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 linear 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. The model generates forecasts and conditional parameter estimates using a probability weighted average over regressions that include progressively more historical data. I employ the MB model to study the predictive ability of the yield curve for quarterly GDP growth. I show evidence of breaks in the predictive relationship, and the MB model outperforms competing breaks models in an out-of-sample forecasting experiment.

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

∗ 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 (2007) 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 regressions that use progressively more data.

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, time-
varying parameter, and STOPBREAK models. In Section 6, I use the MB model predict quarterly GDP growth using the yield curve. I demonstrate the instability in the predictive relationship and show that the MB model outperforms competing breaks models in an out-of-sample forecasting experiment. Section 7 contains concluding remarks.

2. Markov Breaks Model

The MB model is

\[ y_t = x_t' \beta_t + \varepsilon_t, \quad t = 1, 2, \ldots, T \]

where \( \varepsilon_t | (x_t, \beta_t) \sim \text{iidN}(0, \sigma^2) \), \( x_t \) is a \( r \times 1 \) vector of explanatory variables that may include lags of \( y_t \), and \( \beta_t \sim N(\beta_0, \sigma^2 V_0) \) is a random coefficient vector. To conserve degrees of freedom I specify \( V_0 \) to be a diagonal matrix, although I do not impose this constraint in the derivations in Sections 2-4. 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 \( \beta_t \) that is independent of the observed data up to period \( t-1 \) and \( x_t \), i.e., \( (\beta_t | s_t = 1) \perp \mathcal{I} \) where \( \mathcal{I} \) denotes the information in \( (y_1, y_2, \ldots, y_{t-1}, x_1, x_2, \ldots, x_{t-1}) \). Although the coefficient vector \( \beta_t \) is a random variable, the parameters \( (\beta_0, V_0, \sigma^2) \) are 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. Then in Section 2.2, I extend the model to allow breaks in \( \sigma^2 \), before discussing the model’s properties in Section 2.3.

2.1 Likelihood Function

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

\[ L(\theta) = \sum_{t=1}^{T} \log \left( f(y_t | x_t, \mathcal{I}_{t-1}) \right), \]

where \( f \) denotes the density of \( y_t \) conditional on \( x_t \) and \( \mathcal{I}_{t-1} \). To obtain \( f \), I rewrite the model as

\[ y_t = x_t' \beta_{t-j} + \varepsilon_t, \]

(2)
where $t-j$ denotes the period of the most recent break. Then, defining the matrix $B_j = \begin{bmatrix} \beta_1 & \beta_{t-1} & \cdots & \beta_j \end{bmatrix}$, the model becomes

$$y_t = x_t' B_j \xi_t + \varepsilon_t,$$

where $\xi_t$ 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.,

$$\xi_t = \begin{bmatrix} s_t \\ (1-s_t)s_{t-1} \\ \vdots \\ (1-s_1)(1-s_2) \end{bmatrix}.$$

Using this nomenclature, the predictive density is

$$f(y_t \mid x_t, \mathcal{F}_{t-1}) = \sum_{i=1}^{t} f(y_t \mid \xi_{t,i} = 1, x_t, \mathcal{F}_{t-1}) \Pr(\xi_{t,i} = 1 \mid \mathcal{F}_{t-1}) = f(y_t \mid \xi_t, x_t, \mathcal{F}_{t-1})^T \xi_{i|t-1},$$

where $\xi_{i|t-1} = E(\xi_t \mid \mathcal{F}_{t-1})$ and the term $f(y_t \mid \xi_t, x_t, \mathcal{F}_{t-1})$ is a $t$ dimensional vector denoting the conditional density of $y_t$, in which the $i^{th}$ element corresponds to the event $\xi_{t,i} = 1$. Thus, the likelihood function is a probability weighted average of predictive densities that condition on possible dates of the most recent break.

The state variable $\xi_t$ 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 $2^t$. By reducing the state space dimension to $t$, I make 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 $f(y_t \mid \xi_t, x_t, \mathcal{F}_{t-1})$, I begin with the first observation. For $t=1$, we have

$$y_{t=1} \mid x_{1} \sim N \left( x_{1}' \beta_0, \sigma^2 (1 + x_{1}' \nu_{y,x_1}) \right).$$

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 \( \beta \), and

\[ y_2 \mid x_2, \xi_{1,2} = 1, \mathcal{F}_1 \sim N \left( x'_1 \hat{\beta}_0, \sigma^2 \left( 1 + x'_2 \hat{V}_0 x_2 \right) \right). \]

However, if no break occurs in period \( t = 2 \), then the second element of \( \xi_2 \) equals one and \( \beta_2 = \beta_1 \).

