\frac{1}{W_t}\sum\nolimits_{i=1}^{M} w_t^{(i)} N\left(\mu_t(\lambda^{t,(i)}), \Sigma_t(\lambda^{t,(i)})\right)$$

where $W_t = \sum_{i=1}^{M} w_t^{(i)}$, to represent the target distribution $p(x_t|y^t)$. The detailed mixture Kalman filtering algorithm is provided both for the basic model (equations (1) and (2)) and the extended model with stochastic volatility (equations (4) and (5)) with or without RLS in the appendix. To illustrate the adequacy of this method, we present simple illustrative examples. First, the true process for β_t is generated using equations (1) and (2) with mean reversion and time varying probability with the parameters $(r_0,\ r_1,\ \sigma_e,\ \sigma_\eta,\ \rho) = (-1.96, 4, 0.2, 0.2, -0.1)$. The number of observations is 1000. Figure 1 presents a plot the true path of β_t along with the filtered estimates of β_t. One can see a close agreement between the two. Figure 2 considers the more general case with stochastic volatility, where the true processes for β_t and the stochastic volatility are generated using equations (4) and (5) with mean reversion and time varying probability with the parameters $(r_0,\ r_1,\ \phi,\ \sigma_v,\ \sigma_\eta,\ \rho) = (-1.96, 4, 0.95, 0.2, 0.2, -0.1)$. A plot the true path of β_t along with the filtered estimates of β_t are presented in Panel A. The corresponding values for the volatility process are presented in Panel B. Again, the filtered values closely follow the true paths in both cases. While obviously limited, the cases reported are representative of what one can expect in most cases (from unreported additional simulations performed), showing the adequacy of the filtering method adopted.

4.2 Particle smoothing

The particle smoothing algorithm is designed to obtain particle smothers $\{s_t^{(i)}\}_{i=1}^M$ with certain weights $\{w_t^{(i)}\}_{i=1}^M$ from $p(x_t|y^T)$. Godsill et al. (2004) provide a forward-filtering and backward-simulation smoothing procedure. It allows drawing random samples from the joint density $p(x_0, x_1, \ldots x_T|y^T)$, not only the individual marginal smoothing densities $p(x_t|y^T)$. The smoothing algorithm relies on a pre-filtering procedure and a previously obtained set of filters $\{w_t^{(i)}, x_t^{(i)}\}_{i=1}^M$ for each time period. The main ingredients behind the smoothing algorithm are the relations:

$$p(x_1, \ldots, x_T|y^T) = p(x_T|y^T) \prod_{t=1}^{T-1} p(x_t|x_{t+1}, \ldots, x_T, y^T)$$

and

$$\begin{aligned} p(x_t|x_{t+1}, \ldots, x_T, y^T) &= p(x_t|x_{t+1}, y^t) \\ &= \frac{p(x_t|y^t)p(x_{t+1}|x_t)}{p(x_{t+1}|y^t)} \propto p(x_t|y^t)p(x_{t+1}|x_t) \end{aligned}$$

The first equality follows from the Markov property of the model and the second from Bayes' rule. Since random samples $\{x_t^{(i)}\}_{i=1}^M$ from $p(x_t|y^t)$ can be obtained from the mixture Kalman filtering algorithm, $p(x_t|x_{t+1}, \ldots, x_T, y^T)$ can be approximated as $\sum_{i=1}^M w_{t|t+1}^{(i)} \delta_{x_t^{(i)}}(x_t)$ with modified weights

$$w_{t|t+1}^{(i)} = \frac{w_t^{(i)} p(x_{t+1}|x_t^{(i)})}{\sum_{i=1}^M w_t^{(i)} p(x_{t+1}|x_t^{(i)})}.$$

where $\delta_{x_t^{(i)}}(x_t)$ is the Dirac delta function. This procedure is performed in a reverse-time direction conditioning on future states. Given a random sample $\{s_{t+1}, \ldots, s_T\}$ drawn from $p(x_{t+1}, \ldots, x_T|y^T)$, we take one step back and sample s_t from $p(x_t|s_{t+1}, \ldots, s_T, y^T)$. The smoothing algorithm is summarized in the appendix in the context of the various versions of our model.

4.3 MCEM algorithm

Frequentist likelihood-based parameter estimation of conditional linear and Gaussian state space models using the mixture Kalman filters and smoothers is not straightforward. The gradient-based optimizer suffers from a discontinuity problem caused by the resampling. Here, we follow the Monte Carlo Expectation Maximization (MCEM) method proposed by

Olsson et al. (2008). The Basic EM algorithm is a general method to obtain the maximum-likelihood estimates of the parameters of an underlying distribution from a given data set with missing values. Suppose the complete data set is $Z = (Y, X)$, in which Y is observed but X is unobserved, and Θ is the parameter vector. For the joint density $p(z|\Theta) = p(y, x|\Theta) = p(y|\Theta)p(x|y, \Theta)$, we define the complete-data likelihood function by $L(\Theta|Y, X) = p(Y, X|\Theta)$. The original likelihood $L(\Theta|Y)$ is the incomplete-data likelihood. Since X is unobserved and may be generated from an underlying distribution, e.g., the transition equation in a state space model, $L(\Theta|Y, X)$ is indeed a random variable. Therefore, we maximize the expectation of $logL(\Theta|Y, X)$ with respect to X, with the expectation, conditional on Y and some input value for the parameters $\Theta^{(k-1)}$, defined by:

$$Q(\Theta, \Theta^{(k-1)}) = E[logL(\Theta|Y, X)|Y, \Theta^{(k-1)}] = \int logp(Y, x|\Theta)p(x|Y, \Theta^{(k-1)})dx$$

which will permit an iterative procedure to update the values of the parameters Θ. The difference between the MCEM algorithm and the basic EM algorithm is that when evaluating $Q(\Theta, \Theta^{(k-1)})$, the MCEM uses a Monte-Carlo based sample average to approximate the expectation. The Monte Carlo Expectation or E-step is:

$$Q^*(\Theta, \Theta^{(k-1)}) = \frac{1}{M} \sum_{i=1}^{M} log(p(Y, x^{(i)}|\Theta))$$

where $\{x^{(i)}\}_{i=1}^{M}$ are random samples from $p(x|Y, \Theta^{(k-1)})$. Given current parameter estimates, random samples from $p(x|Y, \Theta^{(k-1)})$ are simply the particle smoothers $\{s_t^{(i)}\}_{i=1}^{M}$ obtained as described above. The Maximization or M-step is:

$$\Theta^{(k)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(k-1)})$$

These two steps are repeated until $\Theta^{(k)}$ converges. The rate of convergence has been studied by many researchers; e.g., Dempster et al. (1977), Wu (1983) and Xu and Jordan (1996). In the context of the simple version of our model, the specifics of the algorithm are in the appendix.

Overall, the estimation procedure is summarized as the following steps: Let $\Theta^{(0)}$ be a vector of initial parameter values

1. (Mixture Kalman filtering): obtain mixture Kalman filters $\{x_t^{(i)}\}_{i=1}^{M}$ from $p(x_t|y^t, \Theta^{(k-1)})$, $i = 1, 2, ..., M$, $t = 1, 2, ..., T$;

2. (Particle smoothing): obtain particle smoothers $\{s_t^{(i)}\}_{i=1}^M$ from $p(x_t|y^T, \Theta^{(k-1)})$, $i = 1, 2, ..., M$, $t = 1, 2, ..., T$;

3. (Estimation): evaluate $Q(\Theta, \Theta^{(k-1)})$ using $\{s_t^{(i)}\}_{i=1}^M$ from the previous step and maximize it to obtain updated parameter estimates $\Theta^{(k)}$;

4. Repeat steps 1-3 with k updated to $k+1$ until the parameter estimates converge.

5 In-sample confidence bands and out-of-sample forecast bands

In this section, we propose a simulation based method to construct in-sample confidence bands and out-of-sample forecast bands following a modification of the method proposed by Blasques et al. (2016) who dealt with observation-driven time varying parameter models, for which the observations $\{y_t\}_{t=1}^T$ are given by $y_t \sim p(y_t|\beta_t; \theta)$. In this case, the time-varying parameter β_t follows the updating equation:

$$\beta_{t+1} = \phi(\beta_t, y_t; \theta)$$

where $\phi(.)$ is a differentiable recurrence function and θ is the static parameter. The framework of this paper does not fit in their analysis since it is a parameter-driven time varying model. Perron and Xu (2016) pointed out that the updating equation (process) for the time-varying parameters in parameter-driven models can be written as:

$$\beta_{t+1} = h(\beta_t, \eta_t; \psi)$$

where $\eta_t \sim g_\eta(\psi)$ is the idiosyncratic innovation and ψ is the static parameter. The time-varying parameter β_t follows a recurrence process with its own innovations. Therefore, the in-sample confidence bands need to incorporate both parameter uncertainty and innovation uncertainty. The parameter estimate $\hat{\psi}$ is constructed via Monte Carlo maximum likelihood estimation. Let the estimate of the asymptotic covariance matrix of $\hat{\psi}$ be defined by $\hat{\Sigma} = -\{\partial^2 \log \hat{L}(\hat{\psi})/\partial\psi\partial\psi'\}^{-1}$, where $\hat{L}(\hat{\psi})$ is the Monte Carlo estimate of the likelihood function evaluated at $\hat{\psi}$. The estimate $\hat{\Sigma}$ can be computed numerically. Once an estimate of the asymptotic distribution of $\hat{\psi}$ is obtained, the in-sample confidence bands for $\hat{\beta}_{t+1}$ can be constructed using simulation methods similar to the filtering forecast band method proposed in Blasques et al. (2016). The procedure can be described as follows:

1. Draw M parameter values $\hat{\psi}^{(i)}$ from the asymptotic distribution $\hat{\psi}^{(i)} \sim N(\hat{\psi}, T^{-1}\hat{\Sigma})$; see Olson and Ryden (2008);

2. Given $\hat{\psi}^{(i)}$, and for each time t, draw S sequences $\eta_t^{(1)}, ..., \eta_t^{(S)}$ from the estimated density $\eta_t^{(s)} \sim g_\eta(\hat{\psi}^{(i)})$ for $s = 1, ..., S$ and $t = 1, ..., T$;

3. Given the observations $\eta_t^{(1)}, ..., \eta_t^{(S)}$, the filtered sequence $\hat{\beta}_1^{(s)}, ..., \hat{\beta}_T^{(s)}$ can be determined using the updating function $\hat{\beta}_{t+1}^{(s)} = h(\hat{\beta}_t^{(s)}, \eta_t^{(s)}; \hat{\psi}^{(i)})$;

4. Repeat steps 2-3 for $i = 1, ..., M$ to obtain $M \times S$ filtered paths of $\hat{\beta}_t^{(i),(s)}$;

5. Calculate the appropriate percentiles for each t over the $M \times S$ draws of $\hat{\beta}_t^{(i),(s)}$ to obtain the in-sample confidence bands for $\hat{\beta}_t$.

The procedure to construct the out-of-sample forecast bands for $\hat{\beta}_{t+h}$ is actually the same as described above. We simply need to obtain $M \times S$ extrapolated paths of $\hat{\beta}_{t+h}^{(i),(s)}$ to compute the percentiles. To illustrate, we again use a simple example. The true process for β_t is generated using equations (1) and (2) with mean reversion and time varying probability with the parameters $(r_0,\ r_1,\ \sigma_e,\ \sigma_\eta,\ \rho) = (-1.96, 4, 0.2, 0.2, -0.1)$. The computation of the in-sample and out-of-sample bands are based on $M = 1000$ and $S = 1000$ simulations. The number of observations is 1000 when considering in-sample bands and 500 for out-of sample bands (given the higher computational burden). Panel A of Figure 3 presents the true process β_t, the filtered estimates and the 2.5% ad 97.5% quantiles of the simulated distribution. The confidence bands are quite narrow around the true process showing precisely estimated parameters. Panel B presents the results for the out-of-sample confidence bands. We use the first 300 observations to obtain the parameter estimates. The out-of-sample forecasting starts from the 301th observation. The forecasting horizon is set to be 100 steps. The figure shows the forecasts and the 2.5% ad 97.5% quantiles of the simulated distribution.

