Markov Breaks in Regression Models

Aaron Smith*

Department of Agricultural and Resource Economics

University of California, Davis

Abstract

This article develops a new Markov breaks (MB) model for forecasting and making inference in regression models with stochastic breaks. The MB model permits an arbitrarily large number of abrupt breaks in the regression coefficients and error variance, but it maintains a low-dimensional state space, and therefore it is computationally straightforward. I compare the model to competing breaks models and show that it outperforms them in a Monte Carlo experiment. I employ the MB model to assess the efficacy of the conditional CAPM in pricing US stock returns. Both in and out of sample and for all portfolios under study, the MB model fits monthly stock returns significantly better than several alternative models. Using the estimated MB model to capture time-varying alphas and betas, I show that the momentum effect has persisted with a constant pricing error since 1927, the size effect has been prominent only in short bursts, and the book-to-market effect has been strong since 1980 after being less prominent in the preceding 30 years.

Keywords: structural breaks, Markov switching, forecasting, smoothing, cointegration.

* Department of Agricultural and Resource Economics, University of California, One Shields Ave, Davis, CA 95616, ph: 530-752-2138, fax: 530-752-5614, email: adsmith@ucdavis.edu. I am grateful to Robert Engle for comments and discussions that were crucial in the development of this manuscript.

1. Introduction

Econometricians frequently confront regressions with coefficients that change over time. These changes can occur abruptly, creating a tradeoff in choosing a sample for parameter estimation. A large estimation sample may contain breaks and therefore generate biased parameter estimates and forecasts, whereas a short sample may yield imprecise estimates and forecasts. Conditional on one or two deterministic breaks, Pesaran and Timmermann (2006) demonstrate that the best method for managing this tradeoff depends on the size of a break and the length of the pre- and post-break samples. In this paper, I treat the regression coefficients and the breaks as random variables, which enables the tradeoff to be managed probabilistically by a likelihood function.

I develop the Markov breaks (MB) model, which differs from a conventional Markov switching model (Hamilton 1989, Timmermann 2001) in two ways. First, Markov switching models treat the regression coefficients and error variance as fixed parameters within each regime, whereas I view them as latent random variables. Second, Markov switching models characterize each regime as a distinct state, whereas the MB model specifies a two-state Markov process to generate the breaks. The first state designates a break and the second state indicates no break. This structure of the MB model produces five immediate benefits, (i) it permits a large number of breaks, up to a maximum of a break every period, (ii) it keeps the dimension of the state space under control, and therefore it is computationally straightforward, (iii) it allows direct inference about the probability of a break in period t conditional on the data, (iv) it allows clustering of breaks through the Markov dependence in the state variable, and (v) rather than conditioning on a single break date, it generates forecasts and estimates using a probability weighted average over models that use progressively more data.

I apply the MB model to the capital asset pricing model (CAPM), which posits that the expected return on a financial asset is determined by the sensitivity (beta) of its returns to the return on the market portfolio. Previous research has shown that the CAPM misprices portfolios

based on past returns, firm size and the ratio of book-value to market-value of equity (Jegadeesh and Titman 1993; Fama and French 1992, 1993). However, the conditional CAPM, in which the betas change over time, can hold even when the unconditional model fails (Hansen and Richard 1987). I estimate a conditional CAPM using the MB(*k*) model and show that it outperforms Markov switching and rolling regression models in an out-of-sample prediction exercise. Moreover, I show that the momentum effect has persisted with a constant pricing error since 1927, and that the size and book-to-market (BM) effects have existed sporadically. These results reconcile the contrasting recent conclusions of Lewellen and Nagel (2006), who reject the conditional CAPM over the period 1964-2001, and Ang and Chen (2006), who find little evidence of the BM effect over the period 1926-2001.

The paper proceeds as follows. I develop the MB model and discuss its properties in Section 2, and I derive filtering and smoothing algorithms in Section 3. Section 4 discusses practical implementation of the model and Section 5 compares the model to Markov switching and STOPBREAK models. The CAPM application follows in Section 6, and Section 7 contains concluding remarks.

2. Markov Breaks Model

The MB model is

$$y_t = x_t'\beta_t + \varepsilon_t, \qquad t = 1, 2, \ldots, T \tag{1}$$

where $\varepsilon_t \,|\, (x_t, \beta_t) \sim N(0, \sigma^2)$, x_t is a $r\times 1$ vector of explanatory variables that may include lags of y_t, and $\beta_t \sim N\left(\beta_0, \sigma^2 V_0\right)$ is a random coefficient vector. To conserve degrees of freedom, I specify V_0 to be a diagonal matrix. The Markov random variable $s_t \in \{1,0\}$ defines the breaks, such that $s_t = 1$ denotes a break in period t and $s_t = 0$ denotes no break. I define a break as a new draw of β_t, which is independent of the observed data up to period t–1 and x_t, i.e., $(\beta_t \,|\, s_t = 1) \perp (x_t, \Im_{t-1})$ where $\Im_{t-1}$ denotes the information in $(y_1, y_2, ..., y_{t-1}, x_1, x_2, ..., x_{t-1})$.

Although the coefficient vector β_t is a random variable, the parameter vector (β_0, V_0, σ^2) is fixed, so I analyze the model in a classical likelihood framework. In the next section, I develop the likelihood function along with estimates of the coefficients and break probabilities conditional on past breaks. Following that, I extend the model to allow for breaks in σ^2, before discussing the model's properties.

2.1 Likelihood Function

The log likelihood function for the model in (1) is

$$L(\theta) = \sum_{t=1}^{T} \log\left(f(y_t \mid x_t, \Im_{t-1}) \right),$$

where *f* denotes the density of y_t conditional on x_t and $\Im_{t-1}$. To obtain *f*, I rewrite the model as

$$y_t = x_t' \beta_{t-j} + \varepsilon_t, \tag{2}$$

where *t–j* denotes the period of the most recent break. Then, defining the matrix $B_t \equiv \left[\beta_t \quad \beta_{t-1} \quad \cdots \quad \beta_1\right]$, the model becomes

$$y_t = x_t' B_t \xi_t + \varepsilon_t, \tag{3}$$

where ξ_t denotes a selection vector that has all elements equal to zero except for one element that equals one and indicates the location of the most recent break, i.e.,

$$\xi_t = \begin{bmatrix} s_t \\ (1-s_t)s_{t-1} \\ \vdots \\ (1-s_t)...(1-s_3)s_2 \\ (1-s_t)...(1-s_3)(1-s_2) \end{bmatrix}.$$

Using this nomenclature, the predictive density is

$$\begin{aligned} f(y_t \mid x_t, \Im_{t-1}) &= \sum_{i=1}^{t} f(y_t \mid \xi_{i,t} = 1, x_t, \Im_{t-1}) \Pr(\xi_{i,t} = 1 \mid \Im_{t-1}) \\ &\equiv f(y_t \mid \xi_t, x_t, \Im_{t-1})' \xi_{t|t-1}, \end{aligned}$$

where $\xi_{t|t-1} \equiv E(\xi_t \mid \Im_{t-1})$ and the term $f(y_t \mid \xi_t, x_t, \Im_{t-1})$ is a *t* dimensional vector denoting the conditional density of y_t, in which the i^{th} element corresponds to the event $\xi_{i,t} = 1$. The state

variable ξ_t has dimension t because it keeps track only of the most recent break date before period t. To account for the entire sequence of possible breaks since period 1 would require an unmanageable state space of dimension 2^t. Reducing the state space dimension to t makes analysis of the MB model computationally feasible. In Section 4 I present a truncation method to further reduce the state space dimension and increase computational efficiency.

To construct $f(y_t \mid \xi_t, x_t, \Im_{t-1})$, I begin with the first observation. For t=1, we have

$$y_1 \mid x_1 \sim N\left(x_1'\beta_0, \sigma^2(1+x_1'V_0x_1)\right).$$

For t=2, the state variable is

$$\xi_2 = \begin{bmatrix} s_2 \\ 1-s_2 \end{bmatrix}.$$

If a break occurs in period 2, then we draw a new value of β_t and

$$y_2 \mid x_2, \xi_{1,2} = 1, \Im_1 \sim N\left(x_2'\beta_0, \sigma^2(1+x_2'V_0x_2)\right).$$

However, if no break occurs in period t=2, then the second element of ξ_2 equals one and $\beta_2 = \beta_1$. In this case, we have[1]

$$y_2 \mid x_2, \xi_{2,2} = 1, \Im_1 \sim N\left(x_1'\hat{\beta}_{1|1}, \sigma^2(1+x_2'\hat{V}_{1|1}x_2)\right).$$

Calculating the moments $\hat{\beta}_{1|1}$ and $\hat{V}_{1|1}$ requires updating inference about β_1 using the information in y_1. The model is

$$\begin{bmatrix} y_1 \mid x_1 \\ \beta_1 \mid x_1 \end{bmatrix} \sim N\left(\begin{bmatrix} x_1'\beta_0 \\ \beta_0 \end{bmatrix}, \begin{bmatrix} \sigma^2(1+x_1'V_0x_1) & \sigma^2x_1'V_0 \\ \sigma^2V_0x_1 & \sigma^2V_0 \end{bmatrix}\right),$$

so $\hat{\beta}_{1|1}$ and $\hat{V}_{1|1}$ can be obtained using standard formulas for updating linear projections (Hamilton, 1994, equation 4.5.30). Specifically, $\hat{\beta}_{1|1} = (V_0^{-1} + x_1x_1')^{-1}(V_0^{-1}\beta_0 + x_1y_1)$ and $\hat{V}_{1|1} = (V_0^{-1} + x_1x_1')^{-1}$.

[1] Here, and in the remainder of the paper, I use a hat notation to indicate that the conditioning set includes knowledge of the most recent break date. For example $\hat{\beta}_{t-5|t}$ denotes an estimate of β_t=β_{t-5} conditional on data up to period t and knowledge that the most recent break occurred in period t–5.

Generalizing to period t and conditioning on the most recent break having occurred in period t–i, we have

$$y_t \mid x_t, \xi_{i+1,t} = 1, \Im_{t-1} \sim N\left(x_t'\hat{\beta}_{t-i|t-1}, \sigma^2(1 + x_t'\hat{V}_{t-i|t-1}x_t)\right), \qquad i = 0, 1, \ldots, t\text{–}1, \tag{4}$$

where $\hat{\beta}_{t-i|t-1} = \hat{V}_{t-1|t-1}\left(V_0^{-1}\beta_0 + \sum_{j=1}^{i} x_{t-j}y_{t-j}\right)$ and $\hat{V}_{t-i|t-1} = \left(V_0^{-1} + \sum_{j=1}^{i} x_{t-j}x_{t-j}'\right)^{-1}$. Because β_0 and V_0 can be viewed as prior moments of β_t, the formulas for $\hat{\beta}_{t-i|t-1}$ and $\hat{V}_{t-i|t-1}$ are identical to those for the Bayesian posterior mean and variance of regression coefficients in a model with fixed σ^2 and normally distributed data and priors (Koop 2003, pg. 37).

The break indicator s_t follows a first order Markov process, which implies that ξ_t is also a first order Markov process. This specification allows application of the standard Markov-switching filter to obtain $\xi_{t|t}$ (Hamilton 1989). The filter is

$$\xi_{t|t} = \frac{f(y_t \mid \xi_t, x_t, \Im_{t-1}) * \xi_{t|t-1}}{f(y_t \mid \xi_t, x_t, \Im_{t-1})'\xi_{t|t-1}} \tag{5}$$

where $\xi_{t|t-1} = P_t\xi_{t-1|t-1}$, * denotes element-by-element multiplication, and P_t denotes the $t \times (t-1)$ matrix of transition probabilities

$$P_t = \begin{bmatrix} p_{11} & p_{01} & p_{01} & \cdots & p_{01} \\ p_{10} & 0 & 0 & \cdots & 0 \\ 0 & p_{00} & 0 & \cdots & 0 \\ 0 & 0 & p_{00} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_{00} \end{bmatrix},$$

where $p_{lm} \equiv \Pr(s_t = m \mid s_{t-1} = l)$.

Therefore, the log likelihood function can be easily calculated using the standard Markov switching filter and standard formulas for updating linear projections. I maximize the likelihood with respect to the unknown parameters of the model (β_0, V_0, σ^2, p_{00}, p_{11}) and obtain standard errors through numerical differentiation of the log likelihood function. Next, I extend the model to allow for breaks in σ^2.

2.2 Allowing Breaks in σ^2

The model in (1) constrains the error variance to be constant. This constraint may not hold in many applications, so I specify the more general model

$$y_t = x_t'\beta_t + \varepsilon_t, \qquad t = 1, 2, \ldots, T$$

where $\varepsilon_t \mid (x_t, \beta_t, \sigma_t) \sim N(0, \sigma_t^2)$, $\beta_t \mid \sigma_t \sim N\left(\beta_0, \sigma_t^2 V_0\right)$, $\sigma_t^{-2} \sim G(\sigma_0^{-2}, \eta_0)$, G denotes a Gamma distribution, and $(\beta_t, \sigma_t \mid s_t = 1) \perp (x_t, \Im_{t-1})$. As for the case with constant error variance, the likelihood function is

$$L(\theta) = \sum_{t=1}^{T} \log\left(f(y_t \mid \xi_t, x_t, \Im_{t-1})'\xi_{t|t-1} \right), \tag{6}$$

where $f(y_t \mid \xi_t, x_t, \Im_{t-1})$ is a t dimensional vector denoting the conditional density of y_t, in which the i^{th} element corresponds to the event $\xi_{i,t} = 1$. The unknown parameters to be estimated are β_0, V_0, σ_0^2, η_0, p_{00}, and p_{11}.