In this case, we have

\[ y_2 \mid x_2, \xi_{2,2} = 1, \mathcal{F}_1 \sim N \left( x'_1 \hat{\beta}_0, \sigma^2 \left( 1 + x'_2 \hat{V}_0 x_2 \right) \right). \]

Calculating the moments \( \hat{\beta}_{1|1} \) and \( \hat{V}_{1|1} \) requires updating inference about \( \beta_1 \) using the information in \( y_1 \). The model is

\[
\begin{bmatrix}
    y_1 | x_1 \\
    \hat{\beta}_1 | x_1
\end{bmatrix} \sim N \left( \begin{bmatrix} x'_1 \beta_0 \\ \beta_0 \end{bmatrix}, \begin{bmatrix} \sigma^2 \left( 1 + x'_2 V_0 x_1 \right) & \sigma^2 x'_1 V_0 \\ \sigma^2 V_0 x_1 & \sigma^2 V_0 \end{bmatrix} \right),
\]

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, \( \hat{\beta}_{1|1} = (V_0^{-1} + x_1 x'_1)^{-1} (V_0^{-1} \beta_0 + x_1 y_1) \) and \( \hat{V}_{1|1} = (V_0^{-1} + x_1 x'_1)^{-1} \).

Generalizing to period \( t \) and conditioning on the most recent break having occurred in period \( t-i \), we have

\[ y_i \mid x_i, \xi_{i+1,i} = 1, \mathcal{F}_{i-1} \sim N \left( x'_i \hat{\beta}_{i-\tilde{j}t-1}, \sigma^2 \left( 1 + x'_i \hat{V}_{i-\tilde{j}t-1} x_i \right) \right), \quad i = 0, 1, \ldots, t-1, \quad (4) \]

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

---

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, \( \hat{\beta}_{1|5} \) denotes an estimate of \( \beta_t \) conditional on data up to period \( t \) and knowledge that the most recent break occurred in period \( t-5 \).
The break indicator \( s_t \) follows a first order Markov process, which implies that \( \xi_t \) is also a first order Markov process. This specification allows application of the standard Markov-switching filter to obtain \( \xi_{it} \) (Hamilton 1989). The filter is

\[
\xi_{it} = f(y_t | \xi_t, x_t, \mathcal{A}_{t-1}) \ast \xi_{i(t-1)} \over f(y_t | \xi_t, x_t, \mathcal{A}_{t-1}) \xi_{i(t-1)}
\]

(5)

where \( \xi_{i(t-1)} = P_t \xi_{i(t-1)} \), \( \ast \) 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 \( p_{im} = \Pr(s_t = m | s_{t-1} = l). \)

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 \((\beta_0, V_0, \sigma^2, p_{00}, p_{11})\) and obtain standard errors through numerical differentiation of the log likelihood function. Next, I extend the model to allow for breaks in \( \sigma^2 \).

### 2.2 Allowing Breaks in \( \sigma^2 \)

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

\[
y_t = x_t' \beta_t + \varepsilon_t, \quad t = 1, 2, \ldots, T
\]

where \( \varepsilon_t \mid (x_t, \beta_t, \sigma_t) \sim iidN(0, \sigma_t^2) \), \( \beta_t \mid \sigma_t \sim N(\beta_0, \sigma_t^2 V_0) \), \( \sigma_t^{-2} \sim G(\sigma_0^{-2}, \eta_0) \), \( G \) denotes a Gamma distribution, and \((\beta_t, \sigma_t \mid s_t = 1) \perp (x_t, \mathcal{A}_{t-1})\). As for the case with constant error variance, the likelihood function is
\[ L(\theta) = \sum_{i=1}^{T} \log \left( f(y_i \mid \xi_i, x_i, \mathcal{J}_{t-1}) \right), \]  

(6)

where \( f(y_i \mid \xi_i, x_i, \mathcal{J}_{t-1}) \) is a \( t \) dimensional vector denoting the conditional density of \( y_i \), in which the \( i \)th element corresponds to the event \( \xi_{i,t} = 1 \). The unknown parameters to be estimated are \( \beta_0, V_0, \sigma_0^2, \eta_0, p_{00}, \) and \( p_{11} \).

Applying standard results that are most often used in Bayesian regression analysis with conjugate priors (Koop 2003, pg. 46), \( f(y_i \mid \xi_i, x_i, \mathcal{J}_{t-1}) \) is a \( t \)-distribution with \( \eta_0 + i \) degrees of freedom. Specifically,

\[ y_i \mid x_i, \xi_{i+1,t}, \mathcal{J}_{t-1} \sim t \left( x_i' \hat{\beta}_{t-i+1}, \hat{\sigma}_{t-i+1}^2 (1 + x_i' \hat{\Sigma}_{t-i+1} x_i), \eta_0 + i \right), \]  