6 Simulations

This section aims to demonstrate the reliability and robustness of RLS type models in forecasting even when the model is misspecified. In the simulation setup, we consider eight Data Generating Processes (DGPs).

1. RLS basic model: $y_t = \beta_t + e_t$, with $\beta_t = \beta_{t-1} + K_t\eta_t$, where $K_t \sim Ber(p)$, $e_t \sim N(0, \sigma_e^2)$, $\eta_t \sim N(0, \sigma_\eta^2)$. We set the true parameters to be $\theta = (p, \sigma_e, \sigma_\eta) = (0.05, 0.2, 0.2)$.

2. RLS with mean reversion: The model is the same as in (1), except that the probability of shifts is now a function of some covariate w_t and β_t follows a mean reverting process; i.e., $p_t = \Phi(r_0 + r_1 w_t)$, $\eta_t \sim N(\mu_{\eta_t}, \sigma_\eta^2)$, $\mu_{\eta_t} = \rho(\beta_{t-1} - \bar{\beta}^{(t-1)})$. The true parameters are $\theta = (r_0, r_1, \sigma_e, \sigma_\eta, \rho) = (-1.96, 4, 0.2, 0.2, -0.1)$. The covariate w_t is set to be 1 every

50 observations, 0 otherwise. Doing so, we intentionally set the probability of level shifts to be small most of the time and close to 1 every 50 periods.

3. RLS without mean reversion: The model is that same as in (2), except that the mean reversion parameter ρ is set to be 0.

4. RLS_SV: The model is the same as in (2), except that we add stochastic volatility to the error term of the form $e_t = \sigma_{\varepsilon,t}\varepsilon_t$, $\varepsilon_t \sim N(0,1)$, with $ln\sigma^2_{\varepsilon,t} = \phi ln\sigma^2_{\varepsilon,t-1} + v_{\varepsilon,t}$, where $v_{\varepsilon,t} \sim N(0,\sigma^2_v)$ and independent of ε_t and e_t. The true parameters are $\theta = (r_0, r_1, \phi, \sigma_v, \sigma_\eta, \rho) = (-1.96, 4, 0.95, 0.2, 0.2, -0.1)$.

5. ARMA(1,1) (Autoregressive and Moving Average process): $(1-\phi L)y_t = (1+\theta L)\varepsilon_t$, $\phi = 0.95$, $\theta = -0.5$ and $\varepsilon_t \sim N(0,1)$.

6. ARIMA(1,1,1) (Autoregressive Integrated and Moving Average process): $(1-L)(1-\phi L)y_t = (1+\theta L)\varepsilon_t$, $\phi = 0.1$, $\theta = -0.5$ and $\varepsilon_t \sim N(0,1)$.

7. TVP (Time Varying Parameter Model): $y_t = \beta_t + e_t$ with $\beta_t = \beta_{t-1} + \eta_t$, where $e_t \sim N(0,\sigma^2_e)$ and $\eta_t \sim N(0,\sigma^2_\eta)$ independent of each other. The true parameters are set to be $(\sigma_e, \sigma_\eta) = (0.2, 0.2)$.

8. Markov Switching (MS): We apply a two states regime switching model (e.g., Hamilton, 1994): $y_t = \mu_{S_t} + e_t$, where $e_t \sim N(0,\sigma^2_{S_t})$, $S_t = 1, 2$. Here we assume $[\mu_1, \mu_2] = [0.5, -0.5]$, $[\sigma^2_1, \sigma^2_2] = [1, 2]$ and the transition matrix from state i to state j for $i, j = 1, 2$ is given by:

$$P = \begin{bmatrix} 0.95 & 0.1 \\ 0.05 & 0.9 \end{bmatrix}.$$

In each case, we generate 100 true data paths and 1000 observations for each path. We use the first 800 observations for in-sample estimation and the rest to evaluate out-of-sample forecasting accuracy. The forecasting horizon is up to 60 periods. The forecasting models considered are: the 'RLS_m': the RLS model with mean reversion and time varying probability; 'RLS_SV': the RLS model with mean reversion, time varying probability and stochastic volatility; 'Average': the historical average, namely the average over all observations in the expanding in-sample period; 'Rolling': the average of the last 50 observations of the in-sample period; 'ARMA': an ARMA(1,1) model; 'ARIMA': an ARIMA(1,1,1) model; 'MS': a Markov switching model as described in DGP 8; 'TVP' a Time Varying Parameter

Model as specified in DGP 7. For DGP (1), we also consider the basic RLS model without mean reversion, nor time varying probability, which acts as the benchmark model. For each DGP, we report the relative MSFEs of some other misspecified models with respect to the benchmark model, which is, in all cases, the true model with estimated parameters. The results are summarized in Table 1. Numbers smaller than 1 indicate a better forecasting performance than that obtained with the corresponding true model. Bold numbers indicate the smallest relative cumulative MSFEs for a given DGP and forecast horizon.

Consider first the results in Panels 1-4, for which some type of RLS model is the true DGP. With few exceptions, the best performing forecasting model is the 'RLS_m'. In the few cases for which it is not the best, the preferred one is the 'RLS_SV' for long forecast horizons for DGP-2. The difference are, however, minor between the two. What is especially interesting is that introducing a mean reverting component even when not present leads to better forecasts, see Panels 1 and 3. The 'TVP' and 'Markov Switching' models perform poorly, especially at long-horizons. The 'ARMA' and 'ARIMA' models perform quite well but still produce inferior forecasts compared to the 'RLS_m'. The 'historical average' is prone to severe deficiencies; e.g Panel 1. The 'rolling average' has about twice the RMSE of 'RLS_m' in most cases.

From panels 5-8, even when the true DGP is not RLS, the RLS type models still have robust or even better performance compared to the benchmark model. The 'RLS_m' or 'RLS_SV' are second best (relative to the benchmark model) in most cases. As seen in panel 8, when the true DGP is a two states Markov switching process, the forecasting performances of the RLS models are much better than those of the true model. In cases of model misspecifications, the performances of the various alternative models considered can be very poor; e.g. DGPs 5 and 7 for the 'historical average' and the 'rolling window average', DGP 8 for 'TVP' and DGP 6 for 'Markov Switching'. As for the 'ARMA' and 'ARIMA' models, the performances are considerably robust but still worse than the RLS type models especially under model misspecification.

The results show that our random level shift model with built-in mean reversion always performs nearly as well as the model corresponding to the true DGP, and can even be better (e.g., when the true DGP is ARIMA or Markov Switching). All other forecasting methods perform very poorly in one or more of the cases considered. Hence, our method provides reliable results that are robust to a wide range of processes.

7 Forecasting applications

We consider two forecasting applications pertaining to variables which have been the object of intense attention in the literature: the equity premium and the Treasury Bill rates. The emphasis is on the equity premium. We compare the forecast accuracy of our model relative to the most important forecasting methods applicable for this variable. For this particular series, it turns out that the Time Varying Parameter Model (TVP) performs quite well being a close second best. As shown in the simulations, the TVP model is not robust to a variety of DGPs, while our method is. To illustrate this feature, we also consider the Treasury Bill rate. Our method continues to provide the best forecasts overall, while the TVP model leads to very poor forecasts in most samples considered.

The out-of-sample forecasts are constructed in two steps. The first involves forecasting the covariates w_t using a preliminary model; e.g., using an $AR(k)$ or the random level shift model with a fixed probability of shift. The h-step ahead forecast of the jump probability is then $p_{t+h|t} = \Phi(\hat{r}_0 + \hat{r}_1 w_{t+h|t})$ where $w_{t+h|t}$ is the h-step ahead forecast of w_{t+h} at time t and $(\hat{r}_0, \hat{r}_1)$ are the parameter estimates. Note that one can also forecast the regressors X_t to obtain predicted values denoted by $X_{t+h|t}$. In the applications, we use forecast values for X_{t+h} and w_{t+h} using an $AR(p)$ model with p selected using the Akaike Information Criterion (AIC) with a maximal value of 4.

The second step is to forecast $\{\beta_{t+s}\}_{s=1}^{h}$. The 1-step-ahead forecast is calculated as $\beta_{t+1|t} = E[\beta_{t+1}|I_t] = \sum_{i=1}^{M} w_t^{(i)} f(\beta_{t+1|t}^{(i)})$, where $\beta_{t+1|t}^{(i)}$ is obtained via the Kalman filtering steps. For s step-ahead forecasts, $\beta_{t+h|t} = E[\beta_{t+h}|I_t]$ can be calculated recursively by repeating the filtering algorithm from time $t+1$ to $t+h$, and treating the observations $\{y_{t+s}\}_{s=1}^{h}$ as missing values. We can continue to apply the above algorithm setting $v_t = 0, K_t = 0$ for $t = t+1, ..., t+h$.

Throughout, the out-of-sample forecasting experiments aim at evaluating the experience of a real-time forecaster by performing all model specifications and estimations using data through date t, making a h-step ahead forecast for date $t+h$, then moving forward to date $t+1$ and repeating this through the sub-sample used to construct the forecasts. Unless otherwise indicated, the estimation of each model is recursive, using an increasing data window starting with the same initial observations. The forecasting performance is evaluated using the mean square forecast error (MSFE) criterion defined as

$$MSFE(h) = \frac{1}{T_{out}} \sum_{t=1}^{T_{out}} (\bar{y}_{t,h} - \bar{y}_{t+h|t})^2$$

where T_{out} is the number of forecasts produced, h is the forecasting horizon, $\bar{y}_{t,h} = \sum_{k=1}^{h} y_{t+k}$ and $\bar{y}_{t+h|t} = \sum_{k=1}^{h} y_{t+k|t}$ with y_{t+k} the actual observation at time $t+k$ and $y_{t+k|t}$ its forecast conditional at time t. To ease presentation, the MSFE are reported relative to some benchmark model, usually the most popular forecasting model in the literature. In all cases, we allow mean reversion in the parameters when constructing forecasts using our RLS model.

Remark 4 *The cumulative MSFE defined above gives the same relative measure of forecast performance as root mean squared errors. Our interest is not in the absolute level, so it makes no difference.*

7.1 Equity premium

Forecasts of excess returns at both short and long-horizons are important for many economic decisions. Much of the existing literature has focused on the conditional return dynamics and studied the implications of structural breaks in regression coefficients including the lagged dividend yield, short-term interest rate, term spread and the default premium. However, most of the research has focused on modeling the equity premium assuming a certain number of structural breaks in-sample while ignoring potential out-of-sample structural breaks. Recently, Maheu and McCurdy (2009) studied the effect of structural breaks on forecasts of the unconditional distribution of returns, focusing on the long-run unconditional distribution in order to avoid model misspecification problems. Their empirical evidence strongly argue against ignoring structural breaks for out-of-sample forecasting. We consider using our forecasting model with different specifications. One models the unconditional mean of excess returns incorporating random level shifts in mean, with the time varying jump probabilities influenced by the lagged value of the absolute rate of growth in the earning price (EP) ratio. We also consider a conditional mean model using the dividend yield as the explanatory variable.

Following Jagannathan et al. (2000), we approximate the equity premium of S&P 500 returns as the difference between stock yield and bond yield. The data were obtained from Robert Shiller's website (http://www.econ.yale.edu/~shiller/data.htm). According to Gordon's valuation model, stock returns are the sum of the dividend yields and the expected future growth rate in stock dividends. We use the average dividend growth rate (over the pre-forecasting sample) to proxy for the expected future growth rate. The data consist of monthly series and cover the period from 1871:1 to 2012.5. High quality monthly data are available after 1927, before 1927 the monthly data are interpolated from lower frequency data. We use the 10-years Treasury constant maturity rate (GS10) as the risk free rate.