Applying standard results that are most often used in Bayesian regression analysis with conjugate priors (Koop 2003, pg. 46), $f(y_t \mid \xi_t, x_t, \Im_{t-1})$ is a t-distribution with $\eta_0 + i$ degrees of freedom. Specifically,

$$y_t \mid x_t, \xi_{i+1,t} = 1, \Im_{t-1} \sim t\left(x_t'\hat{\beta}_{t-i|t-1}, \hat{\sigma}^2_{t-i|t-1}(1 + x_t'\hat{V}_{t-i|t-1}x_t), \eta_0 + i \right), \tag{7}$$

for all i = 0, 1, …, t–1, where $\hat{\beta}_{t-i|t-1} = \left(V_0^{-1} + \sum_{j=1}^{i} x_{t-j}x_{t-j}'\right)^{-1}\left(V_0^{-1}\beta_0 + \sum_{j=1}^{i} x_{t-j}y_{t-j}\right)$ and $\hat{V}_{t-i|t-1} = \left(V_0^{-1} + \sum_{j=1}^{i} x_{t-j}x_{t-j}'\right)^{-1}$ as in (4). As is the case in Bayesian regression analysis, the term $\hat{\sigma}^2_{t-i|t-1}$ is a weighted average of the prior variance and the sample variance with an additional term to account for the update in the estimate of β_{t-i}. Specifically,

$$\hat{\sigma}^2_{t-i|t-1} = \frac{\eta_0\sigma_0^2 + \sum_{j=1}^{i}\left(y_{t-j} - x_{t-j}'\hat{\beta}_{t-i|t-1}\right)^2 + \left(\hat{\beta}_{t-i|t-1} - \beta_0\right)' V_0^{-1}\left(\hat{\beta}_{t-i|t-1} - \beta_0\right)}{\eta_0 + i}, \tag{8}$$

(see Koop 2003, pg. 37).

2.3 Model Properties

I close this section with several remarks about the properties of the MB model.

Remark 1. The MB model accommodates any number of breaks, including the special case of a break every period. In this case, the model reduces to the random coefficients model of Hildreth and Houck (1968), and the data provide little information about any particular realization of β_t and σ_t^2. In general, the model produces fewer breaks than observations, and various dynamic patterns can arise from the Markov property of s_t. For example, if p_{11}=0, then breaks never occur in consecutive periods. If p_{11}=1– p_{00}, then the break arrival process is *iid* Bernoulli and therefore exhibits no dependence. If p_{11} is large, then breaks may be clustered, thereby allowing the model to spend several periods in transition from one stable regime to another.

Remark 2. The MB model nests the constant coefficient regression model. However, the standard likelihood ratio test of the null hypothesis of no breaks fails because the transition probability $p_{11} = \Pr(s_t = 1 | s_{t-1} = 1)$ is unidentified under the null hypothesis (Davies 1977). However, there exist a large number of tests in the econometrics literature that have power against changing coefficient models and are therefore applicable in this context, e.g., Andrews and Ploberger (1994), Bai and Perron (1998), and Elliott and Müller (2004). I suggest using these methods to test constant coefficient model against the MB model.

Remark 3. Vuong's (1989) likelihood ratio test for nonnested hypotheses can be used to compare the MB model to competing breaks models such as Markov switching. Alternatively, model selection criteria such as the Akaike information criterion (AIC, Akaike 1973) or the Bayesian information criterion (BIC) could be used. See Section 5.1 for a comparison between the MB model and Markov switching.

Remark 4. Conditional on the parameters, the likelihood function in (6) can be interpreted as a predictive likelihood function. Lauritzen (1974) and Hinkley (1979) developed predictive likelihood theory by using sufficient statistics to remove unknown parameters from the forecast

distribution. In (6)-(8), I use $\beta_{t-i|t-1}$, $\sigma^2_{t-i|t-1}$, and $\xi_{t|t-1}$ to remove the unknown β_{t-i}, σ^2_{t-i}, and ξ_t from the likelihood. Following White (1982), this predictive likelihood quantifies the Kullback-Leibler divergence between the model and the true data generating density conditional on the parameters (Cooley and Parke 1990).

Remark 5. The parameters of the MB model could be chosen *a priori* rather than estimated. For example, a regression model may be stable within an estimation sample, but a forecaster may suspect that a break occurred at the end of the estimation sample or in the forecast period (Andrews 2001). Clark and McCracken (2005) show how breaks can cause poor out-of-sample performance from a model that fits well in sample. Using the MB model and conditional on the chosen parameters, a forecaster would begin the recursive algorithm in (6)-(8) at the suspected end-of-sample break date and calculate forecasts and a predictive likelihood accordingly.

Remark 6. My treatment of the regression coefficients and error variance as random variables evokes Bayesian imagery. However, as in the random coefficients literature, I analyze the model in a classical likelihood framework. Maximum likelihood is convenient because the MB model permits calculation of the likelihood function using the recursive algorithm in (6)-(8), and likelihood maximization using numerical gradient-based methods. Nonetheless, as with any statistical model, the MB model is amenable to a fully Bayesian analysis. A Bayesian approach could treat β_0, V_0, σ^2_0, and η_0 as priors on the distributions of β_t and σ^2_t. Alternatively, it could treat β_0, V_0, σ^2_0, and η_0 as parameters each with their own prior, as in the Markov switching model of Pesaran, Pettenuzzo, and Timmermann (2006).

Remark 7. The user can constrain some coefficients in β_t to be constant over time by setting to zero the appropriate elements of V_0. Moreover, as long as p_{00} and p_{11} can be identified by time variation in one element of β_t or σ^2_t, the null hypothesis of a constant coefficient can be tested by applying a likelihood ratio (LR) or Wald test to the relevant elements of V_0. The null distribution

of these statistics is nonstandard because the element(s) of V_0 being tested are on the boundary of the parameter space. Following Self and Liang (1987) and Andrews (2001), the asymptotic null distribution of the LR and Wald statistics for testing H_0: $V_{10}=\ldots=V_{q0}=0$ mimics the distribution of $\sum_{i=1}^{q} z_i^2 I(z_i > 0)$, where $z_i \sim iidN(0,1)$ and $I()$ is an indicator function. For q=1, 2, 3, 4, and 5, the 5 percent critical values for this test are 2.71, 4.23, 5.44, 6.50, and 7.48, respectively. Similarly, to jointly test the null hypothesis that $x_1,\ldots x_q$ do not belong in the model (i.e., H_0: $\beta_{10}=\ldots=\beta_{q0}=0$, $V_{10}=\ldots=V_{q0}=0$), the asymptotic null distribution of the LR and Wald statistics mimics the distribution of $\sum_{i=1}^{q} z_i^2 I(z_i > 0) + \sum_{i=q+1}^{2q} z_i^2$, which implies critical values of 5.14, 8.02, 10.53, 12.87, and 15.09 for q=1, 2, 3, 4, and 5, respectively.

3. Filtered and Smoothed Inference About β and σ^2

3.1 Filtering

Conditional on $\Im_t$ and the most recent break having occurred in period $t–i$, β_{t-i} has the multivariate t distribution

$$\beta_{t-i} \mid \xi_{i+1,t} = 1, \Im_t \sim t\left(\hat{\beta}_{t-i|t}, \hat{\sigma}^2_{t-i|t}\hat{V}_{t-i|t}, \eta_0 + i + 1\right), \tag{9}$$

for all i = 0, 1, …, t–1 (Koop 2003, pg. 37). From (3) and (1), the regression coefficients can be written as $\beta_t = B_t \xi_t$, which implies

$$\beta_{t|t} \equiv E\left(\beta_t \mid \Im_t\right) = B_{t|t}\xi_{t|t}, \tag{10}$$

where $B_{t|t} \equiv \left[\hat{\beta}_{t|t} \quad \cdots \quad \hat{\beta}_{2|t} \quad \hat{\beta}_{1|t}\right]$. Thus, to estimate the coefficients conditional on data up to t, the model takes a weighted average of estimates that include progressively more past data. The conditional estimate that receives the largest weight is the one that uses all of the data back to the most likely date of the last break.

Similarly to (10), the filtered variance of β_t is

$$\operatorname{var}\left(\beta_t \mid \Im_t\right) = \sum_{i=0}^{t-1} \operatorname{var}\left(\beta_{t-i} \mid \xi_{i+1,t} = 1, \Im_t\right)\xi_{i+1,t|t}$$

where

$$\text{var}(\beta_{t-i} \mid \xi_{i+1,t} = 1, \Im_t) = \frac{\eta_0 + i + 1}{\eta_0 + i - 1} \hat{\sigma}^2_{t-i|t} \hat{V}_{t-i|t} .$$

The scale factor, $(\eta_0 + i + 1)/(\eta_0 + i - 1)$ arises because the *t*-distribution in (9) has $\eta_0 + i + 1$ degrees of freedom, and it implies that finite variance requires $\eta_0 + i > 1$.

To obtain the full distribution of β_t conditional only on the observed data, I integrate ξ_t out of the joint density $f(\beta_t, \xi_t \mid \Im_t)$ to obtain

$$f\left(\beta_t \mid \Im_t\right) = \sum_{i=0}^{t-1} g\left(\beta_t \mid \xi_{i+1,t} = 1, \Im_t\right) \xi_{i+1,t|t} ,$$

where $g\left(\beta_t \mid \xi_{it} = 1, \Im_t\right)$ denotes the density of the *t* distribution in (9). The distribution $f\left(\beta_t \mid \Im_t\right)$ is a mixture of multivariate *t* distributions and may be multi-modal if sufficient uncertainty exists about the location of the most recent break. On the other hand, if we can estimate accurately the location of the most recent break, i.e., if $\xi_{i+1,t|t} \approx 1$ for some *i*, then the distribution of β_t conditional on $\Im_t$ approximates a *t* distribution (or a normal distribution in the case of constant σ^2).

Conditional on the most recent break, σ^{-2}_{t-i} has the Gamma distribution

$$\sigma^{-2}_{t-i} \mid \xi_{i+1,t} = 1, \Im_t \sim G\left(\hat{\sigma}^{-2}_{t-i|t}, \eta_0 + i + 1\right), \tag{11}$$

for all *i* = 0, 1, …, *t*–1 (see Koop 2003, pg. 37). I obtain the distribution of σ_t^{-2} conditional only on the observed data by integrating ξ_t out of the joint density $f(\sigma_t^{-2}, \xi_t \mid \Im_t)$, which yields

$$f\left(\sigma_t^{-2} \mid \Im_t\right) = \sum_{i=0}^{t-1} h\left(\sigma_t^{-2} \mid \xi_{i+1,t} = 1, \Im_t\right) \xi_{i+1,t|t} , \tag{12}$$

where $h\left(\sigma_t^{-2} \mid \xi_{i+1,t} = 1, \Im_t\right)$ denotes the density of the Gamma distribution in (11). From the properties of the inverse-Gamma distribution, the first moment of the conditional variance is

$$E\left(\sigma^2_{t-i} \mid \xi_{i+1,t} = 1, \Im_t\right) = \frac{\eta_0 + i + 1}{\eta_0 + i - 1} \hat{\sigma}^2_{t-i|t} . \tag{13}$$

Combining (12) and (13) produces the filtered variance estimate

$$\sigma_{t|t}^2 = E\left(\sigma_t^2 \mid \Im_t\right) = \sum_{i=0}^{t-1} \frac{\eta_0 + i + 1}{\eta_0 + i - 1} \hat{\sigma}_{t-i|t}^2 \xi_{i+1,t|t} \equiv S_{t|t} \xi_{t|t}, \tag{14}$$

where $S_{t|t} \equiv \left[\frac{\eta_0+1}{\eta_0-1}\hat{\sigma}_{t|t}^2 \quad \cdots \quad \frac{\eta_0+t-1}{\eta_0+t-3}\hat{\sigma}_{2|t}^2 \quad \frac{\eta_0+t}{\eta_0+t-2}\hat{\sigma}_{1|t}^2 \right]$. The filtered variance only exists if η_0>1. This condition requires that the moments of the marginal distribution provide enough information so that one observation is sufficient to identify $\hat{\sigma}_t^2 = E(\sigma_t^2 \mid \xi_{1t} = 1, \Im_t)$.

The filtered estimates in (10) and (14) use information up to the current period *t*. The Markov property of the state variable ξ_t enables convenient forecasting of the coefficients and error variance. The forecasts are

$$\beta_{t+l|t} \equiv E\left(\beta_{t+l} \mid \Im_t\right) = B_{t+l|t}\left(\prod_{j=1}^{l} P_{t+j}\right)\xi_{t|t},$$

$$\sigma_{t+l|t}^2 \equiv E\left(\sigma_{t+l}^2 \mid \Im_t\right) = S_{t+l|t}\left(\prod_{j=1}^{l} P_{t+j}\right)\xi_{t|t},$$

where the conditional estimates are $S_{t+l|t} \equiv \left[\frac{\eta_0}{\eta_0-2}\sigma_0^2 \quad \cdots \quad \frac{\eta_0}{\eta_0-2}\sigma_0^2 \quad \frac{\eta_0+1}{\eta_0-1}\hat{\sigma}_{t|t}^2 \quad \cdots \quad \frac{\eta_0+t}{\eta_0+t-2}\hat{\sigma}_{1|t}^2 \right]$ and $B_{t+l|t} \equiv \left[\beta_0 \quad \cdots \quad \beta_0 \quad \hat{\beta}_{t|t} \quad \cdots \quad \hat{\beta}_{1|t} \right]$. Because post-break values of β_t and σ_t^2 are drawn from a stationary distribution, the long run forecasts are $\lim_{l\to\infty} \beta_{t+l|t} = \beta_0$ and $\lim_{i\to\infty} \sigma_{t+l|t}^2 = \frac{\eta_0}{\eta_0 - 2}\sigma_0^2$.

3.2 Smoothing

In this section, I present an algorithm that uses future observations to smooth the filtered estimates in (10) and (14). I estimate β_t and σ_t^2 by averaging across various estimates that condition on both the date of the last break before period *t* and the date of the next break after period *t*. Therefore, the smoothed coefficient estimate is

$$\beta_{t|T} = \sum_{m=t}^{T-1} \sum_{j=0}^{t} \hat{\beta}_{j|m} \pi_{jm|T} + \sum_{j=0}^{t} \hat{\beta}_{j|T} \vec{\pi}_{jT|T}, \tag{15}$$

where $\pi_{jm|T} \equiv \Pr\left(s_j = 1, n_j = m+1 \mid \Im_T\right)$, $\vec{\pi}_{jT|T} \equiv \Pr\left(s_j = 1, n_j \geq T+1 \mid \Im_T\right)$, and n_j denotes the period of the next break after period *j*, so that, for example, $n_j = j+3$ corresponds to the event $\{s_{j+1} = 0, s_{j+2} = 0, s_{j+3} = 1\}$. The probability term π_{jm} in (15) can be calculated directly from the

smoothed state probabilities $\xi_{t|T}$, which I obtain using the Markov-switching smoother

$$\xi_{t|T} = \xi_{t|t} * \{P'_{t+1}(\xi_{t+1|T} \div \xi_{t+1|t})\},$$

where ÷ denotes element-by-element division (Hamilton 1994).