(7)

for all \( i = 0, 1, \ldots, t-1 \), where \( \hat{\beta}_{t-i-1} = \left( V_0^{-1} + \sum_{j=1}^{i} x_{t-j} x'_{t-j} \right)^{-1} \left( V_0^{-1} \beta_0 + \sum_{j=1}^{i} x_{t-j} y_{t-j} \right) \) and \( \hat{\Sigma}_{t-i-1} = \left( V_0^{-1} + \sum_{j=1}^{i} x_{t-j} x'_{t-j} \right)^{-1} \) as in (4). As is the case in Bayesian regression analysis, the term \( \hat{\sigma}_{t-i-1}^2 \) 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 \( \beta_{t-i} \). Specifically,

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

(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 \( \beta_t \) and \( \sigma_t^2 \). 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), Nyblom (1989), and Elliott and Müller (2006). 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 \( \hat{\beta}_{i-j-1}, \hat{V}_{i-j-1}, \hat{\sigma}^2_{i-1}, \) and \( \hat{\xi}_{i-j-1} \) to remove the unknown \( \beta_{i-j}, \sigma^2_{i-1}, \) and \( \xi_j \) 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.
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 \( \beta_0, V_0, \sigma_0^2, \) and \( \eta_0 \) as priors on the distributions of \( \beta_t \) and \( \sigma_t^2 \). Alternatively, it could treat \( \beta_0, V_0, \sigma_0^2, \) and \( \eta_0 \) 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 \( \beta_t \) 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 \( \beta_t \) or \( \sigma_t^2 \), 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 \( \sum_{i=1}^{q} z_i^2 I(z_i > 0) + \sum_{i=q+1}^{2 q} z_i^2 \), 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 \( \beta \) and \( \sigma^2 \)

#### 3.1 Filtering

Conditional on \( \mathcal{I}_t \) and the most recent break having occurred in period \( t-i \), \( \beta_{t-i} \) has the multivariate \( t \) distribution

\[
\beta_{t-i} \mid \xi_{t+1:t} = 1, \mathcal{I}_t \sim t \left( \hat{\beta}_{t-i}, \hat{\sigma}_{t-i}^2 \hat{V}_{t-i}, \eta_0 + i + 1 \right),
\]

for all \( i = 0, 1, \ldots, t-1 \) (Koop 2003, pg. 37). From (3) and (1), the regression coefficients can be written as \( \beta_i = B_i \xi_i \), which implies that the filtered coefficient estimates are

\[
\beta_{t|t} = E(\beta_t \mid \mathcal{I}_t) = B_{t|t} \xi_{t|t},
\]

where \( B_{t|t} = \left[ \hat{\beta}_{1|t} \cdots \hat{\beta}_{q|t} \right] \). 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 \( \beta_t \) is

\[
\text{var}(\beta_t \mid \mathcal{I}_t) = \sum_{i=0}^{t-1} \text{var}(\beta_{t-i} \mid \xi_{t+1:t} = 1, \mathcal{I}_t) \xi_{t+1:t|t}
\]

where

\[
\text{var}(\beta_{t-i} \mid \xi_{t+1:t} = 1, \mathcal{I}_t) = \frac{\eta_0 + i + 1}{\eta_0 + i - 1} \hat{\sigma}_{t-i}^2 \hat{V}_{t-i|t}.
\]

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 \( \beta_t \) conditional only on the observed data, I integrate \( \xi_t \) out of the joint density \( f(\beta_t, \xi_t \mid \mathcal{I}_t) \) to obtain
\[ f(\beta_t | \mathcal{Z}_t) = \sum_{i=0}^{t-1} g(\beta_t | \xi_{i+1,t} = 1, \mathcal{Z}_t) \xi_{i+1,t}, \]

where \( g(\beta_t | \xi_{it} = 1, \mathcal{Z}_t) \) denotes the density of the \( t \) distribution in (9). The distribution of \( \beta_t | \mathcal{Z}_t \) 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 \( \xi_{i+1,t} \approx 1 \) for some \( i \), then the distribution of \( \beta_t \) conditional on \( \mathcal{Z}_t \) approximates a \( t \) distribution (or a normal distribution in the case of constant \( \sigma^2 \)).

Conditional on the most recent break, \( \sigma_{t-i}^{-2} \) has the Gamma distribution

\[ \sigma_{t-i}^{-2} | \xi_{i+1,t} = 1, \mathcal{Z}_t \sim G(\hat{\sigma}_{t-i}, \eta_0 + i + 1), \quad (11) \]

for all \( i = 0, 1, \ldots, t-1 \) (see Koop 2003, pg. 37). I obtain the distribution of \( \sigma_{t-i}^{-2} \) conditional only on the observed data by integrating \( \xi_t \) out of the joint density \( f(\sigma_{t-i}^{-2}, \xi_t | \mathcal{Z}_t) \), which yields

\[ f(\sigma_{t-i}^{-2} | \mathcal{Z}_t) = \sum_{i=0}^{t-1} h(\sigma_{t-i}^{-2} | \xi_{i+1,t} = 1, \mathcal{Z}_t) \xi_{i+1,t}, \quad (12) \]

where \( h(\sigma_{t-i}^{-2} | \xi_{i+1,t} = 1, \mathcal{Z}_t) \) 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

\[ E(\sigma_{t-i}^{-2} | \xi_{i+1,t} = 1, \mathcal{Z}_t) = \frac{i + \eta_0 + i + 1}{\eta_0 + i + 1} \hat{\sigma}_{t-i}^{-2}. \quad (13) \]