We start with a simple random level shift model without explanatory variables given by:

$$\begin{aligned} y_t &= \beta_t + e_t \\ \beta_t &= \beta_{t-1} + K_t^{\beta}\eta_t \end{aligned} \tag{6}$$

where $e_t \sim i.i.d.N(0, \sigma_e^2)$, $\eta_t \sim i.i.d.N(\mu_{\eta_t}, \sigma_\eta^2)$, $\mu_{\eta_t} = \rho(\beta_{t-1} - \bar{\beta}^{(t-1)})$, $K_t^{\beta} \sim Ber(p_t)$ with $p_t = \Phi(r_0 + r_1 w_t)$. The covariate w_t used to model the time variation in the probability of shifts is the lagged absolute value of the rate of change in the EP ratio. The rational for doing so is that it is expected that large fluctuations in the EP ratio induce a higher probability that excess stock returns will experience a level shift in the unconditional mean.

We also consider a conditional forecasting model that uses the lagged dividend price ratio as the regressor. The specifications are

$$y_t = \beta_{1t} + \beta_{2t} dp_{t-1} + e_t \tag{7}$$

where, with $\beta_t = (\beta_{1t}, \beta_{2t})$, $\beta_t = \beta_{t-1} + K_t^{\beta}\eta_t$, and dp_t is the dividend-price ratio. Lettau and van Nieuwerburgh (2008) analyzed the implications of structural breaks in the mean of the dividend price ratio for conditional return predictability. Xia (2001) studied model instability using a continuous time model relating excess stock returns to dividend yields. They specify β_t to follow an Ornstein–Uhlenbeck process and the ensuing estimates of the time varying coefficient β_{2t} revealed instability of the forecasting relationship. Hence, instabilities have been shown to be of concern when using this conditional forecasting model, which motivates the use of our forecasting model. Besides the addition of the lagged dividend price ratio as regressors, the specifications are the same as for the unconditional mean model (6).

We consider various versions depending on which coefficients are allowed to change and if so whether they change at the same time. These are: 1) the unconditional mean model (6) with level shifts, 2) the conditional mean model (7) with the constant allowed to change ($K_{1t}^{\beta} \neq 0, K_{2t}^{\beta} = 0$), 3) the conditional mean model (7) with the coefficient on the lagged dividend yield allowed to change ($K_{1t}^{\beta} = 0, K_{2t}^{\beta} \neq 0$). We compare our forecasting model with the most popular forecasting models used in the literature. These are: 1) the historical average (used as the benchmark model); 2) a rolling ten-years average; 3) the conditional model with the lagged dividend price ratio as the regressor without changes in the parameter; 4) a rolling version over ten years of the model previously stated in 3); 5) a TVP model with the unconditional mean following a random walk; 6) a two-states regime switching model.

We first consider 1998-2012 as the forecasting period, with forecasting horizons 1, 6, 12, 18, 24, 30 and 36 months. The results are presented in Table 2.1. The first thing to note is

that all three versions involving random level shifts perform very well and are comparable. The best model for short horizons less than 6 months is the conditional mean model (7) with the constant allowed to change ($K_{1t}^{\beta} \neq 0, K_{2t}^{\beta} = 0$), though the difference are quite minor. For longer horizons, the conditional mean model (7) with the coefficient on the lagged dividend yield allowed to change ($K_{1t}^{\beta} = 0, K_{2t}^{\beta} \neq 0$) is the best. What is noteworthy is that our model performs much better than any competing forecasting models except the TVP model. This is especially the case at short-horizons, for which the gain in forecasting accuracy translates into a reduction in MSFE of up to 90% when compared to the conditional model with no breaks (and even more so when compared to the rolling 10 years average or the historical average, the latter performing especially badly). At longer horizons, the conditional mean model (7) with level shifts still perform better than the conditional model with constant coefficients but to a lesser extent. The rolling version of the dividend price ratio model performs better than the one using the full sample for short horizons but less so at long horizons. In no case is it better than any of the versions with random level shifts. We also provide p-values from the Model Confidence Set (MCS) of Hansen et al (2011) with p-values greater than 0.1 indicating that the corresponding model belongs to the 10% model confidence set. The conditional mean model with the coefficient on the lagged dividend yield allowed to change ($K_{1t}^{\beta} = 0, K_{2t}^{\beta} \neq 0$) belongs to the MCS for all forecasting horizons. Other RLS type models and the TVP model belong to the MCS for 1-step-ahead forecasts. This can be viewed as strong evidence that the performance of our RLS model is superior and dominant in forecasting the equity premium compared to most popular candidates in the literature.

To assess the robustness of the results we also consider the forecasting period 1988-1996, given that it offers an historical episode with different features; see Table 2.2. What is noteworthy is that the conditional mean model with constant parameters now performs very poorly with MSFEs more than four times those of the rolling 10 years average. The benchmark historical average performs even worse during this time period. On the other hand, the models with random level shifts continue to perform very well, with MSFEs around 0.2% of the historical average at short horizons, and around 2.5% at longer horizons up to 60 months (i.e., five years). All models with random level shifts have comparable performance at short horizons, but the unconditional mean model (6) with level shifts is best at longer horizons. Meanwhile, the TVP model is also a strong candidate being the best at very short horizons and remaining in the 10% confidence set for all horizons. The conditional mean model with the coefficient on the lagged dividend yield allowed to change also belongs to the

10% MCS for horizons longer than 6 steps.

Given that the results show very impressive improvements in forecast accuracy using our forecasting method and the fact that forecasting the equity premium is important, we performed further sensitivity analyses using a different data set. The data are the same as used in Welch and Goyal (2008) and were downloaded from Amit Goyal's website (http://www.hec.unil.ch/agoyal/). It is also the same dataset used by Pettenuzzo et al. (2014). As we will show, the improvements in forecast accuracy using our framework continue to be large whatever the sampling intervals used (yearly or monthly). We follow the common practice of simply calculating the equity premium as the historical average difference between returns on stocks and returns on risk-free assets. Annualized equity premiums are calculated from monthly data as compounded excess returns. Goyal's data set goes back to 1871 and includes monthly and annual data. The most recent updated dataset ends in 2014. One advantage of using Goyal's dataset is that there are also many other economic variables available for a long span of time. Here, we use the book-to-market ratio as the regressor to help predict the equity premium, which is one of the three factors in Fama and French (1993) three factors model. The Book-to-Market Ratio (BM) is the ratio of book value to market value for the Dow Jones Industrial Average. The covariate used to forecast the level shift probability is the absolute change in the earning-price ratio. In all cases, we use the historical average as the benchmark model, which is claimed as a competitive candidate to beat when forecasting equity premium. To analyze the effect of the oil shock, the forecast period is from 1975 to 2009 for both annual and monthly data.

The results are presented in Tables 3.1 (annual) and 3.2 (monthly). The results show that all three models we propose beat the benchmark for all sampling intervals. We obtain 80% reduction in cumulative MSFE for one-step-ahead forecasts and 58% for longer horizon forecasts. The improvement in forecast accuracy becomes larger when higher frequency data are used. For monthly data, we get 98% reduction in MSFE for short horizon forecasts and almost 74% reduction for long horizon forecasts. We notice that in Table 3.1 (annual data) almost all competing models belong to the 10% model confidence set except for the rolling 10 years average and the constant parameter model with the BM ratio as a regressor for longer horizons forecasts. On the other hand, as seen in Table 3.2 (monthly data) the TVP model for short horizons forecasts and the constant parameter model with a rolling estimation window for longer horizons forecasts are the only competing models that remain in the 10% MCS. With annual data, the total number of observations for out-of-sample forecasting is 34, which is considerably less than for the monthly case. The small samples available for MCS

testing is likely the reason which makes it more difficult to select the "competing" models. Nevertheless, looking at the relative MSFEs indicates relative performances similar to those in Tables 2.1 and 2.2

In addition, we also looked for other regressors such as financial variables with the purpose of further improving forecasting accuracy. Due to the short span of the financial data, we use monthly data from 1990/01/31 to 2008/12/31 for in-sample estimation and forecast from 2009 to 2015 at horizons up to 24 months. The financial variables are the VIX index from the Chicago Board Options Exchange (CBOE) and the returns on the monthly S&P 500 index option. The results are reported in Table 4, which indicate a gradually improving forecasting performance for horizons longer than 12 months with these two additional regressors. For a forecasting horizon of 24 months, the cumulative MSFE of the conditional mean model including financial variables is only half of the MSFE of the conditional mean model with only the dividend-price ratio.

In summary, the evidence provides strong evidence that our forecasting model offers marked improvements in forecast accuracy. It does so at all horizons with results that are robust to different forecasting periods and different data sets. It remains that the TVP model is a close second best forecasting model for the equity premium series analyzed. According to the simulations, the TVP model is much less robust to model mispecification than the RLS-type models. To illustrate this issue, we next consider the issue of forecasting the interest rate.

7.2 Interest rate

Another variable of interest, which has attracted attention from a forecasting perspective, is the U.S. T-bill rate. Various studies have shown that it exhibits structural instability in both mean and variance, see, e.g., Garcia and Perron (1996), Gray (1996), Ang and Bekaert (2002) and Pesaran et al. (2006). We use monthly data on the 3-months Treasury Bill rate from 1947:07-2002:12, obtained from the Federal Bank of St. Louis database. Our data is the same as used in Pesaran et al. (2006). The period prior to 1968:12 is used for in-sample estimation, and we consider forecasting horizons of 12, 24, 36, 48 and 60 months. The basic model adopted is a simple AR(1) process given by:

$$y_t = \beta_{1t} + \beta_{2t} y_{t-1} + \epsilon_t$$

In all cases, we allow mean-reversion in the parameters and the covariate w_t used to model the time-varying probabilities of shifts is the lagged value of the growth rate of GDP when a

single latent Bernoulli variable is present. When two are present, the additional covariate is the lagged value of the absolute change in stock returns ($S\&P$ 500). We consider two possible specifications: 1) 'AR_K_{2t}' ($K_{1t}^{\beta} = 0$); 2) 'AR_K_{1t}, K_{2t}' with K_{1t}^{β} and K_{2t}^{β} allowed to be different latent Bernoulli processes. The performance of the models is assessed relative to four competing forecasting methods[1]: 1) 'Recursive OLS': a recursively estimated first-order autoregression with fixed parameters, used as the benchmark model; 2) 'Rolling 5 years': a first-order autoregression with fixed parameters estimated using a 5-year rolling window; 3) 'Rolling 10 years': a first-order autoregression with fixed parameters estimated using a 10-year rolling window; 4) 'TVP': a time-varying probability model in which $\beta_t = (\beta_{1t},\ \beta_{2t})$ is modelled as a random walk.

The results are presented in Table 5 for various forecast periods and forecast horizons $h = 12, 24, 36, 48, 60$ months. Consider first the results for the longest forecasting period 1968-2002. Here, the best forecasting model for all horizons is the 'AR_K_{1t}, K_{2t}' with both the constant term and the AR coefficient allowed to follow a random level shift process. The gains in forecast accuracy vary between 3% and 8% and increase as the forecasting horizon increases. We then separate the forecasting period into three decades: the 70s, the 80s and the 90s. In the 70s, the 5 years rolling-average is overall the best predictor, though the 'AR_K_{1t}, K_{2t}' model catches up and and is superior at $h = 60$. For the 80s and 90s, the best forecasting model is again the 'AR_K_{1t}, K_{2t}' with both the constant term and the AR coefficient in the AR regression allowed to follow a random level shift process. The improvements in forecast accuracy are on average 5% reduction in MSFE. The 'AR_K_{1t}, K_{2t}' performs the best in 16 out of 20 cases and is in the 10% MCS for 17 out of 20 cases. Note, however, that all models with random level shifts in parameters perform much better than the 'TVP' model in different time periods and different horizons. The gains in forecast accuracy when using the RLS-type models over the 'TVP' model are substantial. For the full sample, they range from a (roughly) 50% reduction at the shortest horizon to a 90% reduction at the longest one. For the 90s sub-sample, the corresponding reductions are of the order of 65% to 96%. The only period in which the 'TVP' model does not perform badly is the 70's, but it is still inferior to the RLS models and not within the 10% MCS. These

[1]We also compared the forecasting performance using Pesaran et al. (2006)'s composite and last regime model proposed in their paper. Our RLS model performs better in most cases for forecasts computed every 12 months as in their paper. We do not include their model in our comparisons set due to computational constraints. Their method is highly computationally intensive and we could not apply it to forecasts computed every months. The sample obtained using forecasts computed every 12 months makes is too small for a valid MCS testing.

results are consistent with the simulations reported in Section 6, which show the RLS-type model to be robust to a wide range of DGPs, while the other forecasting methods are not and can produce very poor forecasts for some DGPs. Overall, the evidence again indicates that important gains in forecast accuracy can be obtained using our forecasting models and that they are robust in the sense that in no case do they perform substantially worse than popular forecasting methods.