The term $\pi_{jm|T}$ can be written as

$$\begin{aligned}\pi_{jm|T} &= \Pr\left(s_j = 1, s_{j+1} = 0, ..., s_m = 0, s_{m+1} = 1 \mid \Im_T\right) \\ &= \Pr\left(s_j = 1, s_{j+1} = 0, ..., s_m = 0 \mid \Im_T\right) - \Pr\left(s_j = 1, s_{j+1} = 0, ..., s_m = 0, s_{m+1} = 0 \mid \Im_T\right) \\ &= \xi_{m-j+1,m|T} - \xi_{m-j+2,m+1|T} \cdot \end{aligned} \tag{16}$$

Thus, the smoothed probability $\pi_{jm|T}$ equals the difference between the $(m–j+1)^{\text{th}}$ element of $\xi_{m|T}$ and the $(m–j+2)^{\text{th}}$ element of $\xi_{m+1|T}$. For the case of no breaks before the end of the sample,

$$\begin{aligned}\vec{\pi}_{jT|T} &= \Pr\left(s_j = 1, n_j \geq T+1 \mid \Im_T\right) \\ &= \Pr\left(s_j = 1, s_{j+1} = 0, ..., s_{T-1} = 0, s_T = 0 \mid \Im_T\right) \\ &= \xi_{T-j+1,T|T} \cdot \end{aligned} \tag{17}$$

In sum, the smoothed estimates are

$$\beta_{t|T} = \sum_{m=t}^{T-1} \sum_{j=0}^{t} \hat{\beta}_{j|m} \left(\xi_{m-j+1,m|T} - \xi_{m-j+2,m+1|T}\right) + \sum_{j=0}^{t} \hat{\beta}_{j|T} \xi_{T-j+1,T|T} \cdot$$

Similarly, for the error variance,

$$\sigma^2_{t|T} = \sum_{m=t}^{T-1} \sum_{j=0}^{t} \frac{\eta_0 + m - j + 1}{\eta_0 + m - j - 1} \hat{\sigma}^2_{j|m} \left(\xi_{m-j+1,m|T} - \xi_{m-j+2,m+1|T}\right) + \sum_{j=0}^{t} \frac{\eta_0 + T - j + 1}{\eta_0 + T - j - 1} \hat{\sigma}^2_{j|T} \xi_{T-j+1,T|T},$$

where $\hat{\sigma}^2_{j|m}$ is as defined in (8).

The distribution of β_t conditional on $\Im_T$ is non-Gaussian because of the possibility of breaks. However, conditional on knowledge of the breaks, the coefficient estimates are *t*-distributed which implies that

$$\begin{aligned} f\left(\beta_t \mid \Im_T\right) = \sum_{m=t}^{T-1} \sum_{j=0}^{t} & g\left(\beta_j \mid s_j = 1, n_j = m+1, \Im_T\right)\left(\xi_{m-j+1,m|T} - \xi_{m-j+2,m+1|T}\right) \\ & + \sum_{j=0}^{t} g\left(\beta_j \mid s_j = 1, n_j \geq T+1, \Im_T\right) \xi_{T-j+1,T|T} \end{aligned}$$

where $g\left(\beta_j \mid s_j = 1, n_t = m+1, \Im_T\right) = g\left(\beta_m \mid \xi_{m-j+1,m} = 1, \Im_m\right)$ denotes the density of the

distribution in (9). It follows that the distribution of β_t conditional on $\Im_T$ approximates multivariate *t* when we can estimate accurately the location of the last break before and the first break after period *t*. Conversely, when there is less certainty about the location of the nearest break, the distribution departs from *t*. Similarly, the distribution of σ_t^{-2} conditional on $\Im_T$ approximates Gamma when we can estimate accurately the location of the last break before and the first break after period *t*, and is given by

$$f\left(\sigma_t^{-2} \mid \Im_T\right) = \sum_{m=t}^{T-1} \sum_{j=0}^{t} h\left(\sigma_j^{-2} \mid s_j = 1, n_j = m+1, \Im_T\right)\left(\xi_{m-j+1,m|T} - \xi_{m-j+2,m+1|T}\right) + \sum_{j=0}^{t} h\left(\sigma_j^{-2} \mid s_j = 1, n_j \geq T+1, \Im_T\right)\xi_{T-j+1,T|T}$$

where $h\left(\sigma_j^{-2} \mid s_j = 1, n_t = m+1, \Im_T\right) = h\left(\sigma_m^{-2} \mid \xi_{m-j+1,m} = 1, \Im_m\right)$ denotes the density of the distribution in (11).

4. Practical Implementation: Truncating the State Space

The state variable ξ_t has dimension *t*, which markedly improves on the dimension 2^t required to keep track of the entire sequence of breaks up to period *t*, but is still computationally demanding for reasonable *t*. However, if the most recent break occurred long in the past, the exact date of the break carries little information for the period *t* likelihood. Therefore, I truncate the state space to have maximum dimension *k*+1, yielding the MB(*k*) model. In this section, I show how to approximate the likelihood function and the filtered coefficient estimates using the truncated state space. I show using Monte Carlo simulations that the approximation error from this truncation of the state space is negligible even when *k* is not too large.

For some *k*, define the state variable

$$\xi_t^k = \begin{bmatrix} s_t \\ (1-s_t)s_{t-1} \\ \vdots \\ (1-s_t)...(1-s_{t-k+2})s_{t-k+1} \\ (1-s_t)...(1-s_{t-k+2})(1-s_{t-k+1}) \end{bmatrix},$$

which is a Markov process with the $(k+1)\times(k+1)$ transition probability matrix

$$P = \begin{bmatrix} p_{11} & p_{01} & p_{01} & \cdots & p_{01} & p_{01} \\ p_{10} & 0 & 0 & \cdots & 0 & 0 \\ 0 & p_{00} & 0 & \cdots & 0 & 0 \\ 0 & 0 & p_{00} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{00} & p_{00} \end{bmatrix}.$$

To generate the filtered state probabilities, I use the standard Markov switching filter. For $t \leq k+1$, I can use the same filter as in (5) because the state-space truncation is not binding. For $t > k+1$, I use

$$\xi_{t|t}^{k} = \frac{f(y_t \mid \xi_t^k, x_t, \Im_{t-1}) * \xi_{t|t-1}^{k}}{\sum_{i=1}^{k+1} f(y_t \mid \xi_{it}^k = 1, x_t, \Im_{t-1}) \xi_{i,t|t-1}^{k}}$$

where $\xi_{t+1|t}^{k} = P\xi_{t|t}^{k}$. The term $f(y_t \mid \xi_t^k, x_t, \Im_{t-1})$ is a $k+1$ dimensional vector denoting the conditional density of y_t, where each element i corresponds to the event $\xi_{it} = 1$.

If a break occurs in the interval $[t-k+1, t]$, then one of the first k elements of ξ_t^k equals one and the predictive distribution $f(y_t \mid \xi_t^k, x_t, \Im_{t-1})$ is the same as in (7). On the other hand, if the most recent break occurred in period $t-k$ or before, then the last element of ξ_t^k equals one and the exact break date is unknown. In this case, forming the predictive distribution $f(y_t \mid \xi_{k+1,t}^k = 1, x_t, \Im_{t-1})$ requires an approximation to the distribution of $(\beta_{t-k}, \sigma_{t-k}^{-2})$ conditional on $\left(\Im_{t-1}, \xi_{k+1,t}^k = 1\right)$. I approximate this distribution using a simple forward recursion.

The event $\xi_{k+1,t}^k = 1$ indicates that no breaks occurred in the window from $t-k+1$ to t, but it does not designate when before period $t-k+1$ the most recent break occurred. Thus, the distribution of β_{t-k} conditional on $(\xi_{k+1,t}^k = 1, \sigma_{t-k}^2, \Im_{t-1})$ is a mixture of normals with weights depending on the expected date of the most recent break before $t-k+1$. Similarly, the distribution of σ_{t-k}^{-2} conditional on $(\xi_{k+1,t}^k = 1, \Im_{t-1})$ is a mixture of gammas. For large values of k, the last k observations in $\Im_{t-1}$ dominate, and there is little difference between the components of these

mixtures. It follows that, as k increases, the distribution of β_{t-k} conditional on $(\xi_{k+1,t}^{k}=1,\sigma_{t-k}^{2},\Im_{t-1})$ approaches normality and the distribution of σ_{t-k}^{-2} conditional on $(\xi_{k+1,t}^{k}=1,\Im_{t-1})$ approaches gamma. I use this large k approximation to generate the recursion for $\beta_{t-k},\sigma_{t-k}^{-2}\mid\xi_{k+1,t}^{k}=1,\Im_{t-1}$ and in turn to approximate the density $f(y_t\mid\xi_{k+1,t}^{k},x_t,\Im_{t-1})$.

Specifically, define the quantities $\bar{\beta}_{t-k|t-1}$, $\bar{V}_{t-k|t-1}$, $\bar{\sigma}_{t-k|t-1}^{2}$, and $\bar{\eta}_{t-k|t-1}$ such that[2]

$$\beta_{t-k}\mid\xi_{k+1,t}^{k}=1,\sigma_{t-k}^{2},\Im_{t-1}\simeq N\left(\bar{\beta}_{t-k|t-1},\sigma_{t-k}^{2}\bar{V}_{t-k|t-1}\right) \tag{18}$$

$$\sigma_{t-k}^{-2}\mid\xi_{k+1,t}^{k}=1,\Im_{t-1}\simeq G\left(\bar{\sigma}_{t-k|t-1}^{-2},\bar{\eta}_{t-k|t-1}\right), \tag{19}$$

where $\simeq$ denotes "approximately distributed as." Combined with the normality of ε_t, the approximations in (18) and (19) imply that

$$y_t\mid x_t,\xi_{k+1,t}^{k}=1,\Im_{t-1}\simeq t\left(x_t'\bar{\beta}_{t-k|t-1},\bar{\sigma}_{t-k|t-1}^{2}(1+x_t'\bar{V}_{t-k|t-1}x_t),\bar{\eta}_{t-k|t-1}\right).$$

To update $\bar{\beta}_{t-k|t-1}$ and $\bar{V}_{t-k|t-1}$ after observing y_t, I use standard formulae for updating linear projections (e.g., Hamilton 1994, pg 99),

$$\bar{\beta}_{t-k|t}=\bar{\beta}_{t-k|t-1}+\bar{V}_{t-k|t-1}x_t\left(1+x_t'\bar{V}_{t-k|t-1}x_t\right)^{-1}(y_t-x_t'\bar{\beta}_{t-k|t-1}) \tag{20}$$

and

$$\bar{V}_{t-k|t}=\bar{V}_{t-k|t-1}-\bar{V}_{t-k|t-1}x_t\left(1+x_t'\bar{V}_{t-k|t-1}x_t\right)^{-1}x_t'\bar{V}_{t-k|t-1}. \tag{21}$$

For the error variance, using (18) and (19) along with standard results that are most often used in Bayesian regression analysis with conjugate priors (Koop 2003, pg. 37), we have

$$\bar{\sigma}_{t-k|t}^{2}=\frac{\bar{\eta}_{t-k|t-1}\bar{\sigma}_{t-k|t-1}^{2}+\left(y_t-x_t'\bar{\beta}_{t-k|t}\right)^{2}+\left(\bar{\beta}_{t-k|t}-\bar{\beta}_{t-k|t-1}\right)'\bar{V}_{t-k|t-1}^{-1}\left(\bar{\beta}_{t-k|t}-\bar{\beta}_{t-k|t-1}\right)}{\bar{\eta}_{t-k|t-1}+1} \tag{22}$$

$$\bar{\eta}_{t-k|t}=\bar{\eta}_{t-k|t-1}+1.$$

[2] Here, and in the remainder of the paper, I use a bar notation to indicate that the conditioning set includes no breaks between the period of interest (period *t*-*k* in this case) and the end of the information set (period *t*-1 in this case).

Equations (20)-(22) show how to obtain $\overline{\beta}_{t-k|t}$, $\overline{V}_{t-k|t}$, $\overline{\sigma}^2_{t-k|t}$, and $\overline{\eta}_{t-k|t}$ from $\overline{\beta}_{t-k|t-1}$, $\overline{V}_{t-k|t-1}$, $\overline{\sigma}^2_{t-k|t-1}$, and $\overline{\eta}_{t-k|t-1}$. For period t–k+1, I obtain $\overline{\beta}_{t-k+1|t}$, $\overline{V}_{t-k+1|t}$, $\overline{\sigma}^2_{t-k+1|t}$, and $\overline{\eta}_{t-k+1|t}$ from $\overline{\beta}_{t-k|t}$, $\overline{V}_{t-k|t}$, $\overline{\sigma}^2_{t-k|t}$, and $\overline{\eta}_{t-k|t}$ by accounting for the possibility of a break in period t–k+1. Specifically, I calculate the parameters of the approximate distribution of $\beta_{t-k+1}, \sigma^{-2}_{t-k+1} \mid \xi^k_{k+1,t+1} = 1, \Im_{t-1}$ as a weighted average of the values conditional on $s_{t-k+1} = 1$ and the values conditional on $s_{t-k+1} = 0$. This approximation closely resembles that used in Kim (1994) for approximating the Kalman filter in a dynamic linear model with Markov switching (see also Harrison and Stevens 1976). For $\overline{\beta}_{t-k+1|t}$, we have

$$\begin{aligned}\overline{\beta}_{t-k+1|t} &= \overline{s}_{t-k+1|t} E(\beta_{t-k+1} \mid \xi_{kt} = 1, \sigma^2_{t-k+1}, \Im_t) + (1 - \overline{s}_{t-k+1|t})\overline{\beta}_{t-k|t} \\ &= \overline{s}_{t-k+1|t}\hat{\beta}_{t-k+1|t} + (1 - \overline{s}_{t-k+1|t})\overline{\beta}_{t-k|t},\end{aligned}$$

where

$$\begin{aligned}\overline{s}_{t-k+1|t} &\equiv \Pr(s_{t-k+1} = 1 \mid s_t = s_{t-1} = \ldots = s_{t-k+2} = 0, \Im_t) \\ &= \frac{\Pr(s_t = s_{t-1} = \ldots = s_{t-k+2} = 0, s_{t-k+1} = 1 \mid \Im_t)}{\Pr(s_t = s_{t-1} = \ldots = s_{t-k+2} = 0 \mid \Im_t)} \\ &= \frac{\xi^k_{k,t|t}}{\xi^k_{k,t|t} + \xi^k_{k+1,t|t}}.\end{aligned}$$

Similarly, for the other parameters, we have $\overline{V}_{t-k+1|t} = \overline{s}_{t-k+1|t}\hat{V}_{t-k+1|t} + (1 - \overline{s}_{t-k+1|t})\overline{V}_{t-k|t}$, $\overline{\sigma}^{-2}_{t-k+1|t} = \overline{s}_{t-k+1|t}\hat{\sigma}^{-2}_{t-k+1|t} + (1 - \overline{s}_{t-k+1|t})\overline{\sigma}^{-2}_{t-k|t}$, and $\overline{\eta}_{t-k+1|t} = \overline{s}_{t-k+1|t}(\eta_0 + k) + (1 - \overline{s}_{t-k+1|t})\overline{\eta}_{t-k|t}$.