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

\[ \sigma_{it}^{-2} = E(\sigma_{t-i}^{-2} | \mathcal{Z}_t) = \sum_{i=0}^{t-1} \frac{i + \eta_0 + i + 1}{\eta_0 + i + 1} \hat{\sigma}_{t-i}^{-2} \xi_{i+1,t} \equiv S_{it} \xi_{it}, \quad (14) \]

where \( S_{it} = \begin{bmatrix} \eta_0^{-1} \hat{\sigma}_{it}^{-2} & \cdots & \eta_0^{-1} \hat{\sigma}_{it}^{-2} \end{bmatrix} \). The filtered variance only exists if \( \eta_0 \geq 1 \). This condition requires that the moments of the marginal distribution provide enough information so that one observation is sufficient to identify \( \hat{\sigma}_t^{-2} = E(\sigma_t^{-2} | \xi_{it} = 1, \mathcal{Z}_t) \).
The filtered estimates in (10) and (14) use information up to the current period $t$. The Markov property of the state variable $\xi_t$ enables convenient forecasting of the coefficients and error variance. The forecasts are

$$\beta_{t+\ell t} = E(\beta_{t+\ell t} | \mathcal{F}_t) = B_{t+\ell t} \left( \prod_{j=1}^{\ell} P_{t+j} \right) \hat{\xi}_{t},$$

$$\sigma_{t+\ell t}^2 = E(\sigma_{t+\ell t}^2 | \mathcal{F}_t) = S_{t+\ell t} \left( \prod_{j=1}^{\ell} P_{t+j} \right) \hat{\xi}_{t},$$

where the conditional estimates are

$$S_{t+\ell t} = \left[ \begin{array}{cccc} \frac{\eta_0}{\eta_0-2} & \ldots & \frac{\eta_{-1}}{\eta_0-2} & \sigma_0^2 \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\eta_{-\ell}}{\eta_0-2} & \ldots & \frac{\eta_{-1}}{\eta_0-2} & \sigma_{-\ell}^2 \\ \end{array} \right]$$

and

$$B_{t+\ell t} = \left[ \begin{array}{cccc} \beta_0 & \ldots & \beta_0 & \hat{\beta}_t & \ldots & \hat{\beta}_t \end{array} \right].$$

Because post-break values of $\beta_t$ and $\sigma_t^2$ are drawn from a stationary distribution, the long run forecasts are $\lim_{\ell \to \infty} \beta_{t+\ell t} = \beta_0$ and $\lim_{\ell \to \infty} \sigma_{t+\ell t}^2 = \frac{\eta_0}{\eta_0-2} \sigma_0^2$.

### 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 $\beta_t$ and $\sigma_t^2$ 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

$$\beta_{\ell t} = \sum_{m=0}^{T-1} \sum_{j=0}^{\ell} \hat{\beta}_{jm} \pi_{jmT} + \sum_{j=0}^{\ell} \hat{\beta}_{\ell t} \bar{\pi}_{\ell tT},$$

where $\pi_{jmT} = \Pr(s_j = 1, n_j = m+1 | \mathcal{F}_T)$, $\bar{\pi}_{\ell tT} = \Pr(s_j = 1, n_j \geq T+1 | \mathcal{F}_T)$, 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 terms in (15) can be calculated directly from the smoothed state probabilities $\xi_{\ell t}$, which I obtain using the Markov-switching smoother

$$\xi_{\ell t} \equiv \xi_{t} \ast \left\{ P_{t+1}^{\ell} \left( \xi_{t+1} \ast \xi_{t+1} \right) \right\},$$

where $\ast$ denotes element-by-element division (Hamilton 1994).

The term $\pi_{jmT}$ can be written as

$$\pi_{jmT} = \Pr(s_j = 1, s_{j+1} = 0, \ldots, s_m = 0, s_{m+1} = 1 | \mathcal{F}_T)$$
Thus, the smoothed probability \( \pi_{jmT} \) equals the difference between the \((m-j+1)\)th element of \( \xi_{mT} \) and the \((m-j+2)\)th element of \( \xi_{m+1T} \). For the case of no breaks before the end of the sample,

\[
\hat{\pi}_{jT|T} = \Pr\left(s_j = 1, n_j \geq T + 1 \mid \mathcal{F}_T\right)
\]

\[
= \Pr\left(s_j = 1, s_{j+1} = 0, \ldots, s_T = 0 \mid \mathcal{F}_T\right)
\]

\[
= \xi_{T-j+1,T|T}.
\]

(17)