8 Conclusion

We proposed a forecasting framework based on modeling the parameters as random level shift processes dictated by a Bernoulli process for the occurrence of shifts and a normal random variable for its magnitude. Some or all of the parameters of the model can be allowed to change and the latent variables that dictate the changes can be common or different for each parameters. Also, the variance of the errors may change in a similar manner. To improve the forecasting performance we augmented the basic model to allow the probability of shifts to be a function of some covariates which can be forecasted and to incorporate a mean-reversion mechanism such that the parameters tend to revert back to the pre-forecast average. Our model can be cast into a conditional linear and Gaussian state space framework for which standard Kalman filter type algorithms cannot be used. To provide a computationally efficient method of estimation, we rely on recent developments on mixture Kalman filtering methods. Simulations show that the proposed model has robust forecasting performance even under model misspecification.

We applied our forecasting model to the equity premium and the Treasury bill interest rates. In each case, we compare the forecast accuracy of our model relative to the most important forecasting methods used applicable for each variable. We also consider different forecasting sub-samples or periods. The results show clear gains in forecasting accuracy, sometimes by a very wide margin.

Finally, note that given the availability of the proper code for estimation and forecasting, the method is very flexible and easy to implement. For a given forecasting model, all that is required by the users are: 1) which parameters (including the variance of the errors if desired) are subject to change; 2) whether the same or different latent Bernoulli processes dictate the timing of the changes in each parameters; 3) which covariates are potential explanatory variables to model the probability of shifts.

References

Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821-856.

Ang, A., Bekaert, G., 2002. Regime switches in interest rates. Journal of Business and Economic Statistics 20, 163-182.

Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica 66, 47-78.

Bai, J., Perron, P., 2003. Computation and analysis of multiple structural change models. Journal of Applied Econometrics 18, 1-22.

Bilmes, J.A., 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Working Paper, International Computer Science Institute.

Blasques, F., Koopman, S.J., Katarzyna, L., Lucas, A., 2016. In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation driven models. International Journal of Forecasting 32, 875–887.

Chen, R., Liu, J.S., 2000. Mixture Kalman filters. Journal of Royal Statistical Society Series B 62, 493-508.

Chow, G.C., 1984. Random and changing coefficient models. In Griliches, Z., Intriligator, M. (Eds.), Handbook of Econometrics, vol. 2. North Holland, Amsterdam, 1213-1245.

Chib, S., 1998. Estimation and comparison of multiple change-point models. Journal of Econometrics 86, 221-241.

Clements, M. and Hendry, D.F., 2006. Forecasting with breaks. In Elliott, G., Granger, C.W.J., Timmermann, A., (Eds.), Handbook of Economic Forecasting, Elsevier Science, Amsterdam, 605-657.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum-likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B 39, 1-38.

Duan, J., Fulop, A., 2011. A stable estimator of the information matrix under EM for dependent data. Statistics and Computing 21, 83-91.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3-56.

Garcia, R., Perron, P., 1996. An analysis of the real interest rate under regime shifts. Review of Economics and Statistics 78, 111-125.

Gerlach, R., Carter, C., Kohn, R., 2000. Efficient Bayesian inference for dynamic mixture models. Journal of the American Statistical Association 95, 819-828.

Giordani, P., Kohn, R., 2008. Efficient Bayesian inference for multiple change-point and mixture innovation models. Journal of Business and Economic Statistics 26, 66-77.

Giordani, P., Kohn, R., van Dijk, D., 2007. A unified approach to nonlinearity, structural change, and outliers. Journal of Econometrics 137, 112-133.

Godsill, S.J., Doucet, A., West, M., 2004. Monte Carlo smoothing for nonlinear time series. Journal of the American Statistical Association 99, 156-168.

Gray, S.F., 1996. Modeling the conditional distribution of interest rates as regime-switching process. Journal of Financial Economics 42, 27-62.

Groen, J.J.J., Paap, R., Ravazzolo, F., 2013. Real-time inflation forecasting in a changing world. Journal of Business and Economic Statistics 31, 29-44.

Hamilton, J.D., 1994. Time Series Analysis. Princeton: Princeton University Press.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79, 453–497.

Harvey, A.C., 2006. Forecasting with unobserved components time series models. In Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting, Elsevier Science, Amsterdam, 327-408.

Hauwe, S., Paap, R., van Dijk, D., 2011. An alternative Bayesian approach to structural breaks in time series models. Tinbergen Institute Discussion Paper.

Hull , J., White, A., 1987. The Pricing of options on assets with stochastic volatilities. Journal of Finance 42, 281-300.

Jagannathan, R., McGrattan, E.R., Scherbina, A., 2000. The declining U.S. equity premium. Federal Reserve Bank of Minneapolis Quarterly Review 24(4): 3-19.

Koop, G., Potter, S.M., 2007. Estimation and forecasting in models with multiple breaks. Review of Economics Studies 74, 763-789.

Lettau, M., Van Nieuwerburgh, S., 2008. Reconciling the return predictability evidence. Review of Financial Studies 21, 1607-1652.

Li, Y., Perron, P., Xu, J., 2016. Modelling exchange rate volatility with random level shifts. Forthcoming in Applied Economics.

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society. Series B 44, 226-233.

Lu, Y.K., Perron, P., 2010. Modeling and forecasting stock return volatility using a random level shift model. Journal of Empirical Finance 17, 138-156.

Maheu, J.M., Gordon, S., 2008. Learning, forecasting and structural breaks. Journal of Applied Econometrics 23, 553-583.

Maheu, J.M., McCurdy, J.H., 2009. How useful are historical data for forecasting the long run equity return distribution? Journal of Business and Economic Statistics 27, 95-112.

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

Nicholls, D.F., Pagan, A.R., 1985. Varying coefficient regression.In Hannan, E.J., Krishnaiah, P.R., Rao, M.M. (Eds.), Handbook of Statistics, vol. 5. North-Holland, Amsterdam, 413-450.

Olsson, J., Cappe, O., Douc, R.,Moulines, E., 2008. Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space models. Bernoulli 14,155–179.

Olsson, J., Ryden, T., 2008. Asymptotic properties of particle filter-based maximum likelihood estimators for state space models. Stochastic Processes and their Applications 118, 649–680.

Pastor, L., Stambaugh, R.F., 2001. The equity premium and structural breaks. Journal of Finance 4, 1207-1231.

Paye, B.S., Timmermann, A., 2006. Instability of return prediction models. Journal of Empirical Finance 13, 274-315.

Perron, P., 2006. Dealing with structural breaks. In Patterson, K., Mills, T.C. (Eds.), Palgrave Handbook of Econometrics, Vol. 1: Econometric Theory, Palgrave Macmilan, Basingstoke, 278-352.

Perron, P., Qu, Z., 2010. Long-memory and level shifts in the volatility of stock market return indices. Journal of Business and Economic Statistics 28, 275-290.

Perron, P., Xu, J., 2016. Comments on "In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation driven models". International Journal of Forecasting 32, 891-892.

Pesaran, M.H., Pettenuzzo, D., Timmermann, A., 2006. Forecasting time series subject to multiple structural breaks. Review of Economic Studies 73, 1057-1084.

Pesaran, M.H., Timmermann, A., 2002. Market timing and return prediction under model instability. Journal of Empirical Finance 9, 495-510.

Pettenuzzo, D., Timmermann, A., 2011. Predictability of stock returns and asset allocation under structural breaks. Journal of Econometrics 164, 60-78.

Pettenuzzo, D., Timmermann, A. Rossen, V., 2014. Forecasting stock returns under economic constraints. Journal of Financial Economics 114, 517-553.

Qu, Z., Perron, P., 2013. A stochastic volatility model with random level shifts and its application to S&P 500 and NASDAQ return indices. Econometrics Journal 16, 309-339.

Rapach, D.E., Wohar, M.E., 2006. Structural breaks and predictive regression models of aggregate U.S. stock returns. Journal of Financial Econometrics 4, 238-274.

Ray, B.K., Tsay, R.S., 2002. Bayesian methods for change-point detection in long-range dependent processes. Journal of Time Series Analysis 23, 867-705.

Rosenberg, B., 1973. Random coefficient models: the analysis of a cross-section of time series by stochastically convergent parameter regression. Annals of Economic and Social Measurement 2, 399-428.

Smith, A., 2012. Markov breaks in regression models. Journal of Time Series Econometrics 4(1), Article 3: 1-33.

Stock, J.H., Watson, M.W., 1996. Evidence on structural instability in macroeconomic time series relations. Journal of Business and Economic Statistics 14, 11-30.

Stock, J.H., Watson, M.W., 2007. Why has U.S. inflation become harder to forecast? Journal of Money, Credit and Banking 39, 3-33.

Varneskov, R.T., Perron P., 2016. Combining long memory and level shifts in modeling and forecasting the volatility of asset returns. Unpublished Manuscript, Department of Economics, Boston University.

Wei, G., Tanner, M.A., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association 85, 699–704.

Welch, I., Goyal A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455-1508.

Wu, C.F.J., 1983. On the convergence properties of the EM algorithm. The Annals of Statistics 11, 95-103.

Xia, Y., 2001. Learning about predictability: the effects of parameter uncertainty on dynamic asset allocation. Journal of Finance 56, 205-246.

Xu, J., 2017. Forecasting macroeconomic variables using common factors with parameter instability. Manuscript in Preparation, Shanghai University of Finance and Economics.

Xu, J., Perron, P., 2014 Modeling and forecasting stock return volatility: level shift model with time varying jump probability and mean reversion. International Journal of Forecasting 30, 449-463.

Xu, L., Jordan, M.I., 1996. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Computation 8, 129-151.

Appendix

A.1 Mixture Kalman filtering algorithm

A.1.1 The basic model: equations (1)-(2)

At $t = 0$, for $i = 1, ..., M$, given initial parameters $\Theta_0 = (r_0, r_1, \sigma_e, \sigma_\eta, \rho)$, draw $K_0^{(i)} \sim Ber(p_0)$, $\beta_0^{(i)}$, $P_0^{(i)} \sim f(\beta_0; \Theta_0)$ and set $w_0^{(i)} = 1/M$. For $t = 1, ..., T$ and $i = 1, ..., M$:

1. Construct the Kalman predictions:

$$\begin{aligned} \beta_{t|t-1}^{(i)} &= \beta_{t-1|t-1}^{(i)} + K_{t-1}^{(i)} \mu_{\eta,t|t-1}^{(i)} \\ P_{t|t-1}^{(i)} &= P_{t-1|t-1}^{(i)} + K_{t-1}^{(i)} \sigma_\eta^2 \end{aligned}$$

   where $\mu_{\eta,t|t-1}^{(i)} = \rho(\beta_{t-1|t-1}^{(i)} - \bar{\beta}^{t-1,(i)})$.

2. Compute the forecast of y_t: $E[f(\beta_{t|t-1})] = \sum_{i=1}^{M} w_{t-1}^{(i)} f(\beta_{t|t-1}^{(i)})$.

3. Compute the importance weights $w_t^{(i)} \sim N(v_t^{(i)}, F_t^{(i)})$, where

$$\begin{aligned} v_t^{(i)} &= y_t - X_t \beta_{t|t-1}^{(i)} \\ F_t^{(i)} &= X_t P_{t-1|t-1}^{(i)} X_t' + \sigma_u^2 \end{aligned}$$

   and set the normalized importance weights as $\hat{w}_t^{(i)} = w_t^{(i)} / \sum_{j=1}^{M} w_t^{(j)}$.