Using the truncated state space, the filtered estimates of β_t and σ^2_t are given by $\beta_{t|t} = B^k_{t|t}\xi^k_{t|t}$ and $\sigma^2_{t|t} = S^k_{t|t}\xi^k_{t|t}$ respectively, where $S^k_{t|t} \equiv \left[\frac{\eta_0+1}{\eta_0-1}\hat{\sigma}^2_{t|t} \quad \cdots \quad \frac{\eta_0+k}{\eta_0+k-2}\hat{\sigma}^2_{t-k+1|t} \quad \frac{\eta_{t-k|t}}{\eta_{t-k|t}-2}\overline{\sigma}^2_{t-k|t}\right]$ and $B^k_{t|t} \equiv \left[\hat{\beta}_{t|t} \quad \cdots \quad \hat{\beta}_{t-k+1|t} \quad \overline{\beta}_{t-k|t}\right]$. In the Appendix, I present the algorithm for generating smoothed estimates of β_t and σ^2_t using the truncated state space.

The expected Kullback-Leibler (KL) information loss from truncating the state space is

$$\delta(k) = T^{-1}E\left(L(\theta) - L_k(\theta)\right) = T^{-1}\sum_{t=1}^{T} E\left(\log\left(\frac{f(y_t \mid \xi_t, x_t, \Im_{t-1})'\xi_{t|t-1}}{f(y_t \mid \xi^k_t, x_t, \Im_{t-1})'\xi^k_{t|t-1}}\right)\right) > 0.$$

To assess the effect of truncating the state space, I simulate from the MB model and estimate $\delta(k)$ using the average likelihood difference. I use the following MB settings:

$$y_t = \beta_{1t} + x_t \beta_{2t} + \varepsilon_t \tag{23}$$

$$x_t \sim iidN(0,1)\,,\ \varepsilon_t \sim iidN(0,\sigma_t^2)$$

$$\beta_{1t} \sim N(1, v_0)\,,\ \ \beta_{2t} \sim N(2, v_0)\,,\ v_0 \in \{0.04, 0.25, 1\}$$

$$\sigma_t^{-2} \sim G(1, \eta_0)\,,\ \eta_0 \in \{20, 10, 5\}$$

$$p_{00} \in \{0.995, 0.95\},\ p_{11}=1-p_{00}.$$

To understand the magnitude of the breaks implied by the chosen values of v_0, consider the variance of the regression signal relative to its variance if there were no breaks, i.e.,

$$\frac{E(\beta_t' x_t x_t' \beta_t)}{E(\beta_t' x_t x_t' \beta_t \mid \beta_t = \beta_0)} = \frac{E(\beta_t'\beta_t)}{\beta_0'\beta_0} = \frac{\beta_0'\beta_0 + 2v_0}{\beta_0'\beta_0} = 1 + 0.4v_0\,,$$

using $\beta_0=(1,2)'$. Thus, for v_0 = 0.04, 0.25, and 1, the breaks account for 1.6 percent, 10 percent, and 40 percent of the signal variance. I label these three v_0 values the small, medium, and large break settings, respectively. From the properties of the Gamma distribution, $\text{var}(\sigma_t^{-2}) = 2\eta_0^{-1}\sigma_0^{-4} = 2\eta_0^{-1}$, so for η_0 = 20, 10, and 5, the variance of σ_t^{-2} equals 0.1, 0.2, and 0.4, respectively.

Figure 1 presents the percentage KL loss from the MB(k) model for six settings. The plots show that the required value of k to ensure zero information loss decreases in the break probability and the break size. If p_{00}=0.95, p_{11}=0.05, v_0=1, and η_0=5 (frequent large breaks), then setting k=20 ensures zero information loss. For the case with rare small breaks (p_{00}=0.995, p_{11}=0.005, v_0=0.04, η_0=20), k=350 is required to ensure zero information loss. However, even in this extreme case, the information loss is less than 1% for $k \geq 10$ and less than 0.1% for $k \geq 80$.

Figure 2 shows how the likelihood function varies with the parameters for various values of k for the setting (p_{00}=0.99, p_{11}=0.01, v_0=1, η_0=5). Panel A varies the transition probability ($1-p_{00}$) and Panel B varies the mean of the distribution from which the slope coefficient is drawn (β_{20}).

The figure shows little variation in the coefficient value that maximizes the likelihood, even for *k* as small as 10. Overall, there appears to be little advantage from choosing large values of *k*, even though correct specification implies choosing *k*=*T*. For most applications, *k*=25 would appear to be an appropriate starting point. If the parameter estimates indicate small and rare breaks, then KL divergence could be further reduced by increasing *k*.

5. Comparison to Other Breaks Models

5.1 Markov Switching

To model stochastic breaks in regression models, Chib (1998) and Timmermann (2001) specify *N*-state Markov switching models with nonrecurring states (MSNR). Under this model

$$y_t = x_t'\beta_{s_t} + \sigma_{s_t}\varepsilon_t,$$

where $\varepsilon_t \sim N(0,1)$, $\beta_{s_t} \in \{\beta_1,...,\beta_N\}$, $\sigma_{s_t} \in \{\sigma_1,...,\sigma_N\}$, and the state variable s_t is an *N*-dimensional reducible Markov chain with transition probability matrix

$$Q = \begin{bmatrix} q_{11} & 0 & 0 & \cdots & 0 \\ 1-q_{11} & q_{22} & 0 & \cdots & 0 \\ 0 & 1-q_{22} & q_{33} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 1-q_{N-1,N-1} & 1 \end{bmatrix}.$$

In contrast to the MB model, this Markov switching model conditions on the coefficients and error variance, treating them as parameters to be estimated. This approach requires the number of possible breaks (*N*–1) to be specified *a priori*, which reduces flexibility for out-of-sample forecasting. Moreover, this approach requires that each regime be long enough to identify the parameters within that regime. In contrast, the MB model may exhibit regimes as short as one period, which can be useful if the process does not change to a new regime immediately, but rather needs a period of adaptation.

To compare the information loss from the MB(*k*) and MSNR models, I estimate the expected predictive likelihood of each model in various parameter settings. I denote the expected predictive

likelihood by $E_{\hat{\theta}}\left(E_Y\left(L(Y;\hat{\theta})\right)\right)$, where the first expectation is taken over the estimated parameters and the second expectation is taken over a predictive sample (Cooley and Parke 1990). Following White (1982), the expected predictive likelihood quantifies the Kullback-Leibler divergence between the model and the true data generating density. To estimate $E_{\hat{\theta}}\left(E_Y\left(L(Y;\hat{\theta})\right)\right)$, I simulate 1000 observations from the process in (23) and use the first 500 to estimate the parameters of the MB(k) and MSNR models. Next, I calculate the predictive likelihood in (6) over the last 500 observations. I repeat this process 100 times and average the predictive likelihood across the iterations.

When calculating the predictive likelihood, I do not re-estimate the parameters of the MB(k) as I iterate through the sample. However, because the MSNR model has no capacity to predict post-break values of β_t and σ_t, I re-estimate the parameters of this model before calculating the predictive likelihood for each observation. Specifically, for the MSNR model I begin by estimating the parameters using the first 500 observations. I use these estimates to calculate the predictive likelihood for observation 501, before re-estimating the parameters using the first 501 observations, calculating the likelihood for observation 502 etc. I assume that the number of states is known when estimating the model, although to reduce computation time I truncate the maximum number of estimated states to be 10. To understand the impact of assuming a known number of states, I also calculate the predictive likelihood for an MSNR model that selects the number of states using the Markov switching criterion (MSC) of Smith, Naik and Tsai (2006).

Table 1 contains the average KL divergence between an MB(25) model and MSNR models with the number of states known and estimated by MSC. For comparison, I also include results from a Markov switching model with recurring states (MS), rolling OLS regressions, and a fixed-sample OLS regression that ignores the possibility of breaks. I assume a normal likelihood function for these comparison models, and I use MSC to select the number of states in the MS model. As for the MB(25) model, I do not re-estimate the parameters of the Markov switching

and fixed OLS models as I iterate through the sample. However, the rolling regressions, by their very nature, must be re-estimated iteratively in the same manner as the MSNR model.

The MB(25) model outperforms all of the other models. For large breaks (v_0=1, η_0=5), the KL divergence exceeds 100 for the MSNR models and is close to 400 for the MS and OLS models. To place these divergence numbers in context, Panel B of Table 1 shows that the MSNR models are more than 10 percent worse, and the MS and OLS models are close to 50 percent worse than the MB(25) for large breaks. For smaller breaks the competing models come closer to the MB(25) model, culminating in less than 5 percent divergence in the small breaks case.

The OLS regression that ignores the breaks always performs worst. This model provides a benchmark for how much can be gained by incorporating the breaks. These potential gains increase in the size and frequency of the breaks. The smaller and more rare are the breaks, the less costly is ignoring those breaks. The second worst model is the Markov switching model with recurring states. This model is misspecified because it assumes that the post-break parameters (β_t, σ_t^2) are drawn from a discrete distribution rather than the continuous normal-gamma distribution. The poor performance of the MS model indicates that this distributional misspecification affects the models performance considerably.

When using MSC to select the number of states in the MSNR model, rolling OLS regressions with 50 observations outperform the MSNR model in all settings. Even when the number of states is assumed known, the MSNR model only outperforms the rolling regressions in the case with small breaks. For the cases with more frequent breaks (p_{10}=0.01), this poor performance may be partially explained by the upper bound of 10 that I place on the number of states in the MSNR model. However, for p_{10}=0.005, this constraint almost never binds and it has a negligible effect on the average KL divergence. Overall, the MSNR model performs poorly relative to both rolling regressions and the MB(25) model.

Panel C Table 1 shows the proportion of times in the 100 iterations that each model generated a larger predictive likelihood than the MB(25) model. Only in the case with the smallest breaks can the best of the alternative models beat the MB(25) model a significant proportion of the time. This result reflects the estimation uncertainty that arises from a small number of in-sample breaks, especially if those breaks are small in magnitude. For the rare breaks case (p_{10}=0.005), the median draw contains only 3 breaks in the estimation sample and many draws contain only 1 break. This low number of breaks causes the MB model's parameters to be imprecisely estimated. Nonetheless, the MB(25) model still estimates the moments of β_t and σ_t^2 well enough to outperform the best alternative models 50 percent of the time for the smallest breaks and 93 percent of the time for large breaks. Moreover, Figure 1 suggests that increasing *k* could improve the MB model by as much as 0.6 percent in the rare breaks case.

5.2 Stochastic Permanent Breaks and Innovation Regime Switching

Consider the intercept-only model $y_t = \beta_t + \varepsilon_t$ with constant error variance. This model has some similar properties to the STOPBREAK model in Engle and Smith (1999) and the innovation regime switching (IRS) model in Kuan, Huang, and Tsay (2005). The MB model specifies $\beta_t = s_t z_t + (1-s_t)\beta_{t-1}$, where $z_t \sim N(\beta_0, V)$ is independent of ε_t. The basic IRS model specifies $\beta_t = s_t y_{t-1} + (1-s_t)\beta_{t-1}$, which has the same state variable but differs from the MB model in the way the post-break intercept is determined. The MB model takes a draw from an independent stationary distribution, whereas the IRS model uses the lagged value of y_t. Thus, the IRS model has no long-run link to a stationary level, and therefore it exhibits permanent breaks. Moreover, its post-break value of β_t is observable. On the other hand, the MB model contains transitory breaks and a latent post-break value of β_t.

Setting *k*=1 yields a MB model with similar properties to the basic STOPBREAK model. Using *k*=1, we have from (10)

$$\beta_{t|t} = s_{t|t}\hat{\beta}_{t|t} + (1 - s_{t|t})\bar{\beta}_{t-1|t} \tag{24}$$

where

$$\hat{\beta}_{t|t} = (1 + V_0^{-1})^{-1}(y_t + \beta_0 V_0^{-1}) = y_t - \frac{(y_t - \beta_0)V_0^{-1}}{1 + V_0^{-1}}, \tag{25}$$

and $\bar{\beta}_{t-1|t} = \beta_{t-1|t-1} + g_{t-1}(y_t - \beta_{t-1|t-1})$, where $g_{t-1} = \bar{V}_{t-1|t-1}/(1 + \bar{V}_{t-1|t-1})$ is the Kalman gain in (20). Putting (24) and (25) together yields

$$\beta_{t|t} = s_{t|t}\left(y_t - \frac{(y_t - \beta_0)V_0^{-1}}{1 + V_0^{-1}}\right) + (1 - s_{t|t})\left(\beta_{t-1|t-1} + g_{t-1}(y_t - \beta_{t-1|t-1})\right) \tag{26}$$

In contrast, Engle and Smith (1999) specify the model $\beta_{t|t} = q_t y_t + (1 - q_t)\beta_{t-1|t-1}$, where q_t is a $\Im_t$-measurable weighting function. Unlike the MB model, the function q_t does not necessarily represent the conditional probability of a break in period *t*.

The recursion in (26) includes two more terms than the STOPBREAK model. First, when no break occurs, the MB(1) model updates the estimate of β_t using the Kalman filter. This feature allows increases precision during stable periods. Second, the MB model uses prior information on the marginal distribution of β_t to modify the estimate of β_t in the event of a break. This tie to β_0 generates mean reversion in the process because the new draws of β_t come from a stationary distribution. If the user were to set V_0 to infinity, then the model would use no information from the marginal distribution to estimate the post-break value of β_t. Like the STOPBREAK model, the MB(1) model would not exhibit long-run mean reversion in this case.