In sum, the smoothed estimates are

\[
\hat{\beta}_{iT|T} = \sum_{m=1}^{T-1} \sum_{j=0}^t \hat{\beta}_{jm}(\xi_{m-j+1,m|T} - \xi_{m-j+2,m+1|T}) + \sum_{j=0}^t \hat{\beta}_{jT}\xi_{T-j+1,T|T}.
\]

Similarly, for the error variance,

\[
\hat{\sigma}_{iT|T}^2 = \sum_{m=1}^{T-1} \sum_{j=0}^t \eta_0 + m - j + 1 \hat{\sigma}_{jm}^2 \left(\xi_{m-j+1,m|T} - \xi_{m-j+2,m+1|T}\right) + \sum_{j=0}^t \eta_0 + T - j + 1 \hat{\sigma}_{jT}^2 \xi_{T-j+1,T|T},
\]

where \( \hat{\sigma}_{jm}^2 \) is as defined in (8).

The distribution of \( \beta_t \) conditional on \( \mathcal{F}_T \) 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

\[
f\left(\beta_t \mid \mathcal{F}_T\right) = \sum_{m=1}^{T-1} \sum_{j=0}^t g\left(\beta_j \mid s_j = 1, n_j = m + 1, \mathcal{F}_T\right) \left(\xi_{m-j+1,m|T} - \xi_{m-j+2,m+1|T}\right)
\]

\[
+ \sum_{j=0}^t \eta_0 + T - j + 1 \hat{\sigma}_{jT}^2 \xi_{T-j+1,T|T}
\]

where \( g\left(\beta_j \mid s_j = 1, n_j = m + 1, \mathcal{F}_T\right) = g\left(\beta_m \mid \xi_{m-j+1,m} = 1, \mathcal{F}_m\right) \) denotes the density of the distribution in (9). It follows that \( \beta_t \) conditional on \( \mathcal{F}_T \) has approximately multivariate \( t \)-distribution 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, \( \beta_t \mid \mathcal{F}_T \) departs from a \( t \)-distribution. Similarly, the distribution of \( \sigma_t^2 \) conditional on \( \mathcal{F}_T \) 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

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

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

4. Practical Implementation: Truncating the State Space

The state variable $\xi_t$ 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

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

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

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

\[
\xi^k_{t+1} = \frac{f(y_t | \xi^k_t, x_t, \mathcal{F}_{t-1}) \ast \xi^k_{t+1}}{\sum_{j=1}^{k+1} f(y_t | \xi^k_j = 1, x_t, \mathcal{F}_{t-1}) \xi^k_{t+1,j-1}}
\]

where \( \xi^k_{t+1,j} = P \xi^k_t \). The term \( f(y_t | \xi^k_t, x_t, \mathcal{F}_{t-1}) \) is a \( k+1 \) dimensional vector denoting the conditional density of \( y_t \), where each element \( i \) corresponds to the event \( \xi^k_t = 1 \).

If a break occurs in the interval \([t-k+1, t] \), then one of the first \( k \) elements of \( \xi^k_t \) equals one and the predictive distribution for \( y_t \) 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 \( \xi^k_t \) equals one and the exact break date is unknown. In this case, forming the predictive distribution \( f(y_t | \xi^k_{k+1,t} = 1, x_t, \mathcal{F}_{t-1}) \) requires an approximation to the distribution of \( (\beta_{t-k}, \sigma_{t-k}^2) \) conditional on \( (\mathcal{F}_{t-1}, \xi^k_{k+1,t} = 1) \). I approximate this distribution using a simple forward recursion.

The event \( \xi^k_{k+1,t} = 1 \) 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 \( \beta_{t-k} \) conditional on \( (\xi^k_{k+1,t} = 1, \sigma_{t-k}^2, \mathcal{F}_{t-1}) \) 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 \( \sigma_{t-k}^2 \) conditional on \( (\xi^k_{k+1,t} = 1, \mathcal{F}_{t-1}) \) is a mixture of gammas. For large values of \( k \), the last \( k \) observations in \( \mathcal{F}_{t-1} \) dominate, and there is little difference between the components of these mixtures. It follows that, as \( k \) increases, the distribution of \( \beta_{t-k} \) conditional on \( (\xi^k_{k+1,t} = 1, \sigma_{t-k}^2, \mathcal{F}_{t-1}) \) approaches normality and the distribution of \( \sigma_{t-k}^2 \) conditional on \( (\xi^k_{k+1,t} = 1, \mathcal{F}_{t-1}) \) approaches gamma. I use this large \( k \) approximation to generate the recursion for \( \beta_{t-k}, \sigma_{t-k}^2 | \xi^k_{k+1,t} = 1, \mathcal{F}_{t-1} \) and in turn to approximate the density \( f(y_t | \xi^k_{k+1,t}, x_t, \mathcal{F}_{t-1}) \).
Specifically, define the quantities \( \beta_{t-k|y} \), \( \bar{V}_{t-k|y} \), \( \sigma^2_{t-k|y} \), and \( \bar{n}_{t-k|y} \) such that\(^2\)