4. Resample M samplers $\{\beta_{t-1|t-1}^{(i)}, P_{t-1|t-1}^{(i)}, K_{t-1}^{(i)}\}_{i=1}^{M}$ with probabilities $\{\hat{w}_t^{(i)}\}_{i=1}^{M}$ and for $i = 1, ..., M$ set $w_t^{(i)} = 1/M$.

5. Draw $K_t^{(i)} \sim Ber(p_t)$ and construct the following steps of the Kalman filter:

$$\begin{aligned} \beta_{t|t-1}^{(i)} &= \beta_{t-1|t-1}^{(i)} + K_t^{(i)} \mu_{\eta,t|t-1}^{(i)} \\ P_{t|t-1}^{(i)} &= P_{t-1|t-1}^{(i)} + K_t^{(i)} \sigma_\eta^2 \\ v_t^{(i)} &= y_t - X_t \beta_{t|t-1}^{(i)} \\ F_t^{(i)} &= X_t P_{t-1|t-1}^{(i)} X_t' + \sigma_u^2 \\ \beta_{t|t}^{(i)} &= \beta_{t|t-1}^{(i)} + P_{t|t-1}^{(i)} X_t' F_t^{-1,(i)} v_t^{(i)} \\ P_{t|t}^{(i)} &= P_{t|t-1}^{(i)} - P_{t|t-1}^{(i)} X_t' F_t^{-1,(i)} P_{t|t-1}^{(i)} \end{aligned}$$

6. Compute the filtered estimate: $E[f(\beta_t)] = \sum_{i=1}^{M} \hat{w}_{t-1}^{(i)} f(\beta_{t|t}^{(i)})$.

A.1.2 The stochastic volatility model: equations (1)-(3), $K_t^\sigma = 1$

At $t = 0$, for $i = 1, ..., M$, given initial parameters $\Theta_0 = (r_0, r_1, \phi, \sigma_v, \sigma_\eta, \rho)$, let $ln\sigma_{\epsilon,t}^2 = z_t$, draw $K_0^{(i)} \sim Ber(p_0)$, $z_0^{(i)} \sim f(z_0; \Theta_0)$, $\beta_0^{(i)}$, $P_0^{(i)} \sim f(\beta_0; \Theta_0)$ and set $w_0^{(i)} = 1/M$. The procedure is then the same as described above, escept that in step 3, we have

$$F_t^{(i)} = X_t P_{t-1|t-1}^{(i)} X_t' + \exp(z_{t-1}^{(i)})$$

and in step 5, one also draws $z_t^{(i)} \sim N(\phi z_{t-1}^{(i)}, \sigma_v^2)$ and use

$$F_t^{(i)} = X_t P_{t-1|t-1}^{(i)} X_t' + \exp(z_t^{(i)}).$$

A.2 Particle smoothing algorithm: basic model

Consider the weighted samplers obtained from the filtering algorithm $\{w_t^{(i)}, \beta_t^{(i)}, K_t^{\beta(i)}\}_{i=1}^M$ for $i = 1, \ldots, M$, and $t = 1, \ldots, T$. Let $\{s_{\beta,t}^{(j)}, s_{K_1,t}^{(j)}\}_{j=1}^M$ be a set of particle smoothers. First set $s_{\beta,T}^{(j)} = \beta_T^{(i)}$ and $s_{K_1,T}^{(j)} = K_T^{\beta(i)}$ with probability $(1/M)$. Then, for $t = T-1, T-2, \ldots, 1$, compute

$$w_{t|t+1}^{(i)} \propto w_t^{(i)} p(s_{\beta,t+1}^{(j)} | \beta_t^{(i)}) \propto \{p_{t+1} exp(-\frac{(s_{\beta,t+1}^{(j)} - \beta_t^{(i)} - \mu_\eta)^2}{2\sigma_\eta^2})\}^{s_{K_1,t+1}^{(j)}} \{1 - p_{t+1}\}^{1 - s_{K_1,t+1}^{(j)}}$$

for $i = 1, \ldots, M$, and let $s_{\beta,t}^{(j)} = \beta_t^{(i)}$ and $s_{K_1,t+1}^{(j)} = K_t^{\beta(i)}$ with probability $w_{t|t+1}^{(i)}$. Repeat the steps above decreasing from $t-1$ until 1 to obtain $\{s_{\beta,t}^{(j)}, s_{K_t^\beta,t+1}^{(j)}\}$ as approximations to $p(\beta_t, K_t^\beta | y^{(T)})$, for $j = 1, \ldots, M$.

A.3 The stochastic volatility model: equations (1)-(3), K_t^σ unrestricted.

1. For $i = 1, \ldots, M$, given initial parameters $\Theta^0 = (p_1, p_2, \phi, \sigma_\eta.\sigma_v, \rho)$, generate $K_0^{\beta,(i)} \sim Ber(p_1)$, then $\beta_0^{(i)} \sim K_0^{\beta(i)} N(0, \sigma_\eta^2)$. Also generate $K_0^{\sigma,(i)} \sim Ber(p_2)$, then with $\ln \sigma_{\epsilon,t}^2 \equiv z_t$, $z_0^{(i)} = K_0^{\sigma,(i)} N(0, \sigma_v^2)$. Set the initial weights to $w_0^{(i)} = (1/M)$.

2. For $t = 1, \ldots, T$, generate $K_t^{\beta,(i)} \sim Ber(p_1)$ and $\beta_t^{(i)} = \beta_{t-1}^{(i)} + K_t^{\beta,(i)} N(\mu_{\eta,t}, \sigma_\eta^2)$, $\mu_{\eta,t} = \rho(\beta_{t-1}^{(i)} - \bar{\beta}^{(i),(t-1)})$, where $\bar{\beta}^{(i),(t-1)}$ is the average of all the particle filters from $t = 1$ to time $t-1$. Also generate $K_t^{\sigma,(i)} \sim Ber(p_2)$ and $z_t = \phi z_{t-1} + K_t^{\sigma,(i)} N(0, \sigma_v^2)$.

3. Compute

$$w_t^{(i)} \propto p(y_t | x_t^{(i)}) w_{t-1}^{(i)} \propto \frac{1}{\sqrt{2\pi \exp(z_t)}} exp\left\{ -\frac{(y_t - X_t \beta_t^{(i)})^2}{2 \exp(z_t)} \right\},$$

for $i = 1, \ldots, M$, and set the normalized importance weights as $\hat{w}_t^{(i)} = w_t^{(i)} / \sum_{i=1}^M w_t^{(i)}$.

4. Resample $\{\beta_t^{(i)}, K_t^{\beta,(i)}, z_t^{(i)}, K_t^{\sigma,(i)}\}_{i=1}^M$ with probability $\hat{w}_t^{(i)}$, and set $w_t^{(i)} = (1/M)$.

5. Repeat steps 1-4 increasing from $t+1$ until T.

A.4 MCEM: basic model

For the E-step, the complete likelihood of $\{\beta_1, \ldots, \beta_T, K_1^\beta, \ldots, K_T^\beta, y_1, \ldots, y_T\}$ is

$$f(\beta, K_1, Y) = \prod_{t=1}^{T} f(\beta_t|\beta_{t-1}, K_t^\beta) \prod_{t=1}^{T} f(K_t^\beta) \prod_{t=1}^{T} f(y_t|\beta_t, K_t^\beta)$$

$$= \{\prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_\eta^2}} exp(-\frac{(\beta_t - \beta_{t-1} - \mu_{\eta t})^2}{2\sigma_\eta^2})\}^{K_t^\beta} \prod_{t=1}^{T} p_t^{K_t^\beta} (1-p_t)^{1-K_t^\beta} \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_e^2}} exp(-\frac{(y_t - X_t\beta_t)^2}{2\sigma_e^2})$$

The log-likelihood function is:

$$\begin{aligned} -2logf(\beta, K^\beta, Y) &= \sum_{t=1}^{T} K_t^\beta [log(\sigma_\eta^2) + \frac{(\beta_t - \beta_{t-1} - \mu_{\eta t})^2}{\sigma_\eta^2}] \\ &\quad -2\sum_{t=1}^{T} [K_t^\beta log(p_t) + (1 - K_t^\beta) log(1-p_t)] \\ &\quad + \sum_{t=1}^{T} [log(\sigma_e^2) + \frac{(y_t - X_t\beta_t)^2}{\sigma_e^2}] \end{aligned}$$

The expectation of the complete log-likelihood function with respect to the unknown state variables β, K^β given Y and the current parameter estimates $\Theta^{(k-1)}$ is the objective function to be maximized. For the Monte Carlo EM algorithm, we approximate the expectation by Monte Carlo sample average with random samples drawn from $p(\beta_t, K_t^\beta|y^T)$ obtained using the particle smoothing algorithm. Then,

$$\begin{aligned} Q(\Theta, \Theta^{(k-1)}) &= E[-2logf(\beta, K^\beta, Y)|Y, \Theta^{(k-1)}] \\ &= \frac{1}{M} \sum_{i=1}^{M} \{\sum_{t=1}^{T} K_t^{\beta(i)} [log(\sigma_\eta^2) + \frac{(\beta_t^{(i)} - \beta_{t-1}^{(i)} - \mu_{\eta t})^2}{\sigma_\eta^2}] \\ &\quad -2\sum_{t=1}^{T} [K_t^{\beta(i)} log(p_t) + (1 - K_t^{\beta(i)}) log(1-p_t)] \\ &\quad + \sum_{t=1}^{T} [log(\sigma_e^2) + \frac{(y_t - X_t\beta_t^{(i)})^2}{\sigma_e^2}]\} \end{aligned}$$

For the M-step, the objective function becomes the usual log-likelihood function of Θ. Hence, standard maximum likelihood estimates are obtained by solving the first order conditions.

Remark 5 *For the full model with stochastic volatility, the estimation methodology is the same. The difference is that instead of having two state variables, we now have three, namely* $\{\beta_t, K_t^\beta, ln\sigma_{\epsilon,t}^2\}$. *Similarly, if different parameters are allowed to vary independently, we simply add the additional latent variables* $(\beta_{jt}, K_{jt}^\beta)$.

A.5 Selection of the initial values and construction of the standard errors

In order to speed up the convergence of the estimation algorithm, we can use information from the data to provide better initial parameter values. Consider the simple model

$$\begin{aligned} y_t &= \beta_t + e_t \\ \beta_t &= \beta_{t-1} + K_t^\beta \eta_t \end{aligned}$$

where $\eta_t \sim N(0, \sigma_\eta^2)$, $e_t \sim N(0, \sigma_e^2)$ and $K_t^\beta \sim Ber(p)$. The initial parameter values are set to $p^{(0)}$, $\sigma_\eta^{2^{(0)}} = |var(y - y_{-2}) - var(y - y_{-1})|$ and $\sigma_e^{2^{(0)}} = (var(y - y_{-1}) - p^{(0)}\sigma_\eta^{2^{(0)}})/2$. We set $p^{(0)}$ according to prior judgment about the frequency of the jumps.