6. The Conditional CAPM

The central tenet of asset pricing theory states that risk premia in asset returns depend on the sensitivity of an asset to aggregate risk. The CAPM (Sharpe 1964, Lintner 1965) is the most well known asset-pricing model; it measures the aggregate risk premium by the expected excess return on a market portfolio. A vast literature exists on tests of the unconditional CAPM, which assumes a constant beta, where beta represents the sensitivity of an asset to the market portfolio. This

literature reveals that the constant-beta CAPM misprices portfolios based on stocks with the largest returns in recent periods, portfolios based on company size, and portfolios based on the ratio of book equity to market equity (Fama and French 1992, 1993, Jegadeesh and Titman 1996). However, the conditional CAPM, in which the betas change over time, can hold even when the unconditional model fails (Hansen and Richard 1987). In this section, I show that the MB(*k*) model fits well in this setting, and I use it to determine whether the conditional CAPM correctly prices momentum, size, and book-to-market (BM) portfolios.

I estimate the model

$$r_t = \alpha_t + \beta_t r_{M,t} + \varepsilon_{it}, \tag{27}$$

where r_t denotes the excess return on some portfolio, $r_{M,t}$ denotes the excess return on the market portfolio, $\varepsilon_t \mid (r_{M,t}, \alpha_t, \beta_t, \sigma_t) \sim N(0, \sigma_t^2)$, $\alpha_t \mid \sigma_t \sim N\left(\alpha_0, \sigma_t^2 V_\alpha\right)$, $\beta_t \mid \sigma_t \sim N\left(\beta_0, \sigma_t^2 V_\beta\right)$, and $\sigma_t^{-2} \sim G(\sigma_0^{-2}, \eta_0)$. The intercept term in (27) is known as Jensen's alpha and the model correctly prices the portfolio in period *t* if α_t=0.

I estimate (27) for monthly stock returns on three long-short portfolios representing the momentum, size and BM effects. These portfolios were constructed by Fama and French.[3] The momentum portfolio return equals the difference between the return on stocks with high returns over the most recent year (top 30 percent) and stocks that had low returns over the most recent year (bottom 30 percent). The size portfolio return equals the difference between the returns on large stocks (greater than median size) and small stocks (less than median size). Similarly, the BM portfolio return is the difference between returns on high BM stocks (top 30 percent) and low BM stocks (bottom 30 percent). The composition of the momentum portfolio is updated monthly, whereas the size and BM portfolios are updated annually. The excess return on the market equals the value-weighted return on all NYSE, AMEX, and NASDAQ stocks minus the one-month Treasury bill rate. The data span the period 1927:01-2006:02.

[3] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. I use the monthly series labeled UMB, SMB, and HML.

I begin in Section 6.1 by establishing the presence of breaks and showing that the MB model outperforms its competitors in an out-of-sample forecasting exercise. Then, in Section 6.2 I examine the conditional CAPM.

6.1 Econometric Model Specification and Forecasting Comparison

To assess whether the coefficients in (27) change over time, I apply three hypothesis tests: (i) the exponential statistic of Andrews-Ploberger (1994), which is designed to maximize average asymptotic power against the alternative of a single break in the coefficients; (ii) Elliott and Müller's (2006) *J*-test, which is asymptotically optimal across a broad class of alternative models; and (iii) Bai and Perron's (1998) sequential *F*-test, which estimates the number of breaks. I report these statistics in Table 2. For all three portfolios, each test rejects the null hypothesis of constant coefficients in (27).

Bai and Perron's sequential procedure estimates the number of breaks by sequentially testing the null hypothesis of L breaks against the alternative of L+1 breaks until the null cannot be rejected. This procedure suggests the presence of 6 breaks for the momentum portfolio, and 4 breaks each for the size and BM portfolios. However, for the size portfolio, the null hypothesis of L=6 breaks is rejected against the alternative of L=7, and for the BM portfolio, the null of L=5 is rejected against the alternative of L=6. Thus, there is some evidence of additional breaks in these two portfolios. Table 2 lists the estimated breaks dates from the Bai-Perron procedure. Each portfolio exhibits a depression-era break in 1932, and none of the portfolios contain breaks in the 1980s or 1990s. Between 1932 and 1980, the breaks tend not to coincide across the portfolios.

To assess the ability of the MB model to capture these breaks, I estimate (27) using data up to December 1970, and use the data from 1971–2006 to evaluate post-sample forecasting performance conditional on $r_{M,t}$. As in the simulation exercise in Section 5.1, I compare the MB model to Markov switching models, rolling regressions, and fixed-sample OLS. Table 3 shows the estimated Kullback-Leibler information loss from applying these alternative models rather

than an MB(24) model. I estimate the KL loss using AIC (Akaike 1973) for the estimation sample and the predictive likelihood for the post sample forecasting period (Cooley and Parke, 1990). In addition, I present in Table 3 the mean-squared forecast errors of the alternative models relative to the MB(24) model for the out-of-sample period.

From a model selection perspective, Burnham and Anderson (2002, p. 170) offer the following guidelines to assess degrees of confidence in alternative models: KL divergence between 0-2 indicates a substantial empirical support for a model; KL divergence between 4-7 suggests considerably less support; and KL divergence > 10 implies essentially no empirical evidence in favor of that model. By these criteria, the MB(24) dominates all of the other models both in and out of sample. The only case with KL divergence less than 10 is the MS model for the size portfolio in the out-of-sample period. Testing the null hypothesis of equal predictive likelihood values (Vuong 1989) mimics this conclusion, with the MB(24) model being significantly better out of sample than all models except the MS model for the size portfolio.

By the MSFE criterion, the MB(24) model also markedly outperforms the alternative methods for the momentum portfolio. The other two portfolios show less pronounced MSFE differences. For the size portfolio, none of the MSFE differences are statistically significant using a *t*-test (West 1996). For the BM portfolio, the MB(24) model significantly outperforms the fixed OLS and MS models, with insignificant MSFE differences relative to the MSNR and rolling OLS models. Overall, given that KL divergence is a well accepted measure of information loss (Akaike 1973, White 1982, Cooley and Parke 1990), these results indicate that the MB(24) fits the data significantly better than competing breaks models, both in and out of sample.

6.2 Testing the Conditional CAPM

To examine the conditional CAPM, I estimate the model (27) using the full sample from 1927:01-2006:12, both with and without the intercept term, and present the results in Table 4. The conditional CAPM holds if $\alpha_t=0$ for all t. To test the null hypothesis that the conditional CAPM

holds, I use LR and Wald tests as outlined in Remark 7. The momentum portfolio does not conform to CAPM. Moreover, the estimated pricing error is constant at 0.82 throughout the sample because of the small value of V_α, which denotes the variance of the marginal distribution of α_t. A likelihood ratio test confirms this result, with the LR statistic of 90.47 far exceeding the 5 percent critical value of 5.14.

The BM portfolio more strongly rejects the conditional CAPM than the size portfolio, although the evidence is somewhat mixed for both portfolios. The average pricing error α_0 does not significantly differ from zero for either portfolio, with *t*-statistics of 0.31 and 1.31 far below the 5 percent critical value of 1.96. The Wald statistic for testing the null hypothesis that V_α=0, takes the values 0.55 and 2.43 for the size and BM portfolios. Given 5 and 10 percent critical values of 2.71 and 1.64, these tests show some evidence against CAPM for the BM portfolio but not for the size portfolio. However, for both portfolios, the LR test rejects at 5 percent the joint null hypothesis that α_0 and V_α equal zero, which implies rejection of the conditional CAPM.

Using a model with time-varying betas and constant alpha, Ang and Chen (2006) find little evidence that alpha differs significantly from zero for the BM portfolio over the period 1927-2001. Their result coheres with the result in Table 4 that the average pricing error α_0 does not differ significantly from zero for this portfolio. However, consistent with the results of Lewellen and Nagel (2006), V_α significantly exceeds zero for this portfolio (at the 10 percent significance level). Therefore, the *conditional* alphas differ from zero in some periods, and the conditional CAPM misprices the BM portfolio in at least some periods. Periods of overpricing offset periods of underpricing such that the average level of mispricing is close to zero.

The MB(24) model estimates a large number of breaks for all three portfolios. On average, breaks occur in one third of periods for the size portfolio. This result emanates from the large estimated value of 0.88 for p_{11}, which indicates that a break in one period usually follows a break in the previous period. For the momentum and BM portfolios, the estimated break probabilities

are similar at 0.17 and 0.13, although the breaks process displays different dynamics across the two portfolios. The momentum portfolio rarely exhibits breaks in consecutive periods, as shown by the estimated value of 0.05 for p_{11}. In fact, the null hypothesis that $p_{11}=1-p_{00}$ cannot be rejected for this portfolio, indicating the absence of time dependence in the break process. For the BM portfolio, $p_{11}=0.5$ and $p_{00}=0.92$ indicating longer periods without breaks than the momentum portfolio, but when breaks do occur they occur in clusters.

Figure 3 plots the smoothed values $\alpha_{t|T}$ and $\beta_{t|T}$ for the three portfolios and illustrates the wide variety of dynamic structures the MB model can accommodate. I also generated these curves using $k=120$ in the MB model, and the curves were almost identical to those in Figure 3. In contrast to its constant $\alpha_{t|T}$, the momentum portfolio displays substantial volatility in $\beta_{t|T}$. This pattern likely reflects frequent changes in the composition of the portfolio because the cohort of firms with the largest returns changes frequently. One notable feature of the $\beta_{t|T}$ curve is its persistent drops to about -1 after the stock market crashes in 1929 and 2001. These drops reveal that stocks with the largest price drops during the crashes correlate more closely with the market in the 3 years or so following the crash. In spite of these fluctuations in $\beta_{t|T}$, the pricing error $\alpha_{t|T}$ remains constant.

For the size portfolio, the $\beta_{t|T}$ fluctuates between zero and 0.5 for most of the sample. Between 1943 and 1965, it is smooth in between several abrupt breaks. Outside of this period, $\beta_{t|T}$ exhibits greater volatility. The $\alpha_{t|T}$ term follows a similar pattern, with a calm period from 1943-65 and more volatile periods before 1943 and after 1965. The average pricing error ($\alpha_{t|T}$) is close to zero, as expected from the small estimate of α_0. The CAPM performs worst in the periods from 1927-29, 1965-83, and 1997-2006, when the average absolute price error $|\alpha_{t|T}|$ equals 0.48, 0.55, and 0.57 respectively. In the other periods in the sample, 1930-64 and 1984-96, the absolute pricing error equals 0.20 and 0.22. Moreover, during both the 1930-64 and 1984-96 periods, the likelihood for the MB model without an intercept exceeds the likelihood for the MB model with

an intercept. This result indicates that the size effect was negligible in these periods. The resumption of the size effect after 1996 coincides with the beginning of the prolonged bull market in US equities. Mispricing was negative initially, indicating greater than expected returns on large stocks. This pattern reversed in mid 1999 at the height of the dotcom boom, when the premium shifted to small stocks, where it remained through the 2001 crash until late 2003.

For the BM portfolio, $\beta_{t|T}$ declined steadily throughout the sample, and the pricing error was smallest when $\beta_{t|T}$ was close to zero. In the pre-1950 period when $\beta_{t|T}$ was mostly positive, the absolute pricing error $|\alpha_{t|T}|$ averaged 0.46, whereas it averaged 0.25 from 1951-79. The pricing error was especially small at 0.11 during the 1950s, the so-called golden age of the US economy. The 1950s constitute the only decade in the sample for which the no-intercept likelihood exceeds the likelihood for the model with an intercept. In the middle of 1980, $\beta_{t|T}$ dropped dramatically from zero to -0.7, and it remained negative for the rest of the sample. During the period from 1980-2006, which coincides with the great moderation in the US economy, the absolute pricing error averaged 0.46, which exceeds the 0.25 value for the 1951-79 period. The mean BM portfolio return was similar across these two periods; it equalled 0.37 in the 1951-79 period and 0.41 in the 1980-2006 period. Thus, mispricing of the BM portfolio by CAPM in the 1980s and 1990s arose because growth stocks became relatively more sensitive to the market portfolio, but did not see a commensurate relative increase in average returns.

Overall, allowing time variation in the alphas and betas does not explain the pricing anomalies in the CAPM. In particular, the momentum effect remains strong. The size effect has been prominent only in short bursts and the BM effect has been strong since 1980 after being less prominent in the preceding 30 years.

7. Conclusion

In this article I develop the MB(*k*) model for estimation and forecasting in regressions with changing coefficients and error variances. I parameterize the model using a two-state hidden

Markov process, which allows me to apply the standard Markov switching filter. Furthermore, I use two features of the model to keep the state space of low dimension. First, evaluating the likelihood in a particular period requires knowledge only of the most recent break date; it does not require knowledge of the entire sequence of break dates up to that period. Second, if the most recent break occurred long in the past, the exact date of the break carries little information for the period t likelihood. The resulting MB(k) model outperforms competing breaks models both in Monte Carlo simulations and in an application to the conditional CAPM.

The MB model generates conditional parameter estimates and forecasts by averaging over models that include progressively more historical data. This feature provides a link to the forecast combination literature (Timmermann 2006), in which averaging across models often improves forecasting performance. Moreover, it explains why the model performs well even when the breaks are small and therefore difficult to identify. Further research into the links between forecast combination and the MB model will further improve forecasting and inference in the presence of breaks and model uncertainty.

Appendix: Algorithm for Filtered and Smoothed Inference

Model:

$$y_t = x_t'\beta_t + \varepsilon_t, \qquad t = 1, 2, \ldots, T$$

where $\varepsilon_t \mid (x_t, \beta_t, \sigma_t) \sim N(0, \sigma_t^2)$, $(\beta_t, \sigma_t \mid s_t = 1) \perp (x_t, \Im_{t-1})$, $\beta_t \mid \sigma_t \sim N\left(\beta_0, \sigma_t^2 V_0\right)$, $\sigma_t^{-2} \sim G(\sigma_0^{-2}, \eta_0)$, and $\Im_{t-1}$ denotes the information in $(y_1, y_2, ..., y_{t-1}, x_1, x_2, ..., x_{t-1})$. Define the Markov random variable $s_t \in \{0,1\}$ such that s_t=1 implies a new draw of (β_t, σ_t) in period t, and s_t=0 implies $\beta_t = \beta_{t-1}$ and $\sigma_t = \sigma_{t-1}$.