\[
\beta_{t-k} | \xi^{k+1}_{k,t} = 1, \sigma^2_{t-k}, \mathbb{S}_{t-1} \sim N \left( \bar{\beta}_{t-k|y}, \sigma^2_{t-k|y} \bar{V}_{t-k|y} \right)
\]

\[
\sigma^2_{t-k} | \xi^{k+1}_{k,t} = 1, \mathbb{S}_{t-1} \sim G \left( \sigma^2_{t-k|y}, \bar{n}_{t-k|y} \right).
\]

where \( \approx \) denotes “approximately distributed as.” Combined with the normality of \( \varepsilon_t \), the approximations in (18) and (19) imply that

\[
y_t | x_t, \xi^{k+1}_{k,t} = 1, \mathbb{S}_{t-1} \approx t \left( x_t' \bar{\beta}_{t-k|y} - \sigma^2_{t-k|y} (1 + x_t' \bar{V}_{t-k|y} x_t), \bar{n}_{t-k|y} \right).
\]

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

\[
\bar{\beta}_{t-k|y} = \bar{\beta}_{t-k|y} + \bar{V}_{t-k|y} x_t (1 + x_t' \bar{V}_{t-k|y} x_t)^{-1} (y_t - x_t' \bar{\beta}_{t-k|y})
\]

and

\[
\bar{V}_{t-k|y} = \bar{V}_{t-k|y} - \bar{V}_{t-k|y} x_t (1 + x_t' \bar{V}_{t-k|y} x_t)^{-1} x_t' \bar{V}_{t-k|y}.
\]

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

\[
\bar{\sigma}^2_{t-k|y} = \bar{n}_{t-k|y - 1} \bar{\sigma}^2_{t-k|y} + \left( y_t - x_t' \bar{\beta}_{t-k|y} \right)^2 + \left( \bar{\beta}_{t-k|y} - \bar{\beta}_{t-k|y - 1} \right)' \bar{V}_{t-k|y - 1}^{-1} \left( \bar{\beta}_{t-k|y} - \bar{\beta}_{t-k|y - 1} \right)
\]

\[
\bar{n}_{t-k|y} = \bar{n}_{t-k|y - 1} + 1.
\]

Equations (20)-(22) show how to obtain \( \bar{\beta}_{t-k|y} \), \( \bar{V}_{t-k|y} \), \( \bar{\sigma}^2_{t-k|y} \), and \( \bar{n}_{t-k|y} \) from \( \bar{\beta}_{t-k|y - 1} \), \( \bar{V}_{t-k|y - 1} \), \( \bar{\sigma}^2_{t-k|y - 1} \), and \( \bar{n}_{t-k|y - 1} \). For period \( t-k+1 \), I obtain \( \bar{\beta}_{t-k+1|y} \), \( \bar{V}_{t-k+1|y} \), \( \bar{\sigma}^2_{t-k+1|y} \), and \( \bar{n}_{t-k+1|y} \) from \( \bar{\beta}_{t-k|y} \), \( \bar{V}_{t-k|y} \), \( \bar{\sigma}^2_{t-k|y} \), and \( \bar{n}_{t-k|y} \) by accounting for the possibility of a break in period \( t-k+1 \). Specifically, I calculate the parameters of the approximate distribution of \( \beta_{t-k+1}, \sigma^2_{t-k+1} | \xi^{k+1}_{k+1,t} = 1, \mathbb{S}_{t-1} \) as a

\[\text{Here, and in the remainder of the paper, I use a bar notation to indicate that the conditioning set includes no breaks between the period of interest (period } t-k \text{ in this case) and the end of the information set (period } t-1 \text{ in this case).}\]
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 \( \bar{\beta}_{t-k+1|t} \), we have

\[
\bar{\beta}_{t-k+1|t} = \bar{s}_{t-k+1|t} E(\beta_{t-k+1} \mid \xi_{t|t} = 1, \sigma^2_{t-k+1}, \mathcal{F}_t) + (1 - \bar{s}_{t-k+1|t}) \bar{\beta}_{t-k|t},
\]

where

\[
\bar{s}_{t-k+1|t} = \Pr(s_{t-k+1} = 1 \mid s_t = s_{t-1} = \ldots = s_{t-k+2} = 0, \mathcal{F}_t)
\]

\[
= \frac{\Pr(s_t = s_{t-1} = \ldots = s_{t-k+2} = 0, s_{t-k+1} = 1 \mid \mathcal{F}_t)}{\Pr(s_t = s_{t-1} = \ldots = s_{t-k+2} = 0 \mid \mathcal{F}_t)}
\]

\[
= \frac{\xi^k_{t,\mathcal{F}_t}}{\xi^k_{t,k,\mathcal{F}_t} + \xi^k_{t+1,k,\mathcal{F}_t}}.
\]