To construct the standard errors of the estimates, Louis (1982) provides a way of obtaining the information matrix when using the EM algorithm. It is given by

$$\begin{aligned} I &= \sum_{t=1}^{T} E[B(\chi_t, \hat{\Theta})|\chi_t] - \sum_{t=1}^{T} E[S(\chi_t, \hat{\Theta})S^T(\chi_t, \hat{\Theta})|\chi] \\ &\quad -2\sum_{t<k}^{T} E[S(\chi_t, \hat{\Theta})|\chi]E[S(\chi_k, \hat{\Theta})|\chi]' \end{aligned}$$

where $S(\chi_t, \hat{\Theta})$ and $B(\chi_t, \hat{\Theta})$ are the first and second order derivatives, respectively, and χ refers to the complete data set including observed data and unobserved state variables. Since simulations are used in the EM algorithm, this may cause discontinuities, in which case this method is unstable and cannot always provide a positive definite covariance matrix. Duan and Fulop (2011) proposed a stable estimator of the information matrix applicable to the EM algorithm. They estimate the variance using the smoothed individual scores. Define $a_t(\Theta) = E[\partial log f(x_t|\chi_{t-1}, \Theta)/\partial\Theta|Y, \Theta]$, then the estimate of the information matrix is

$$\hat{I} = \Omega_0 + \sum_{j=1}^{l} w(l)(\Omega_j + \Omega_j')$$

where $\Omega_j = \sum_{t-1}^{T-j} a_t(\hat{\Theta})a_{t+j}(\hat{\Theta})'$ and $w(j) = 1 - j/(l+1)$. This method is easy to compute and does not require evaluations of the second-order derivatives of the complete data log-likelihood.

Table 1: Forecasting Comparisons from Simulated Models

	h=1	h=4	h=8	h=12	h=18	h=24	h=36	h=48	h=60
Panel 1: RLS Basic									
RLS_m	**0.93**	**0.87**	**0.85**	**0.84**	**0.87**	**0.89**	**0.92**	**0.94**	**0.95**
RLS_SV	0.97	1.04	1.06	1.07	1.07	1.08	1.06	1.05	1.05
Average	6.66	13.92	15.67	15.50	14.47	12.92	10.04	8.07	6.71
Rolling	1.22	1.57	1.71	1.74	1.76	1.71	1.58	1.47	1.40
ARMA	0.95	0.95	0.99	1.05	1.16	1.25	1.42	1.54	1.65
ARIMA	0.93	**0.87**	0.86	0.86	0.91	0.94	1.01	1.06	1.11
MS	2.92	5.42	6.05	6.02	5.71	5.21	4.25	3.58	3.12
TVP	2.58	4.72	5.30	5.39	5.23	4.90	4.20	3.73	3.41
Panel 2: RLS mean reverting									
RLS_SV	1.03	**1.03**	**1.02**	**1.01**	**1.01**	**0.98**	**0.93**	**0.91**	**0.90**
Average	1.52	2.04	2.00	1.91	1.75	1.56	1.31	1.16	1.16
Rolling	1.45	2.05	2.21	2.14	1.97	1.75	1.41	1.24	1.12
ARMA	1.32	1.78	1.96	1.97	1.94	1.86	1.73	1.69	1.67
ARIMA	**1.02**	1.07	1.11	1.11	1.09	1.07	1.00	0.95	0.92
MS	1.02	1.06	1.12	1.17	1.25	1.32	1.45	1.62	1.82
TVP	1.43	2.04	2.28	2.44	2.59	2.68	2.93	3.21	3.57
Panel 3: RLS no mean reverting									
RLS_m	**0.98**	**0.93**	**0.87**	**0.88**	**0.91**	**0.95**	**1.00**	**1.01**	**1.01**
RLS_SV	1.52	2.33	2.60	2.82	2.67	2.49	2.16	1.94	1.77
Average	2.39	4.38	4.92	5.23	4.61	4.00	3.04	2.46	2.06
Rolling	1.61	2.54	2.86	3.11	2.93	2.72	2.31	2.04	1.84
ARMA	0.98	0.94	0.89	0.91	0.95	0.99	1.05	1.06	1.06
ARIMA	0.98	0.94	0.87	0.88	0.92	0.96	1.02	1.04	1.05
MS	2.40	4.40	4.93	5.23	4.59	3.97	3.00	2.42	2.02
TVP	3.18	6.35	7.29	7.77	6.88	6.01	4.63	3.79	3.21
Panel 4: RLS_SV									
RLS_m	**1.00**	**0.99**	**0.98**	**0.98**	**0.99**	**0.96**	**0.94**	**0.92**	**0.95**
Average	1.07	1.19	1.34	1.48	1.50	1.50	1.42	1.27	1.10
Rolling	1.01	1.04	1.09	1.14	1.14	1.15	1.24	1.31	1.35
ARMA	1.09	1.20	1.34	1.46	1.47	1.46	1.36	1.20	1.03
ARIMA	1.04	1.02	1.04	1.09	1.12	1.14	1.17	1.20	1.24
MS	1.07	1.18	1.33	1.45	1.47	1.46	1.38	1.23	1.06
TVP	1.75	3.95	6.70	8.67	9.80	10.09	9.76	8.49	7.64

continued

Panel 5: ARMA (AR=0.95,MA=-0.5)									
RLS_m	1.06	1.12	**1.08**	**1.07**	**1.05**	**1.04**	1.02	1.04	**1.08**
RLS_SV	1.09	1.15	1.11	1.09	1.08	1.06	**1.02**	**1.03**	1.09
Average	4.23	4.30	3.06	2.47	2.14	2.15	2.66	2.98	3.08
Rolling	4.88	5.01	3.52	2.76	2.22	2.02	2.05	1.94	1.71
ARIMA	1.03	1.10	1.18	1.28	1.44	1.62	2.11	2.64	3.00
MS	3.50	3.64	2.68	2.21	1.89	1.83	2.14	2.46	2.66
TVP	**1.01**	**1.05**	1.13	1.22	1.39	1.59	1.90	2.27	2.76
Panel 6: ARIMA (AR=0.1,MA=-0.5,d=1)									
RLS_m	**1.01**	1.02	1.03	1.04	1.05	1.07	1.10	1.12	1.14
RLS_SV	1.05	**0.99**	**0.94**	**0.95**	**0.93**	**0.92**	**0.90**	**0.89**	**0.88**
Average	48.84	47.96	44.16	41.30	34.31	31.08	26.62	22.56	18.59
Rolling	3.69	3.31	3.09	3.02	2.61	2.45	2.27	2.07	1.81
ARMA	1.01	1.04	1.08	1.11	1.18	1.23	1.29	1.28	1.23
MS	8.78	8.56	8.11	7.88	6.78	6.27	5.50	4.80	4.05
TVP	1.14	1.19	1.27	1.23	1.20	1.18	1.16	1.12	1.09
Panel 7: Time Varying Parameter Model									
RLS_m	1.00	1.01	1.01	1.01	**1.01**	**1.01**	**1.01**	**1.01**	**1.01**
RLS_SV	1.62	1.65	1.45	1.35	1.28	1.24	1.19	1.16	1.14
Average	47.60	47.49	33.11	25.85	19.49	15.47	10.96	8.65	7.17
Rolling	7.09	7.18	5.31	4.37	3.55	3.02	2.41	2.10	1.87
ARMA	1.03	1.10	1.19	1.27	1.39	1.49	1.64	1.78	1.87
ARIMA	**1.00**	**1.01**	**1.01**	**1.01**	1.01	1.02	1.02	1.02	1.02
MS	37.44	37.25	25.92	20.18	15.16	11.99	8.44	6.60	5.43
Panel 8: Markov Switching Model									
RLS_m	0.92	0.76	**0.66**	0.66	0.75	0.82	0.89	0.86	0.88
RLS_SV	0.93	0.80	0.72	0.68	**0.64**	**0.60**	**0.53**	**0.48**	**0.46**
Average	0.95	0.84	0.78	0.74	0.71	0.68	0.62	0.58	0.55
Rolling	0.95	0.86	0.82	0.82	0.79	0.77	0.71	0.64	0.63
ARMA	0.93	0.78	0.71	0.69	0.69	0.67	0.63	0.59	0.56
ARIMA	**0.91**	**0.75**	0.66	**0.66**	0.72	0.77	0.80	0.76	0.77
TVP	1.82	3.04	4.10	4.35	4.82	5.16	6.06	6.59	6.95

Note: This table reports the relative cumulative MSFEs with respect to the true model. In each case, we generate 100 true data paths and 1000 observations for each path. We use the first 800 observations for in-sample estimation and leave the rest of the data to evaluate out-of-sample forecasting accuracy. The forecasting horizon is 60. In panel A, the true model is the basic random level shift model with three parameters $(\mathrm{p}, \sigma_e, \sigma_\eta)$. In panel B, the true model is the RLS model with mean reversion with five parameters $(r_0, r_1, \sigma_e, \sigma_\eta, \rho)$. Detailed explanation for each parameter is introduced in the simulation setup. In panel C, the true model is the same as the one in panel B except that the mean reverting parameter ρ is set to be 0. In panel D, the true model is the RLS model with stochastic volatility. The benchmark model for each panel is the one with correct model specification. Numbers smaller than 1 indicate better forecasting performance than the corresponding benchmark model. The bold numbers in each panel stand for the smallest relative cumulative MSFEs.

Table 2.1: Equity Premium Forecasting Comparisons for the Period 1998-2012

(Monthly Data; Shiller Dataset)

	Cumulative MSFE						
	h=1	h=6	h=12	h=18	h=24	h=30	h=36
Historical average	10.27	356	1364	2955	5010	7390	9981
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Relative Cumulative MSFE						
Rolling 10 years	0.14	0.16	0.18	0.19	0.20	0.23	0.25
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dividend_no break	0.12	0.13	0.14	0.15	0.16	0.18	0.21
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dividend_rolling	0.09	0.10	0.11	0.11	0.12	0.12	0.13
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
TVP	0.01[a]	0.03	0.04	0.06	0.08	0.10	0.12
	(0.69)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Regime Switching	0.26	0.30	0.33	0.35	0.38	0.42	0.47
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Level Shift	0.01[a]	0.03	0.04	0.06	0.08	0.10	0.12
	(0.38)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dividend_K1t	0.01[a,]*	0.01	0.02	0.03	0.03	0.04	0.05
	(1.00)	(0.07)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Dividend_K2t	0.01[a]	0.01[a,]*	0.02[a,]*	0.02[a,]*	0.03[a,]*	0.03[a,]*	0.04[a,]*
	(0.38)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)

Note: This table reports the relative MSFEs with respect to the benchmark model, which is the historical average. Numbers in the parentheses are p-values of the model confidence set of Hansen et al. (2011). Numbers with superscript "a" indicate the models which belong to the 10% model confidence set using all comparisons. Numbers with an asterisk refer to the model with the smallest MSFE amongst all models. 'Rolling 10 years' refers to forecasting using historical averaged data with a window size fixed at 10 years. 'Dividend_no break' refers to the fixed parameter OLS regression of the equity premium on a constant and the lagged dividend-price ratio with full in-sample data. 'Dividend_rolling' is the same OLS regression using rolling 10 years in-sample data. 'TVP' stands for the time varying parameter model in which the unconditional mean of the equity premium is modelled as a random walk. 'Regime switching' is the two-state Markov regime switching model. 'Level shift' is the unconditional mean model with level shifts and mean reversion; 'Dividend_K_{1t}' is the conditional mean model with a constant term and the lagged dividend-price ratio as regressor and the constant term follows a level shift process with mean reversion; 'Dividend_K_{2t}' is the conditional mean model with a constant term and the lagged dividend-price ratio as regressor and the coefficient of the lagged dividend-price ratio follows a level shift process with mean reversion.