Filtered state probabilities:

$$\xi_{t|t}^k = \frac{f(y_t \mid \xi_t^k, x_t, \Im_{t-1}) * \xi_{t|t-1}^k}{\sum_{i=1}^{k+1} f(y_t \mid \xi_{it}^k = 1, x_t, \Im_{t-1}) \xi_{i,t|t-1}^k}, \qquad \xi_{t+1|t}^k = \begin{cases} P_t \xi_{t|t}^k & t \le k+1 \\ P \xi_{t|t}^k & t > k+1 \end{cases},$$

where

$$P_t = \begin{bmatrix} p_{11} & p_{01} & p_{01} & \cdots & p_{01} \\ p_{10} & 0 & 0 & \cdots & 0 \\ 0 & p_{00} & 0 & \cdots & 0 \\ 0 & 0 & p_{00} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_{00} \end{bmatrix}, \quad P = \begin{bmatrix} p_{11} & p_{01} & p_{01} & \cdots & p_{01} & p_{01} \\ p_{10} & 0 & 0 & \cdots & 0 & 0 \\ 0 & p_{00} & 0 & \cdots & 0 & 0 \\ 0 & 0 & p_{00} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{00} & p_{00} \end{bmatrix},$$

$p_{lm} = \Pr(s_t = m \mid s_{t-1} = l)$, $\xi_{1|0}^k = 1$, and * denotes element-by-element multiplication. The matrix P_t has dimension $t \times (t-1)$, and Γ has dimension $(k+1) \times (k+1)$.

Conditional density for i=0,1,…,k–1:

$$f(y_t \mid x_t, \xi_{i+1,t}^k = 1, \Im_{t-1}) = \frac{\Gamma\left((1+i+\eta_0)/2\right)\left(1 + \frac{(y_t - x_t'\hat{\beta}_{t-i|t-1})^2}{\hat{\sigma}_{t-i|t-1}^2 (1 + x_t'\hat{V}_{t-i|t-1}x_t')(i+\eta_0)}\right)^{-(1+i+\eta_0)/2}}{\Gamma\left((i+\eta_0)/2\right)\left(\pi\hat{\sigma}_{t-i|t-1}^2 (1 + x_t'\hat{V}_{t-i|t-1}x_t')(i+\eta_0)\right)^{1/2}},$$

$$\hat{\beta}_{t-i|t-1} = \hat{V}_{t-i|t-1}\left(V_0^{-1}\beta_0 + \sum_{j=1}^{i} x_{t-j} y_{t-j}\right),$$

$$\hat{V}_{t-i|t-1} = \left(V_0^{-1} + \sum_{j=1}^{i} x_{t-j} x_{t-j}'\right)^{-1},$$

$$\hat{\sigma}^2_{t-i|t-1} = \frac{\eta_0\sigma_0^2 + \sum_{j=1}^{i}\left(y_{t-j} - x'_{t-j}\hat{\beta}_{t-i|t-1}\right)^2 + \left(\hat{\beta}_{t-i|t-1} - \beta_0\right)' V_0^{-1}\left(\hat{\beta}_{t-i|t-1} - \beta_0\right)}{\eta_0 + i},$$

where ξ_{it} denotes the i^{th} element of ξ_t and $\Gamma()$ denotes the gamma function.

Approximate conditional density for *i=k*:

$$f(y_t \mid x_t, \xi^k_{i+1,t} = 1, \Im_{t-1}) = \frac{\Gamma\left(\frac{1+i+\overline{\eta}_{t-i|t-1}}{2}\right)\left(1 + \frac{(y_t - x_t'\overline{\beta}_{t-i|t-1})^2}{\overline{\sigma}^2_{t-i|t-1}(1 + x_t'\overline{V}_{t-i|t-1}x_t')(i+\overline{\eta}_{t-i|t-1})}\right)^{-(1+i+\overline{\eta}_{t-i|t-1})/2}}{\Gamma\left((i+\overline{\eta}_{t-i|t-1})/2\right)\left(\pi\overline{\sigma}^2_{t-i|t-1}(1 + x_t'\overline{V}_{t-i|t-1}x_t')(i+\overline{\eta}_{t-i|t-1})\right)^{1/2}}$$

$$\overline{\beta}_{t-k+1|t} = \overline{s}_{t-k+1|t}\hat{\beta}_{t-k+1|t} + (1-\overline{s}_{t-k+1|t})\left(\overline{\beta}_{t-k|t-1} + \overline{V}_{t-k|t-1}x_t\left(1 + x_t'\overline{V}_{t-k|t-1}x_t\right)^{-1}(y_t - x_t'\overline{\beta}_{t-k|t-1})\right)$$

$$\overline{V}_{t-k+1|t} = \overline{s}_{t-k+1|t}\hat{V}_{t-k+1|t} + (1-\overline{s}_{t-k+1|t})\left(\overline{V}_{t-k|t-1} - \overline{V}_{t-k|t-1}x_t\left(1 + x_t'\overline{V}_{t-k|t-1}x_t\right)^{-1}x_t'\overline{V}_{t-k|t-1}\right)$$

$$\overline{\sigma}^{-2}_{t-k+1|t} = \overline{s}_{t-k+1|t}\hat{\sigma}^{-2}_{t-k+1|t}$$
$$+(1-\overline{s}_{t-k+1|t})\left(\frac{\overline{\eta}_{t-k|t-1}\overline{\sigma}^2_{t-k|t-1} + \left(y_t - x_t'\overline{\beta}_{t-k|t}\right)^2 + \left(\overline{\beta}_{t-k|t} - \overline{\beta}_{t-k|t-1}\right)'\overline{V}^{-1}_{t-k|t-1}\left(\overline{\beta}_{t-k|t} - \overline{\beta}_{t-k|t-1}\right)}{\overline{\eta}_{t-k|t-1} + 1}\right)^{-2}$$

$$\overline{\eta}_{t-k+1|t} = \overline{s}_{t-k+1|t}(\eta_0 + k) + (1-\overline{s}_{t-k+1|t})\left(\overline{\eta}_{t-k|t-1} + 1\right)$$

$$\overline{s}_{t-k+1|t} = \frac{\xi^k_{k,t|t}}{\xi^k_{k,t|t} + \xi^k_{k+1,t|t}} \quad .$$

Filtered coefficient estimates:

$$\beta_{t|t} = \sum_{i=0}^{k-1}\hat{\beta}_{t-i|t}\xi^k_{i+1,t|t} + \overline{\beta}_{t-k|t}\xi^k_{k+1,t|t} \equiv B^k_{t|t}\xi^k_{t|t}$$

where $B^k_{t|t} \equiv \left[\hat{\beta}_{t|t} \quad \cdots \quad \hat{\beta}_{t-k+1|t} \quad \overline{\beta}_{t-k|t}\right]$.

Filtered error variance estimates:

$$\sigma^2_{t|t} = E\left(\sigma_t^2 \mid \Im_t\right) = \sum_{i=0}^{k-1}\frac{\eta_0 + i + 1}{\eta_0 + i - 1}\hat{\sigma}^2_{t-i|t}\xi^k_{i+1,t|t} + \frac{\overline{\eta}_{t-k|t}}{\overline{\eta}_{t-k|t} - 2}\overline{\sigma}^2_{t-k|t}\xi^k_{i+1,t|t} \equiv S^k_{t|t}\xi^k_{t|t}$$

where $S^k_{t|t} \equiv \left[\frac{\eta_0+1}{\eta_0-1}\hat{\sigma}^2_{t|t} \quad \cdots \quad \frac{\eta_0+k}{\eta_0+k-2}\hat{\sigma}^2_{t-k+1|t} \quad \frac{\eta_{t-k|t}}{\eta_{t-k|t}-2}\overline{\sigma}^2_{t-k|t}\right]$.

Smoothed state probabilities:

$$\xi^k_{t|T} = \xi^k_{t|t} * \{P'_{t+1}(\xi^k_{t+1|T} \div \xi^k_{t+1|t})\}$$

where $\div$ denotes element-by-element division.

Smoothed coefficient estimates:

$$\begin{aligned}
E(\beta_t \mid \Im_T) &= E(\beta_t \mid n_t = t+1, \Im_T)\Pr(n_t = t+1 \mid \Im_T) + E(\beta_t \mid n_t > t+1, \Im_T)\Pr(n_t > t+1 \mid \Im_T) \\
&= \sum_{m=t}^{t+k-2} E(\beta_t \mid n_t = m+1, \Im_m)\Pr(n_t = m+1 \mid \Im_T) + E(\beta_t \mid n_t \geq t+k, \Im_T)\Pr(n_t \geq t+k \mid \Im_T) \\
&= \sum_{i=0}^{1}\sum_{m=t}^{t+k-2} E(\beta_t \mid s_t = i, n_t = m+1, \Im_m)\Pr(s_t = i, n_t = m+1 \mid \Im_T) \\
&\qquad + \sum_{i=0}^{1} E(\beta_t \mid s_t = i, n_t \geq t+k, \Im_T)\Pr(s_t = i, n_t \geq t+k \mid \Im_T)
\end{aligned}$$

where n_t denotes the date of the next break after period t. The second equality uses the fact that $E(\beta_t \mid s_t = i, n_t = m+1, \Im_T) = E(\beta_t \mid s_t = i, n_t = m+1, \Im_m)$ because of the break in period m+1. I approximate $E(\beta_t \mid s_t = i, n_t \geq t+k, \Im_T)$ by $E(\beta_t \mid s_t = i, n_t \geq t+k, \Im_{t+k-1})$, which implies that the smoothed coefficient estimates are:

$$\beta_{t|T} = \hat{B}_t \Pi_{t|T} + \bar{B}_t \Lambda_{t|T}$$

where

$$\hat{B}_t \equiv \left[\hat{\beta}_{t|t} \quad \hat{\beta}_{t|t+1} \quad \ldots \quad \hat{\beta}_{t|t+k-1}\right], \qquad \Pi_{t|T} \equiv \left[\pi_{tt|T} \quad \ldots \quad \pi_{t,t+k-2|T} \quad \vec{\pi}_{t,t+k-1|T}\right]',$$

$$\hat{\beta}_{t|m} \equiv E(\beta_t \mid s_t = 1, n_t = m+1, \Im_m), \qquad \pi_{tm|T} \equiv \Pr(s_t = 1, n_t = m+1 \mid \Im_T),$$

$$\vec{\pi}_{tm|T} \equiv \Pr(s_t = 1, n_t \geq m+1 \mid \Im_T)$$

$$\bar{B}_t \equiv \left[\bar{\beta}_{t|t} \quad \bar{\beta}_{t|t+1} \quad \ldots \quad \bar{\beta}_{t|t+k-1}\right], \qquad \Lambda_{t|T} \equiv \left[\lambda_{tt|T} \quad \ldots \quad \lambda_{t,t+k-2|T} \quad \vec{\lambda}_{t,t+k-1|T}\right]',$$

$$\bar{\beta}_{t|m} \equiv E(\beta_t \mid s_t = 0, n_t = m+1, \Im_m), \qquad \lambda_{tm|T} \equiv \Pr(s_t = 0, n_t = m+1 \mid \Im_T),$$

$$\vec{\lambda}_{tm|T} \equiv \Pr(s_t = 0, n_t \geq m+1 \mid \Im_T).$$

Notes:

(i) This smoother averages over various estimates of β_t conditional on whether a break occurred in period t and on when the next break occurs.

(ii) The dimension of the state variable ξ^k_t dictates the dimension of $\hat{B}_t$, $\bar{B}_t$, Π_t, and Λ_t.

(iii) Near the end of the sample, there are fewer than k–1 observations after t, which reduces the dimension of $\hat{B}_t$, $\bar{B}_t$, Π_t, and Λ_t. Specifically, for t>T–k+1, $\hat{B}_t \equiv \left[\hat{\beta}_{t|t} \quad \ldots \quad \hat{\beta}_{t|T}\right]$, $\bar{B}_t \equiv \left[\bar{\beta}_{t|t} \quad \ldots \quad \bar{\beta}_{t|T}\right]$, $\Pi_{t|T} \equiv \left[\pi_{tt|T} \quad \ldots \quad \pi_{t,T-1|T} \quad \vec{\pi}_{t,T|T}\right]'$, and $\Lambda_{t|T} \equiv \left[\lambda_{tt|T} \quad \ldots \quad \lambda_{t,T-1|T} \quad \vec{\lambda}_{t,T|T}\right]'$.

Conditional coefficient estimates for smoothing:

Elements of $\hat{B}_t$: $\hat{\beta}_{t|m} = \left(V_0^{-1} + \sum_{j=t}^{m} x_j x_j'\right)^{-1} \left(V_0^{-1}\beta_0 + \sum_{j=t}^{m} x_j y_j\right)$

Forward recursions for elements of $\bar{B}_t$:

For all $m = t, \ldots, t+k-2$: $\bar{\beta}_{t|m} = \dfrac{\hat{\beta}_{t-1|m}\pi_{t-1,m|T} + \bar{\beta}_{t-1|m}\lambda_{t-1,m|T}}{\pi_{t-1,m|T} + \lambda_{t-1,m|T}}$

For $\bar{\beta}_{t|t+k-1}$, I use the linear approximation in (20) and (21). I present this recursion above in the section "*approximate conditional density for i=k*", so do not repeat it here.

Initialize recursions using $\bar{\beta}_{1|m} = \hat{\beta}_{1|m}$.