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

Using the truncated state space, the filtered estimates of \( \beta \) and \( \sigma^2 \) are given by \( \beta_{t|t} = B_{t|t} \xi_{t|t} \) and \( \sigma^2_{t|t} = S_{t|t} \xi_{t|t} \) respectively, where \( S_{t|t} = \begin{bmatrix} \eta_0 + k \sigma^2_{t-k+1} & \cdots & \frac{\eta_0 + k}{\eta_{k-1} + k} \sigma^2_{t-k+1} & \frac{\eta_0 + k}{\eta_{k-1} + k} \sigma^2_{t-k} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\eta_0 + k}{\eta_{k-1} + k} \sigma^2_{t-k+1} & \cdots & \eta_{k-1} + k \sigma^2_{t-k} & \eta_{k-1} + k \sigma^2_k \\ \end{bmatrix} \) and \( B_{t|t} = \begin{bmatrix} \hat{\beta}_{t|t} & \cdots & \hat{\beta}_{t-k+1|t} & \hat{\beta}_{t-k|t} \end{bmatrix} \). In the Appendix, I summarize this filtering algorithm and present the algorithm for generating smoothed estimates of \( \beta \) and \( \sigma^2 \) using the truncated state space.

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

\[
\delta(k) = T^{-1} E \left( L(\theta) - L_k(\theta) \right) = T^{-1} \sum_{t=1}^{T} E \left( \log \frac{f(y_t \mid \xi_{t|t}, x_t, \mathcal{F}_{t-1}) \xi_{t|t}^y}{f(y_t \mid \xi_{t|t}^y, x_t, \mathcal{F}_{t-1}) \xi_{t|t}^y} \right) > 0.
\]

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:

\[
y_t = \beta_{1t} + x_t \beta_{2t} + \epsilon_t
\]

\[
x_t \sim iidN(0,1), \ \epsilon_t \sim iidN(0, \sigma^2_t)
\]
\[ \beta_t \sim N(1, v_0), \quad \beta_{2t} \sim N(2, v_0), \quad v_0 \in \{0.04, 0.25, 1\} \]
\[ \sigma_t^{-2} \sim G(1, \eta_0), \quad \eta_0 \in \{20, 10, 5\} \]
\[ p_{00} \in \{0.995, 0.95\}, p_{11} = 1 - p_{00}. \]

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.,

\[
\frac{E(\beta'_i x_i' \beta_i)}{E(\beta'_i x_i' \beta_i | \beta_i = \beta_0)} = \frac{E(\beta'_i \beta_i)}{\beta_0' \beta_0} = \frac{\beta_0' \beta_0 + 2v_0}{\beta_0' \beta_0} = 1 + 0.4v_0,
\]

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 \( \eta_0 = 20, 10, \) and \( 5 \), the variance of \( \sigma_t^{-2} \) 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 \( \eta_0 = 5 \) (frequent large breaks), then setting \( k = 20 \) ensures zero information loss. For the case with rare small breaks \( (p_{00} = 0.995, p_{11} = 0.005, v_0 = 0.04, \eta_0 = 20), k = 350 \) 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 \( (p_{00} = 0.99, p_{11} = 0.01, v_0 = 1, \eta_0 = 5) \). 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 \( (\beta_{20}) \). 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

$$y_t = x_t' \beta_s + \sigma_s \epsilon_t,$$

where $\epsilon_t \sim N(0,1)$, $\beta_s \in \{\beta_1, \ldots, \beta_N\}$, $\sigma_s \in \{\sigma_1, \ldots, \sigma_N\}$, 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.

If previously observed states can recur as in Hamilton’s (1989) model, then the transition probability matrix $Q$ can be full. Such a Markov switching model with recurring states (MSR) shares with the MB model the property of stationarity, assuming ergodicity of the Markov chain. However, the MSR model is restricted by the number of possible values the coefficients and error variance can take. In the MB model a break causes a new draw from a continuous distribution,
whereas in the MSR model a break generates a draw from a discrete distribution. Thus, the MSR model only forecasts well after breaks when the process reverts to a previously observed state. Moreover, as the number of states increases, the MSR model loses parsimony because it requires estimation of $N(N-1)$ transition probability parameters as well as $N$ different coefficient vectors and error variance values.

5.2 Time-Varying Parameter

Unlike Markov switching models, the time-varying parameter (TVP), or stochastic coefficient, model of Cooley and Prescott (1973) incorporates a continuous state space. Under this model

$$y_t = x_t'\beta_t + \sigma_t \epsilon_t,$$

where $\beta_t = A\beta_{t-1} + \omega_t$, $\epsilon_t \sim N(0,1)$, $\nu_t \sim N(0,1)$, and $E(\epsilon_t, \nu_t) = 0$ for all $t, s$. Cooley and Prescott (1973) specify a constant error variance but more recent applications also allow time-varying volatility through either a stochastic volatility or a GARCH model. Time-varying parameter models have been applied in a wide variety of settings (see Cogley and Sargent 2005, Stock and Watson 1996, and the references therein). In many applications, the matrix $A$ is set to an identity.