Table 2.2: Equity Premium Forecasting Comparisons for the Period 1988-1996

(Monthly Data; Shiller Dataset)

	Cumulative MSFE					
	h=1	h=12	h=24	h=36	h=48	h=60
Historical average	22.86	3228	12445	27030	47108	72598
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Relative Cumulative MSFE					
Rolling 10 years	0.076	0.081	0.090	0.102	0.110	0.115
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dividend_no break	0.303	0.272	0.235	0.205	0.171	0.143
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dividend_rolling	0.019	0.021	0.028	0.036	0.049	0.060
	(0.00)	(0.07)	(0.01)	(0.00)	(0.00)	(0.00)
TVP	0.002[a,*]	0.013[a]	0.018[a]	0.021[a]	0.027[a]	0.031[a]
	(1.00)	(0.72)	(0.37)	(0.17)	(0.19)	(0.25)
Regime Switching	0.626	0.622	0.615	0.608	0.605	0.602
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Level Shift	0.002	0.013[a,*]	0.016[a,*]	0.018[a,*]	0.022[a,*]	0.025[a,*]
	(0.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
Dividend_K1t	0.005	0.021	0.027	0.037	0.049	0.060
	(0.00)	(0.07)	(0.01)	(0.00)	(0.00)	(0.00)
Dividend_K2t	0.004	0.015[a]	0.018[a]	0.022[a]	0.027[a]	0.029[a]
	(0.00)	(0.22)	(0.48)	(0.28)	(0.37)	(0.44)

See Notes to Table 2.1

Table 3.1: Equity Premium Forecasting Comparisons for the Period 1975-2009

(Post Oil Shock; Annual Data; Welch & Goyal Dataset)

	Cumulative MSFE				
	h=1	h=2	h=3	h=4	h=5
Historical average	29.37	109.43	224.08	363.85	522.17[a]
	(0.01)	(0.05)	(0.01)	(0.01)	(0.19)
	Relative Cumulative MSFE				
Rolling 10 years	0.67	0.76	0.87	0.99	1.11
	(0.02)	(0.05)	(0.06)	(0.01)	(0.00)
BM_no break	0.92	0.94	0.99	1.07	1.16
	(0.01)	(0.05)	(0.02)	(0.01)	(0.00)
BM_rolling	0.38	0.39[a]	0.43[a]	0.50[a]	0.57[a]
	(0.08)	(0.35)	(0.27)	(0.47)	(0.33)
TVP	0.19[a,]*	0.27[a]	0.36[a]	0.46[a]	0.58[a]
	(1.00)	(0.85)	(0.81)	(0.47)	(0.29)
Regime Switching	0.78	0.77	0.77[a]	0.77[a]	0.77[a]
	(0.01)	(0.05)	(0.12)	(0.28)	(0.28)
Level Shift	0.19[a]	0.27[a,]*	0.35[a]	0.45[a]	0.56[a]
	(0.52)	(1.00)	(0.81)	(0.47)	(0.37)
BM_K1t	0.27[a]	0.29[a]	0.32[a,]*	0.37[a,]*	0.42[a,]*
	(0.14)	(0.85)	(1.00)	(1.00)	(1.00)
BM_K2t	0.35	0.35[a]	0.37[a]	0.41[a]	0.46[a]
	(0.08)	(0.35)	(0.48)	(0.47)	(0.37)

Note: The models are the same as in Tables 2.1-2.2, except that ‘BM’ replaces “dividend’. We use annual data from 1947-2014. Data before 1975 are used for in-sample estimation and the forecasting horizon is 5 years ahead. The Book-to-Market Ratio (BM) is the ratio of book value to market value for the Dow Jones Industrial Average. The covariate used to forecast level shift probability is the absolute changes in the earning-price ratio.

Table 3.2: Equity Premium Forecasting Comparisons for the Period 1975-2009

(Post Oil Shock; Monthly Data; Welch & Goyal Dataset)

	Cumulative MSFE					
	h=1	h=12	h=24	h=36	h=48	h=60
Historical average	29.19	4112	15895	34205	56932	82170
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Relative Cumulative MSFE					
Rolling 10 years	0.526	0.597	0.674	0.751	0.837	0.934
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BM_nobreak	0.857	0.904	0.945	0.985	1.030	1.084
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BM_rolling	0.201	0.226	0.236	0.243	0.252	0.264[a]
	(0.00)	(0.00)	(0.00)	(0.00)	(0.09)	(0.78)
TVP	0.014[a,*]	0.065[a]	0.132	0.196	0.270	0.353
	(1.00)	(0.16)	(0.00)	(0.00)	(0.00)	(0.00)
Regime Switching	0.699	0.691	0.680	0.669	0.659	0.651
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Level Shift	0.014[a]	0.065[a]	0.132	0.196	0.270	0.353
	(0.94)	(0.16)	(0.00)	(0.00)	(0.00)	(0.00)
BM_K1t	0.014[a]	0.054[a,*]	0.106[a,*]	0.155[a,*]	0.207[a,*]	0.257[a,*]
	(0.94)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
BM_K2t	0.015[a]	0.067[a]	0.127[a]	0.192[a]	0.262	0.324
	(0.92)	(0.16)	(0.20)	(0.10)	(0.09)	(0.07)

Note: The models are the same as in Tables 2.1-2.2, except that 'BM' replaces "dividend'. We use monthly data from 1921/03/31-2014/12/31. Data before 1975 are used for in-sample estimation and the forecasting horizon is 60 months ahead.

Table 4: Equity Premium Forecasting Comparisons for the Period 2009-2015

(Monthly Data; Welch & Goyal Dataset)

	Cumulative MSFE						
	h=1	h=4	h=8	h=12	h=16	h=20	h=24
Historical average	54.45	800	2859	5794	9244	12843	16176
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Relative Cumulative MSFE						
Rolling 10 years	0.67	0.67	0.68	0.69	0.70	0.72	0.74
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dividend_no break	0.91	0.93	0.96	0.97	0.98	0.99	0.99
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dividend_rolling	0.25	0.28	0.32	0.35	0.37	0.40	0.42
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
TVP	0.01[a]	(0.02[a]	0.04[a]	0.07[a]	0.10[a]	0.16[a]	0.24
	(0.38)	(0.51)	(0.65)	(0.96)	(0.66)	(0.17)	(0.00)
Regime Switching	0.64	0.63	0.61	0.60	0.58	0.57	0.55
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Level Shift	0.02	0.03	0.06	0.10	0.16	0.24	0.36
	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
DP_LS	0.01[a,]*	0.02[a,]*	0.04[a,]*	0.06[a,]*	0.10[a]	0.16[a]	0.24
	(1.00)	(1.00)	(1.00)	(1.00)	(0.66)	(0.17)	(0.01)
DP_VIX_SPX_LS	0.03	0.05	0.06[a]	0.07[a]	0.08[a,]*	0.10[a,]*	0.12[a,]*
	(0.02)	(0.07)	(0.44)	(0.96)	(1.00)	(1.00)	(1.00)

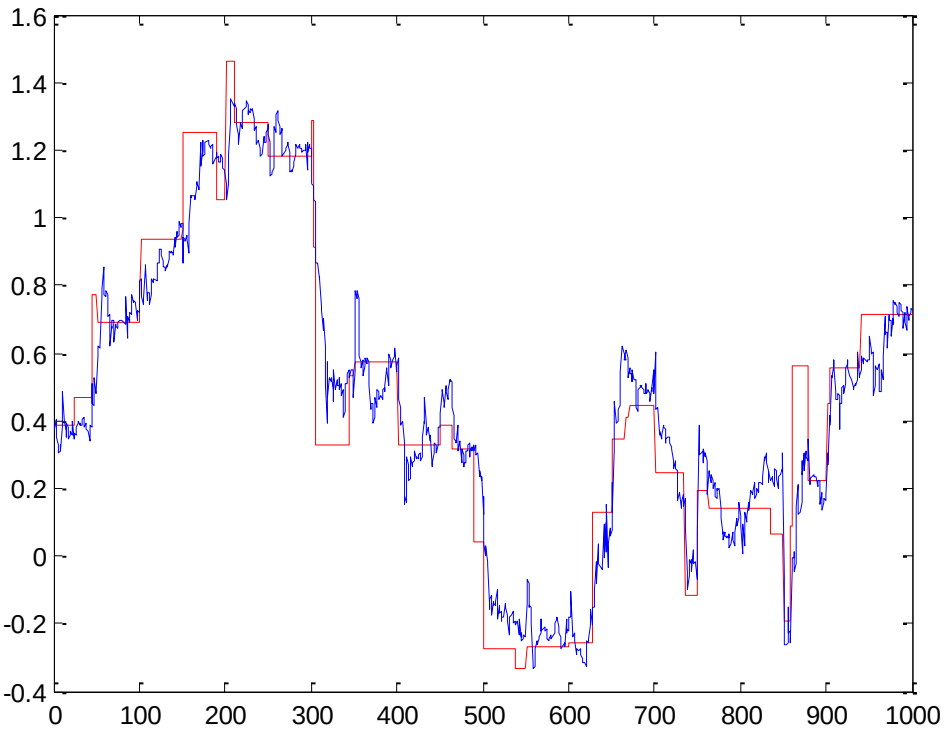
Note: We use monthly data from 1990/01/31-2015/07/31. Data before 2009 are used for in-sample estimation and the forecasting horizon is 24 months ahead. 'Level Shift' is the unconditional mean model with level shifts and mean reversion; 'DP_LS' is the conditional mean model with a constant term and the lagged dividend-price ratio as regressors with the constant term following a level shift process with mean reversion; 'DP_VIX_SPX_LS' is the conditional mean model with a constant term, the lagged dividend-price ratio, the VIX index and the returns on the monthly S&P 500 index option as regressors with the constant term following a level shift process with mean reversion.

Table 5: Treasury Bill Rate Forecasting Comparisons

	h=12		h=24		h=36		h=48		h=60	
MSFE 1968-2002										
Recursive OLS	2.07	(0.06)	2.67[a]	(0.13)	3.10[a]	(0.11)	3.32[a]	(0.12)	3.44	(0.08)
Relative MSFE										
Rolling 5 years	1.12	(0.06)	1.38	(0.00)	2.22	(0.00)	4.06	(0.00)	8.35	(0.00)
Rolling 10 years	1.04	(0.06)	1.03	(0.00)	1.06	(0.00)	1.10	(0.00)	1.23	(0.01)
TVP	1.72	(0.00)	3.44	(0.00)	5.06	(0.00)	6.79	(0.00)	8.68	(0.00)
AR_K2t	1.02	(0.06)	1.01[a]	(0.13)	1.00[a]	(0.11)	1.00[a]	(0.12)	0.99[a]	(0.10)
AR_K1t,K2t	0.97[a,]*	(1.00)	0.97[a,]*	(1.00)	0.95[a,]*	(1.00)	0.94[a,]*	(1.00)	0.92[a,]*	(1.00)
MSFE 1970s										
Recursive OLS	2.05	(0.01)	2.99	(0.00)	3.11	(0.00)	2.77	(0.00)	2.43	(0.00)
Relative MSFE										
Rolling 5 years	0.93[a,]*	(1.00)	0.78[a,]*	(1.00)	0.76 [a,]*	(1.00)	0.78[a,]*	(1.00)	0.85[a]	(0.65)
Rolling 10 years	0.97	(0.35)	0.84	(0.00)	0.80	(0.00)	0.80	(0.04)	0.93	(0.02)
TVP	0.98	(0.01)	1.21	(0.00)	1.58	(0.00)	1.89	(0.00)	1.58	(0.00)
AR_K2t	1.03	(0.00)	1.01	(0.00)	1.00	(0.00)	1.00	(0.00)	1.02	(0.00)
AR_K1t,K2t	0.97[a]	(0.35)	0.98	(0.00)	0.96	(0.00)	0.89	(0.00)	0.82[a,]*	(1.00)
MSFE 1980s										
Recursive OLS	2.76[a]	(0.55)	3.32[a]	(0.66)	4.17[a]	(0.13)	4.84[a]	(0.17)	5.23[a]	(0.20)
Relative MSFE										
Rolling 5 years	1.20[a]	(0.22)	1.73[a]	(0.43)	2.76	(0.02)	4.71	(0.01)	9.20	(0.00)
Rolling 10 years	1.02[a]	(0.53)	1.06[a]	(0.66)	1.07	(0.06)	1.05	(0.06)	1.15	(0.09)
TVP	1.75	(0.00)	3.68	(0.00)	4.80	(0.00)	5.56	(0.00)	6.21	(0.00)
AR_K2t	1.01[a]	(0.55)	1.00[a]	(0.66)	1.00[a]	(0.26)	0.99[a]	(0.63)	0.99[a]	(0.64)
AR_K1t,K2t	0.98[a,]*	(1.00)	0.98[a,]*	(1.00)	0.95[a,]*	(1.00)	0.95[a,]*	(1.00)	0.93[a,]*	(1.00)
MSFE 1990s										
Recursive OLS	1.23	(0.04)	1.59	(0.00)	1.74	(0.00)	1.63	(0.04)	1.39[a]	(0.52)
Relative MSFE										
Rolling 5 years	1.17	(0.00)	1.23	(0.00)	1.17	(0.00)	1.28	(0.00)	1.63	(0.00)
Rolling 10 years	1.25	(0.00)	1.36	(0.00)	1.44	(0.00)	1.74	(0.00)	2.20	(0.00)
TVP	2.69	(0.00)	5.78	(0.00)	9.33	(0.00)	15.12	(0.00)	24.91	(0.00)
AR_K2t	1.02	(0.01)	1.01	(0.00)	1.01	(0.00)	1.01	(0.04)	0.99[a]	(0.95)
AR_K1t,K2t	0.94[a,]*	(1.00)	0.89[a,]*	(1.00)	0.92[a,]*	(1.00)	0.95[a,]*	(1.00)	0.99[a,]*	(1.00)

Note: This table reports the relative MSFEs with respect to the benchmark model, which is the recursive OLS. 'Recursive OLS' refers to the OLS model with an expanding estimation window; 'Rolling 5 years and 10 years' refer to OLS models with window lengths set at 5 years and 10 years; 'TVP' stands for the time varying parameter model; 'AR_LS' is the AR(1) model allowing for level shifts in the constant term; 'AR_LS_SV' incorporates stochastic volatility into the error term. 'AR_K1t,K2t' allows for both the constant term and the AR coefficient to follow a level shift process with two different latent variables and mean reversion.