Conditional probabilities for smoothing:

From (16) and (17),

$$\pi_{tm|T} = \xi^k_{m-t+1,m|T} - \xi^k_{m-t+2,m+1|T}$$

$$\vec{\pi}_{tm|T} = \xi^k_{m-t+1,m|T}$$

Similarly,

$$\begin{aligned}
\lambda_{tm|T} &= \Pr\left(s_t = 0, s_{t+1} = 0, \ldots, s_m = 0, s_{m+1} = 1 \mid \Im_T\right) \\
&= \Pr\left(s_t = 0, s_{t+1} = 0, \ldots, s_m = 0 \mid \Im_T\right) - \Pr\left(s_t = 0, s_{t+1} = 0, \ldots, s_m = 0, s_{m+1} = 0 \mid \Im_T\right) \\
&= \Pr\left(s_{t-1} = 1, s_t = 0, \ldots, s_m = 0 \mid \Im_T\right) + \Pr\left(s_{t-1} = 0, s_t = 0, \ldots, s_m = 0 \mid \Im_T\right) \\
&\qquad - \Pr\left(s_{t-1} = 1, s_t = 0, \ldots, s_{m+1} = 0 \mid \Im_T\right) - \Pr\left(s_{t-1} = 0, s_t = 0, \ldots, s_{m+1} = 0 \mid \Im_T\right) \\
&= \sum_{j=m-k+1}^{t-1} \Pr\left(s_j = 1, s_{j+1} = 0, \ldots, s_m = 0 \mid \Im_T\right) + \Pr\left(s_{m-k+1} = 0, s_{m-k+2} = 0, \ldots, s_m = 0 \mid \Im_T\right) \\
&\quad - \sum_{j=m-k+2}^{t-1} \Pr\left(s_j = 1, s_{j+1} = 0, \ldots, s_{m+1} = 0 \mid \Im_T\right) - \Pr\left(s_{m-k+2} = 0, s_{m-k+3} = 0, \ldots, s_{m+1} = 0 \mid \Im_T\right) \\
&= \sum_{j=m-t+2}^{k+1} \xi^k_{j,m|T} - \sum_{j=m-t+3}^{k+1} \xi^k_{j,m+1|T},
\end{aligned}$$

and

$$\vec{\lambda}_{tm|T} = \Pr\left(s_t = 0, s_{t+1} = 0, \ldots, s_m = 0 \mid \Im_T\right) = \sum_{j=m-t+2}^{k+1} \xi^k_{j,m|T}.$$

Smoothed error variance:

$$\sigma^2_{t|T} = \hat{S}_t \Pi_{t|T} + \overline{S}_t \Lambda_{t|T}$$

where

$$\hat{S}_t \equiv \left[\frac{\eta_0+1}{\eta_0-1}\hat{\sigma}^2_{t|t} \quad \frac{\eta_0+2}{\eta_0}\hat{\sigma}^2_{t|t+1} \quad \cdots \quad \frac{\eta_0+k}{\eta_0+k-2}\hat{\sigma}^2_{t|t+k-1} \right],$$

$$\overline{S}_t \equiv \left[\frac{\overline{\eta}_{t|t}}{\overline{\eta}_{t|t}-2}\overline{\sigma}^2_{t|t} \quad \frac{\overline{\eta}_{t|t+1}}{\overline{\eta}_{t|t+1}-2}\overline{\sigma}^2_{t|t+1} \quad \cdots \quad \frac{\overline{\eta}_{t|t+k-1}}{\overline{\eta}_{t|t+k-1}-2}\overline{\sigma}^2_{t|t+k-1} \right]$$

Conditional error variance estimates for smoothing:

Elements of $\hat{S}_t$: $$\hat{\sigma}^2_{t|m} = \frac{\eta_0\sigma_0^2 + \sum_{j=t}^{m}\left(y_j - x_j'\hat{\beta}_{t|m}\right)^2 + \left(\hat{\beta}_{t|m} - \beta_0\right)' V_0^{-1}\left(\hat{\beta}_{t|m} - \beta_0\right)}{\eta_0 + m - t + 1}$$

Forward recursions for elements of $\overline{S}_t$:

For all $m = t, \ldots, t+k-2$:

$$\overline{\sigma}^{-2}_{t|m} = \frac{\hat{\sigma}^{-2}_{t-1|m}\pi_{t-1,m|T} + \overline{\sigma}^{-2}_{t-1|m}\lambda_{t-1,m|T}}{\pi_{t-1,m|T} + \lambda_{t-1,m|T}}$$

$$\overline{\eta}_{t|m} = \frac{(\eta_0 + m - t + 2)\pi_{t-1,m|T} + \overline{\eta}_{t-1|m}\lambda_{t-1,m|T}}{\pi_{t-1,m|T} + \lambda_{t-1,m|T}}$$

For $\overline{\sigma}^{-2}_{t|t+k-1}$ and $\overline{\eta}_{t|t+k-1}$, I use the approximation in (22). I present this recursion above in the section "*approximate conditional density for i=k*", so do not repeat it here.

Initialize recursions using $\overline{\sigma}^2_{1|m} = \hat{\sigma}^2_{1|m}$ and $\overline{\eta}_{1|m} = (\eta_0 + m)$.

References

Akaike, H., 1973. *Information Theory and an Extension of the Maximum Likelihood Principle*. In 2nd International Symposium on Information Theory, B.N. Petrov and F. Csaki (Eds.), 267-281. Budapest: Akademia Kiado.

Andrews, D.W.K., 2001, "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69(3): 683-734.

Andrews, D.W.K. and W. Ploberger, 1994, "Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative," *Econometrica*, 62(6): 1383-1414.

Ang, A. and J. Chen, 2006, "CAPM Over the Long Run: 1926-2001," *Journal of Empirical Finance*, forthcoming.

Bai, J. and P. Perron, 1998, "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66(1): 47-78.

Burnham, K.P. and D.R. Anderson, 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd Ed). New York: Springer.

Chib, S., 1998, "Estimation and Comparison of Multiple Change-Point Models," *Journal of Econometrics*, 86(2): 221-241.

Clark, T.E. and M.W. McCracken, 2005, "The Power of Tests of Predictive Ability in the Presence of Structural Breaks," *Journal of Econometrics*, 124(1): 1-31.

Cooley, T.F., and W.R. Parke, 1990, "Asymptotic Likelihood-Based Prediction Functions," *Econometrica*, 58(5): 1215-34.

Davies, R.B., 1977, "Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternative," *Biometrika*, 64(2): 247-54.

Elliott, G. and U.K. Müller, 2006, "Optimally Testing General Breaking Processes in Linear Time Series Models," *Review of Economic Studies*, forthcoming..

Engle, R.F. and A.D. Smith, 1999, "Stochastic Permanent Breaks," *Review of Economics and Statistics*, 81(4): 553-574.

Fama, E.F. and K.R. French, 1992, "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47(2): 427-65.

Fama, E.F. and K.R. French, 1993, "Common Risk Factors in the Returns on Stock and Bonds," *Journal of Financial Economics*, 33(1): 3-56.

Hamilton, J.D., 1989, "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57(2): 357-84.

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

Hansen, L.P. and S.F. Richard, 1987, "The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models," *Econometrica*, 55(3): 587-613.

Harrison, P.J. and C.F. Stevens, 1976, "Bayesian Forecasting," *Journal of the Royal Statistical Society, Series B*, 38(3): 205-247.

Hildreth, C. and J.P. Houck, 1968, "Some Estimators for a Linear Model with Random Coefficients," *Journal of the American Statistical Association*, 63(322): 584–95.

Hinkley, D., 1979, "Predictive Likelihood," *The Annals of Statistics*, 7(4): 718-728.

Jegadeesh, N. and S. Titman, 1993, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance* 48(1): 65-91.

Kim, C. J., 1994, "Dynamic Linear Models with Markov-Switching," *Journal of Econometrics*, 60(1): 1–22.

Koop, G., 2003, *Bayesian Econometrics*, John Wiley & Sons: West Sussex.

Kuan, C.M., Y.L. Huang, and R.S. Tsay, 2005, "Switching Permanent and Transitory Innovations," *Journal of Business & Economic Statistics*, 23(4): 443-454.

Lauritzen, S.L., 1974, "Sufficiency, Prediction, and Extreme Models," *Scandinavian Journal of Statistics*, 1: 128-134.

Lewellen, J. and S. Nagel, 2006, "The Conditional CAPM does not Explain Asset Pricing Anomalies," *Journal of Financial Economics*, forthcoming.

Lintner, J., 1965, "Security Prices, Risk, and Maximal Gains from Diversification," *Journal of Finance*¸ 20(4): 587-615.

Pesaran, M.H, D. Pettenuzzo , and A. Timmermann, 2006, "Forecasting Time Series Subject to Multiple Structural Breaks," *Review of Economic Studies*, forthcoming.

Pesaran, M.H. and A. Timmermann, 2006, "Selection of Estimation Window in the Presence of Breaks,".*Journal of Econometrics*, forthcoming.

Self, S.G. and K.Y. Liang, 1987, "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under non-Standard Conditions," *Journal of the American Statistical Association*, 82(398): 605–610.

Sharpe, W.F., 1964, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *Journal of Finance*, 19(3): 425-442.

Smith, A., P.A. Naik, and C.L. Tsai, 2006, "Markov-Switching Model Selection using Kullback-Leibler Divergence," *Journal of Econometrics*, forthcoming.

Timmermann, A., 2001, "Structural Breaks, Incomplete Information, and Stock Prices," *Journal of Business and Economic Statistics*, 19(3): 299-314.

Timmermann, A., 2006. "Forecast combinations," in *Handbook of Economic Forecasting* (Edited by Elliott, G., C.W.J. Granger, and A. Timmermann) North Holland.

Vuong, Q.H., 1989, "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57(2): 307-33.

West, K.D., 1996, "Asymptotic Inference about Predictive Ability," *Econometrica*, 64(5): 1067-84.

White, H., 1982, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50(1): 1-25.

Table 1: Monte Carlo Comparison of Markov Breaks Model to Alternatives

Break Probability	Break Size	Ave LLF MB(25)	MSNR (true)	MSNR (MSC)	MS (MSC)	Rolling OLS(50)	Rolling OLS(200)	OLS
			Panel A: Average Kullback-Leibler Divergence from MB(25)					
p_{10}=0.005	v_0=.04, η_0=20	–741.6	-0.4	6.5	15.5	2.9	0.9	22.7
p_{10}=0.01	v_0=.04, η_0=20	–741.8	2.7	6.7	15.8	4.4	3.7	21.0
p_{10}=0.005	v_0=.25, η_0=10	–777.6	19.7	35.4	79.7	7.9	30.2	113.7
p_{10}=0.01	v_0=.25, η_0=10	–770.1	28.9	57.5	102.2	21.8	59.8	131.0
p_{10}=0.005	v_0=1, η_0=5	–810.2	102.4	125.1	391.4	47.3	152.7	395.6
p_{10}=0.01	v_0=1, η_0=5	–795.8	108.1	161.4	378.7	76.0	207.8	422.2
			Panel B: Percentage Kullback-Leibler Divergence from MB(25)					
p_{10}=0.005	v_0=.04, η_0=20		-0.0	0.9	2.1	0.4	0.1	3.1
p_{10}=0.01	v_0=.04, η_0=20		0.1	0.9	2.1	0.6	0.5	2.8
p_{10}=0.005	v_0=.25, η_0=10		2.5	4.5	10.3	1.0	3.9	14.6
p_{10}=0.01	v_0=.25, η_0=10		3.8	7.5	13.3	2.8	7.8	17.0
p_{10}=0.005	v_0=1, η_0=5		12.6	15.4	48.3	5.8	18.8	48.8
p_{10}=0.01	v_0=1, η_0=5		13.6	20.3	47.6	9.6	26.1	53.1
			Panel C: Proportion of Draws in which MB(25) is superior					
p_{10}=0.005	v_0=.04, η_0=20		0.50	0.70	0.94	0.67	0.54	0.95
p_{10}=0.01	v_0=.04, η_0=20		0.72	0.79	0.96	0.77	0.69	0.97
p_{10}=0.005	v_0=.25, η_0=10		0.82	0.88	0.98	0.81	0.84	0.99
p_{10}=0.01	v_0=.25, η_0=10		0.90	0.99	1.00	0.95	0.99	1.00
p_{10}=0.005	v_0=1, η_0=5		0.93	0.96	1.00	0.95	0.97	1.00
p_{10}=0.01	v_0=1, η_0=5		0.96	0.98	1.00	0.98	0.96	1.00

Note: This table shows average predictive likelihood for MB(25) model, and the performance of other models relative to the MB(25) model. I generated the elements from an average over 100 draws of 1000 observations from the model in equation (23). For each draw, I used the first 500 observations for parameter estimation and last 500 observations for post-sample predictive likelihood calculation conditional on x_t. Panel A contains the average difference over the 100 draws between the predictive likelihood of the MB(25) model and the alternative model, Panel B expresses the elements in Panel A as a percentage of the corresponding average predictive likelihood for the MB(25) model, and Panel C shows the proportion of draws for which the MB(25) predictive likelihood exceeds the predictive likelihood of the alternative model. *MSNR(true)* denotes Markov switching model with nonrecurring states and number of states known, *MSNR(MSC)* denotes Markov switching model with nonrecurring states and number of states selected by Markov switching criterion, *MS(MSC)* denotes Markov switching model with recurring states and number of states selected by MSC, *Rolling OLS(X)* denotes fixed-window rolling regressions of length X, and *OLS* denotes one-time OLS regression that ignores breaks.

Table 2: Tests for Changing Coefficients in CAPM Regressions

	Momentum	Size	Book to Market	Crit. Values
Elliott-Müller (*J* statistic)	–20.18*	–9.77*	–57.41*	–8.36
Andrews-Ploberger (Exp statistic)	14.94*	7.42*	25.53*	2.08
Bai-Perron (*F* statistics)				
UDMax	66.69*	54.22*	302.74*	13.27
sup F(1|0)	53.43*	54.22*	302.74*	12.89
sup F(2|1)	38.92*	22.85*	74.37*	14.50
sup F(3|2)	34.72*	42.80*	37.98*	15.42
sup F(4|3)	21.11*	29.50*	32.59*	16.16
sup F(5|4)	20.55*	13.68	14.45	16.61
sup F(6|5)	34.91*	13.68	23.80*	17.02
sup F(7|6)	12.24	17.53*	13.20	17.27
sup F(8|7)	11.49	13.64	10.06	17.55
Bai-Perron estimated break dates	Jul-32	Jul-32	May-32	
	Nov-37	Apr-47	Jun-36	
	Nov-42	Sep-64	Oct-55	
	May-70	Oct-80	May-80	
	Mar-75			
	Dec-01			

Note: Tests apply to the intercept and slope in the regression model in (27) using monthly returns from 1927:01-2006:02. See footnote 3 for a description of the data.