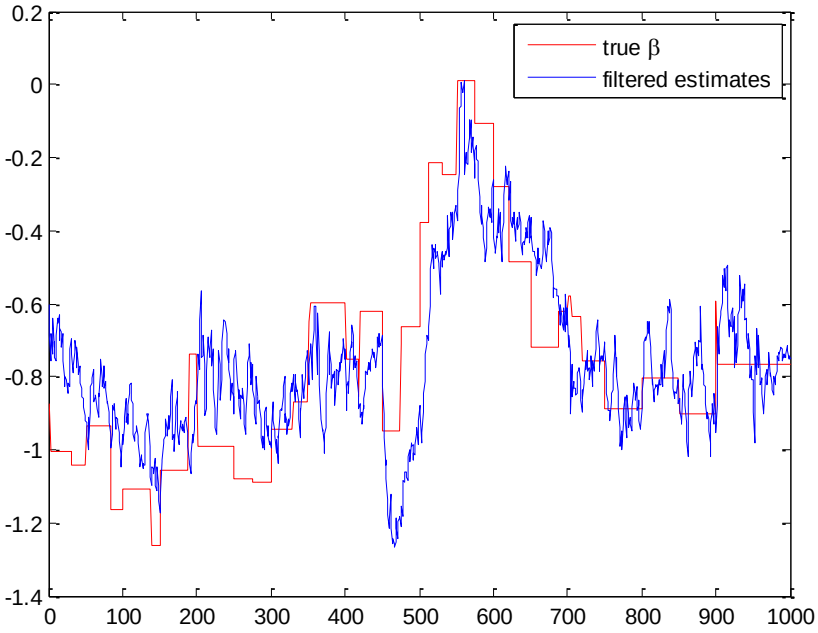
Figure 1: Mixture Kalman Filtered Estimates and True Parameter Process



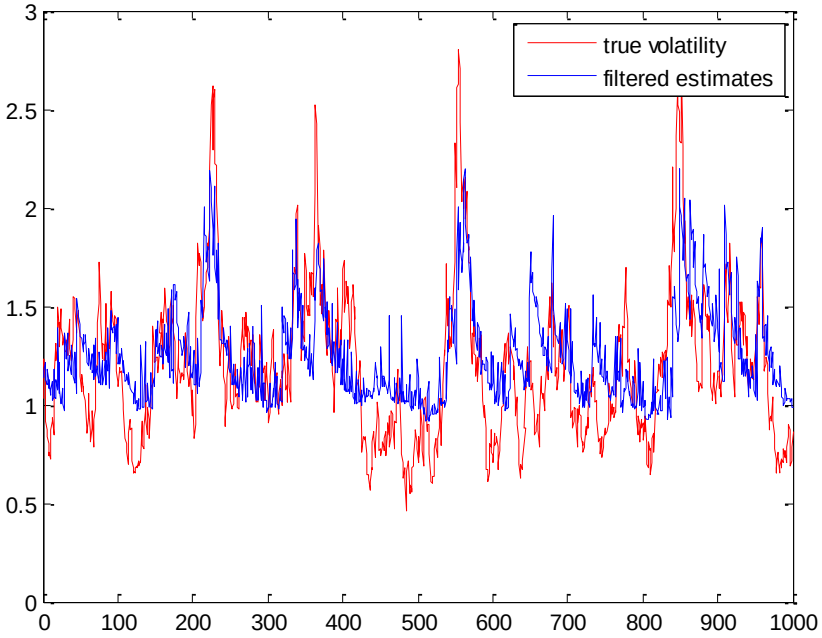
Note: The true process for β_t is generated using equations (1) and (2) with mean reversion and time varying probability with the parameters $(r_0, r_1, \sigma_e, \sigma_\eta, \rho) = (-1.96, 4, 0.2, 0.2, -0.1)$. The number of observations is 1000. The red solid line is the true β_t parameter process; the blue solid line is the corresponding filtered estimates of β_t using the mixture Kalman filter.

Figure 2: Mixture Kalman Filtered Estimates and True Parameter and Stochastic Volatility Processes

Panel A: True and Filtered Estimates of the Parameter Process



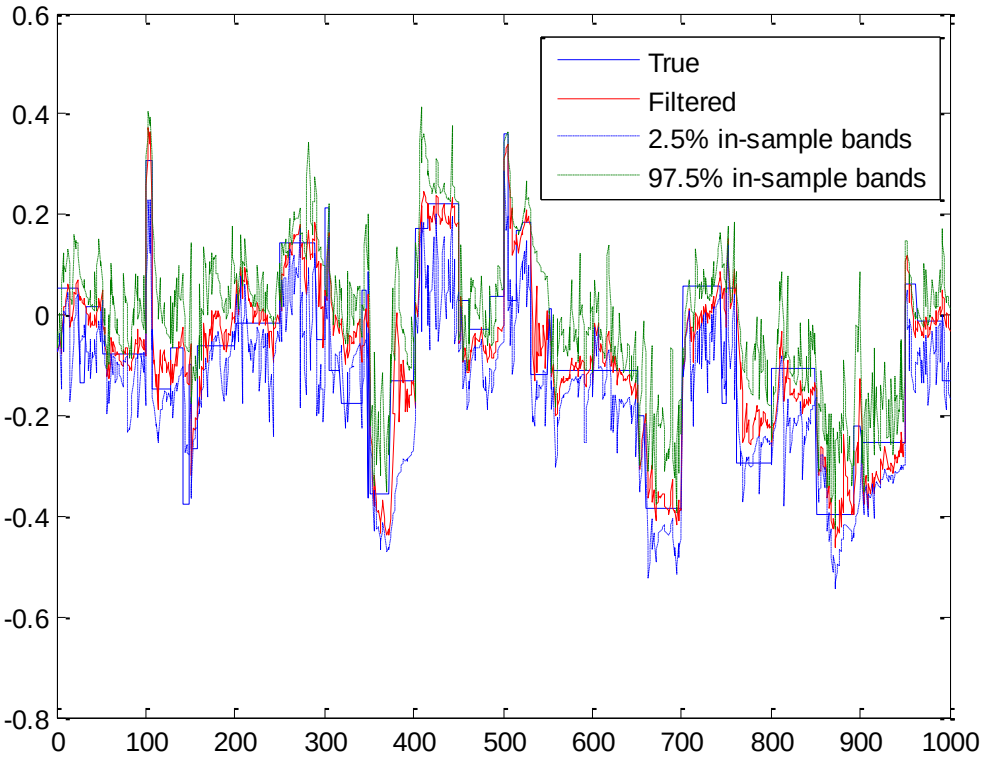
Panel B: True and Filtered Estimates of the Stochastic Volatility Process



Note: The true β_t and the stochastic volatility processes are generated using equations (4) and (5) with mean reversion and time varying probability with the parameters $(r_0, r_1, \phi, \sigma_v, \sigma_\eta, \rho) = (-1.96, 4, 0.95, 0.2, 0.2, -0.1)$. The number of observations is 1000.

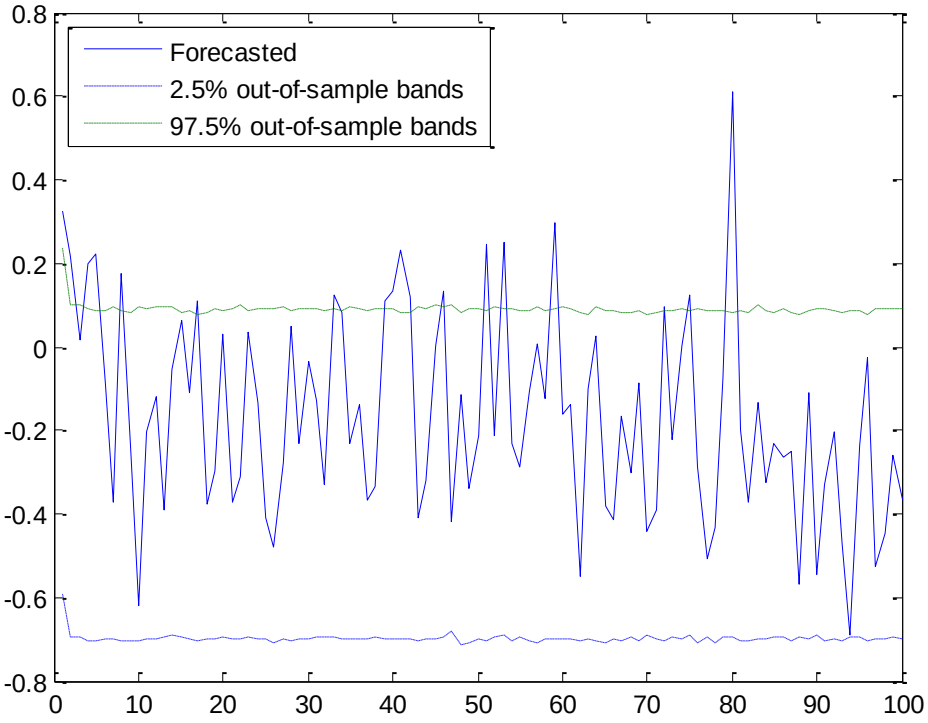
Figure 3: In-Sample Confidence Bands and Forecast Bands for the Parameter Process

Panel A: In-Sample Confidence Bands for the Parameter Process



Note: The true β_t process is generated using equations (1) and (2) with mean reversion and time varying probability with the parameters $(r_0, r_1, \sigma_e, \sigma_\eta, \rho) = (-1.96, 4, 0.2, 0.2, -1)$. The number of observations is 1000. The blue solid line is the true β_t parameter process; while the red solid line is the filtered estimates of the true β_t parameter process. The two dashed lines represent the 2.5% and 97.5% percentiles of the simulated parameter paths. The computation of the in-sample bands are based on M=1000 and S=1000 simulations.

Panel B: Out-of-Sample Forecast Bands for the Parameter Process



Note: The true β_t process is generated using equation (1) and (2) with mean reversion and time varying probability with the parameters $(r_0, r_1, \sigma_e, \sigma_\eta, \rho) = (-1.96, 4, 0.2, 0.2, -1)$. The number of observations is 500. We use the first 300 observations to obtain the parameter estimates. The out-of-sample forecasts start from the 301th observation. The forecasting horizon is set to be 100 steps. The blue solid line is the true data. The two dashed lines represent the 2.5% and 97.5% percentiles of the extrapolated paths. The computation of the out-of-sample bands are based on M=1000 and S=1000 simulations.