Table 3: Forecasting Performance in CAPM Regressions

	MB(24)	Performance Relative to MB(24)				
		MSNR (MSC)	MS (MSC)	Rolling OLS(24)	Rolling OLS(120)	OLS
Momentum Portfolio						
AIC difference	–1329.3	94.6	10.2			216.7
KL divergence	–1084.0	181.5* (5.05)	67.1* (2.91)	106.2* (5.53)	126.2* (5.63)	142.2* (7.34)
Relative MSFE	13.78	1.29* (4.46)	1.21* (2.38)	1.31* (3.80)	1.30* (4.21)	1.42* (4.17)
Size Portfolio						
AIC difference	–1215.1	82.6	34.1			156.4
KL divergence	–1033.7	100.2* (2.77)	5.7 (0.61)	86.2* (2.75)	51.5* (2.81)	49.3* (2.40)
Relative MSFE	10.30	0.97 (-0.86)	0.96 (-1.55)	1.05 (1.59)	0.99 (-0.27)	0.96 (-1.15)
BM Portfolio						
AIC difference	–1245.9	60.3	13.1			139.2
KL divergence	–996.0	53.2* (2.78)	63.7* (3.28)	57.9* (2.64)	41.3* (2.32)	196.9* (8.43)
Relative MSFE	7.86	1.02 (0.32)	1.45* (2.60)	1.05 (1.02)	0.97 (-0.69)	2.01* (6.29)

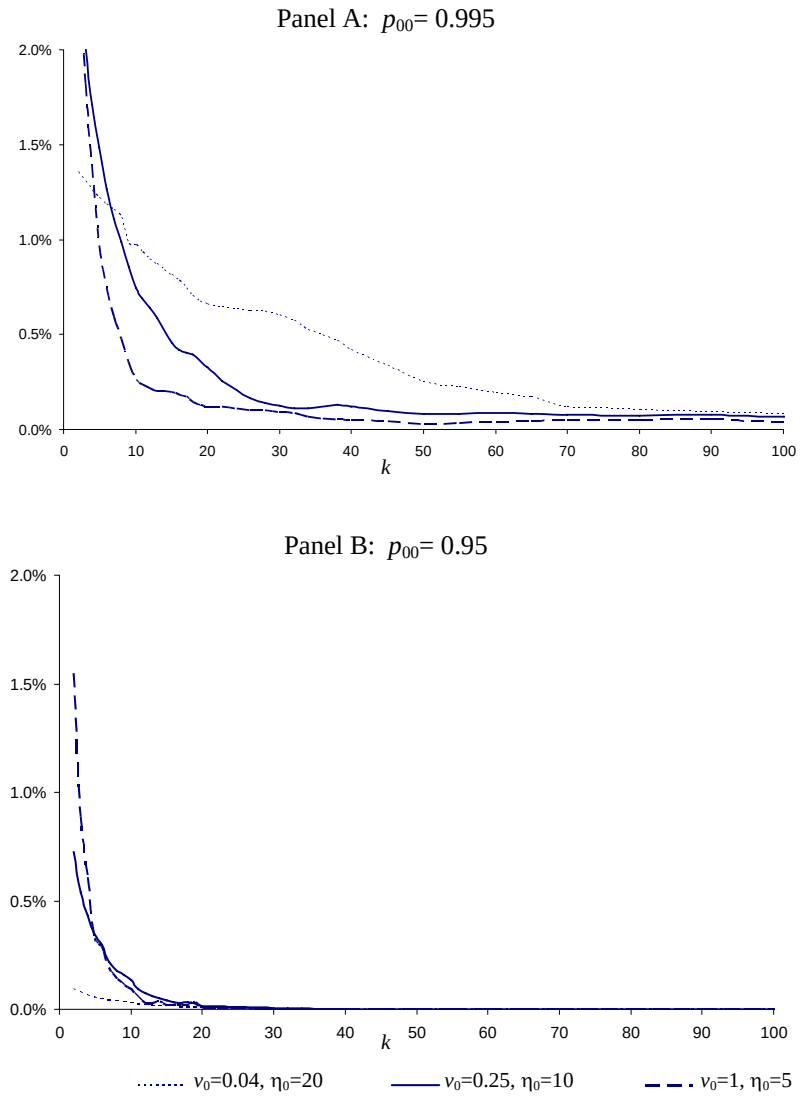
Note: I estimate the regression in (27) using the sample 1927:01-1970:12, and forecast over the period 1971:01-2006:02 conditional on $r_{M,t}$. See the Table 1 notes for a description of the alternative models and footnote 3 for a description of the data. For each portfolio, the MB(24) column shows AIC (in sample), predictive log likelihood (out of sample), and mean squared forecast error (out of sample). For the alternative models, the entries denote (i) the difference between the AIC for the MB(24) model and the AIC for the alternative model (AIC difference), (ii) the difference between the predictive log likelihood for the MB(24) model and the alternative model (KL divergence), and (iii) the ratio of the MSFE of the alternative model to the MB(24) model (relative MSFE). Defining K as the number of estimated parameters, $AIC = L(\theta) - 2K$. Below the out-of-sample stistics in parentheses are t-statistics for testing a zero difference between the MB(24) model and the alternative model; a * superscript denotes significance at 5 percent.

Table 4: MB(24) Estimates for CAPM Regressions (1927:01-2006:12)

	Momentum		Size		Book to Market	
α_0		0.82		0.05		0.14
		(0.09)		(0.15)		(0.11)
V_α		0.00		0.16		0.09
		(0.00)		(0.22)		(0.06)
β_0	0.11	0.04	0.16	0.20	0.02	0.00
	(0.05)	(0.04)	(0.06)	(0.06)	(0.06)	(0.02)
V_β	0.03	0.04	0.03	0.03	0.02	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
σ_0	2.17	1.99	1.84	1.76	1.90	1.82
	(0.13)	(0.11)	(0.14)	(0.15)	(0.25)	(0.14)
η_0	3.98	4.00	4.35	4.39	4.43	4.40
	(0.80)	(0.76)	(0.87)	(0.84)	(1.17)	(0.97)
p_{11}	0.00	0.05	0.90	0.88	0.58	0.50
	(0.00)	(0.23)	(0.05)	(0.05)	(0.54)	(0.00)
p_{00}	0.84	0.80	0.95	0.93	0.93	0.92
	(0.03)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
Pr(s_t=1)	0.13	0.17	0.33	0.35	0.14	0.13
t-stat H_0: $p_{11}=1-p_{00}$	-0.82	-0.32	6.45	7.57	1.19	8.08
Wald stat H_0: $V_\alpha=0$		0.00		0.55		2.43
LR stat H_0: $\alpha_0=V_\alpha=0$		90.47		10.82		17.39
LLF	–2441.27	–2396.04	–2232.92	–2227.51	–2241.42	–2232.72

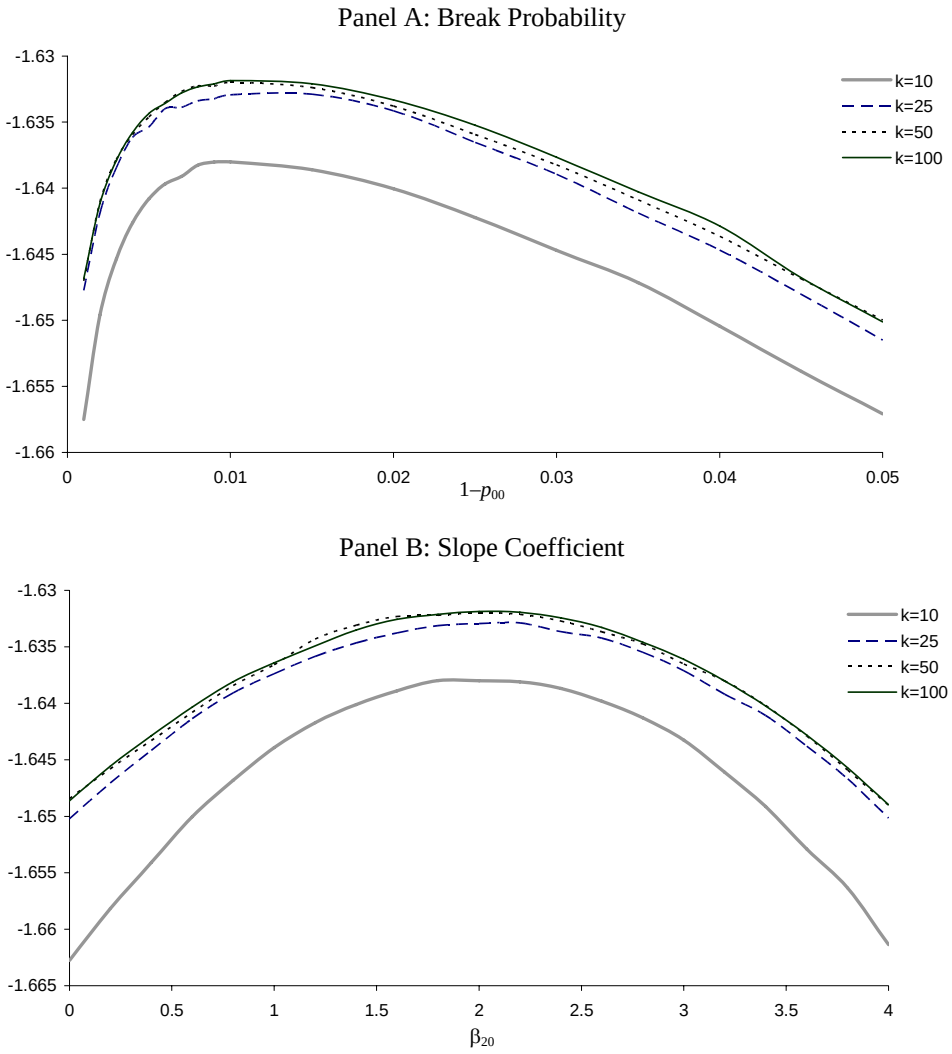
Note: I estimate the MB(24) model in (27) using the sample 1927:01-2006:02. See the Table 1 notes for a description of the alternative models and footnote 3 for a description of the data. I list robust standard errors in parentheses below the parameter estimates. The 5 percent critical value for the Wald test of H_0:V_α=0 is 2.71 and the 5 percent critical value for the LR test of H_0:$\alpha_0=V_\alpha=0$ is 5.14.

Figure 1: Percentage Expected Kullback-Leibler Loss from Truncating State Space



Note: Curves show percentage KL divergence between MB(k) model and nontrunctated MB model for the model in (23). Curves created from an average over a generated sample of size 20,000 with the parameters held at their true values.

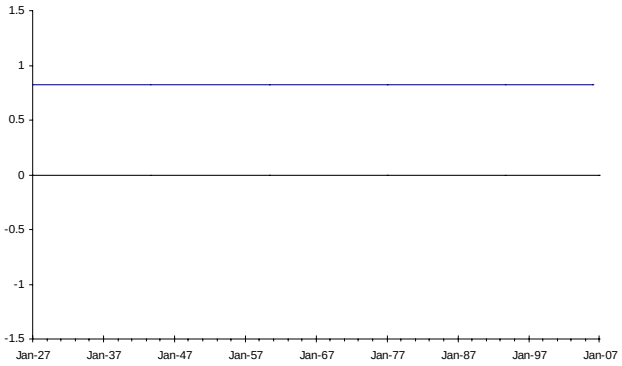
Figure 2: Likelihood Function for Various *k*



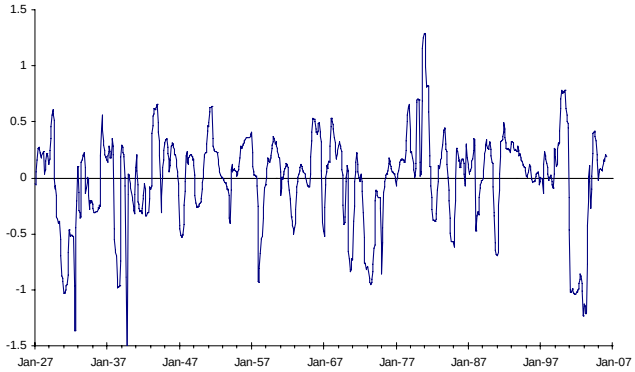
Note: Curves show value of average log likelihood as one parameter varies, holding the other parameters at their true values, for the model in (23). Curves created from an average over a generated sample of size 20,000.

Figure 3: Smoothed Coefficient Estimates

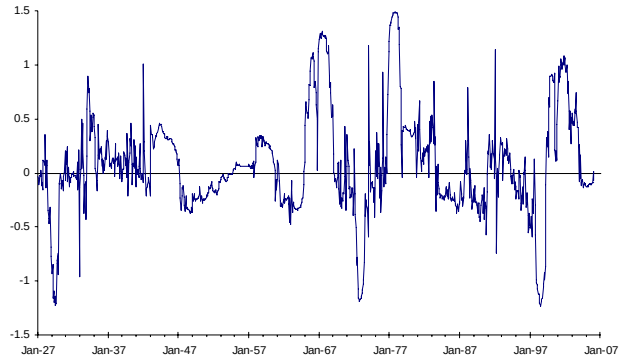
Momentum Portfolio: $\alpha_{t|T}$



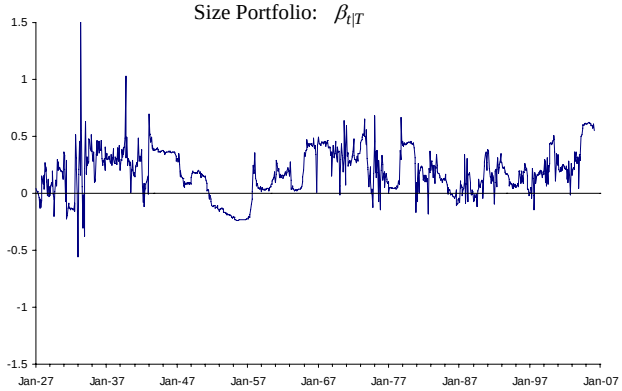
Momentum Portfolio: $\beta_{t|T}$



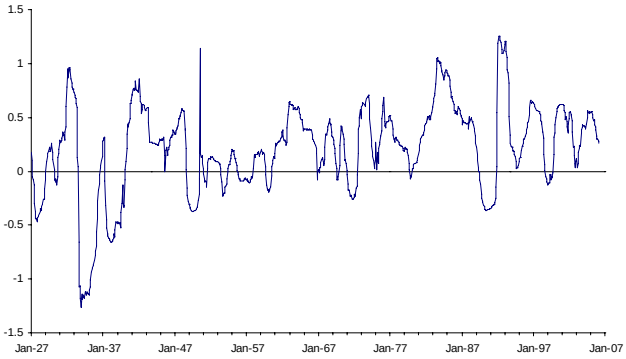
Size Portfolio: $\alpha_{t|T}$



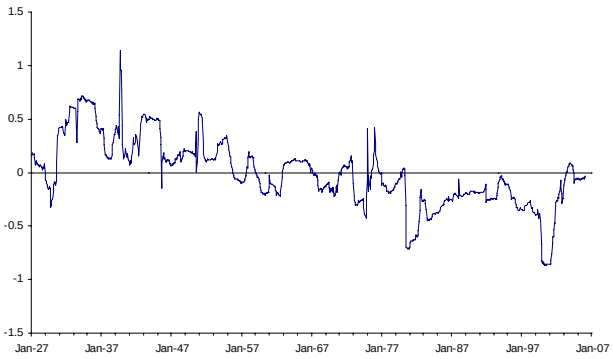
Size Portfolio: $\beta_{t|T}$



BM Portfolio: $\alpha_{t|T}$



BM Portfolio: $\beta_{t|T}$



Note: The graphs show smoothed coefficient estimates for the MB(24) model in (27) for three different portfolios. See the footnote 3 for details on the data set.