A stochastic volatility model for $\sigma_t$ specifies the autoregressive process

$$\ln \sigma_t = \rho \ln \sigma_{t-1} + \theta u_t,$$

where $u_t \sim N(0,1)$ and $\rho$ is sometimes set to one. Calculating the likelihood function for this model requires solving a nonlinear filtering problem for which no closed form solution exists. Consequently, applied users estimate the model using simulation methods such as Markov Chain Monte Carlo (Kim, Shephard and Chib 1998). Alternatively, a GARCH(1,1) model specifies

$$\sigma_t = \omega_0 + \omega_1 \sigma_{t-1} + \omega_2 \left( (y_{t-1} - x_{t-1}'\beta_{t-1})^2 | y_{t-1}, y_{t-2}, ..., x_{t-1}, x_{t-2}, ... \right)$$
Because this model specifies $\sigma_t$ to be measurable with respect to the history of $y_t$ and $x_t$, the Kalman filter can be applied directly to calculate the likelihood (Chou, Engle, and Kane 1992). Moreover, the conditional moment $E((y_{t-1} - x'_{t-1}\beta_{t-1})^2 | y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots)$ can be obtained directly from the Kalman filter. In an application to daily exchange rates Kim, Shephard, and Chib (1998) show that the two volatility models fit the data equally well.

In contrast to the MB and Markov-switching models, the family of TVP models imposes a smooth evolution process on the coefficients and error variance. Coupled with the continuous state space, this constraint allows the model to remain parsimonious even when the coefficients and error variances shift frequently. The MB model also retains parsimony when the process shifts frequently. However, if the process exhibits only occasional shifts, then the TVP estimates will be too volatile between breaks and may not react quickly when breaks occur.

5.3 Stochastic Permanent Breaks and Innovation Regime Switching

Consider the intercept-only MB 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 $z_t \sim N(\beta_0, \sigma^2 V_0)$ is independent of $\varepsilon_t$. 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 $\beta_t$ is observable. On the other hand, the MB model contains transitory breaks and a latent post-break value of $\beta_t$.

Setting $k=1$ yields a MB(1) model with similar properties to the basic STOPBREAK model. Using $k=1$, we have from (10)
\[
\beta_{it} = s_{it} \hat{\beta}_{it} + (1-s_{it}) \bar{\beta}_{t-1|t}
\]

where
\[
\hat{\beta}_{it} = (1+V_{0}^{-1})^{-1}(y_{t} + \beta_{0}V_{0}^{-1}) = y_{t} - \frac{(y_{t} - \beta_{0})V_{0}^{-1}}{1+V_{0}^{-1}},
\]

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

\[
\beta_{it} = s_{it} \left( y_{t} - \frac{(y_{t} - \beta_{0})V_{0}^{-1}}{1+V_{0}^{-1}} \right) + (1-s_{it}) \left( \beta_{t-1|t-1} + g_{t-1}(y_{t} - \beta_{t-1|t-1}) \right)
\]

In contrast, Engle and Smith (1999) specify the model \( \beta_{it} = q_{t}y_{t} + (1-q_{t})\beta_{t-1|t-1} \), where \( q_{t} \) is a \( \mathcal{F}_{t} \)-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 \( \beta_{t} \) using the Kalman filter. This feature increases precision during stable periods. Second, the MB model uses prior information on the marginal distribution of \( \beta_{t} \) to modify the estimate of \( \beta_{t} \) in the event of a break. This tie to \( \beta_{0} \) generates mean reversion in the process because the new draws of \( \beta_{t} \) 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 \( \beta_{t} \). Like the STOPBREAK model, the MB(1) model would not exhibit long-run mean reversion in that case.

6. Output Growth and the Yield Curve

A flattening yield curve tends to predict slower macroeconomic growth up to six quarters into the future (see, for example, Stock and Watson 1989). However, this predictive relationship is unstable. Giacomini and Rossi (2006) show evidence of forecast breakdowns in the predictive ability of the yield spread for GDP growth, and Estrella, Rodrigues, and Schich (2003) show evidence of a break in the predictive ability of the yield spread for industrial production growth.
In this section, I use the MB model to predict GDP growth up to six quarters ahead. In doing so, I assess the stability of the predictive relationship, and I compare predictive ability across several models.

6.1 Model Specification

The model is

\[ y_t = \beta_0 + \beta_1 x_t + \epsilon_t, \quad t = 1, 2, \ldots, T \]  

where \( \epsilon_t \sim iidN(0, \sigma^2) \), \( x_t \) denotes the 10-year Treasury bond yield minus the 3-month Treasury bill yield, and \( y_t = 400 \times \ln(GDP_t / GDP_{t-1}) \) denotes the annualized quarterly change in the logarithm of seasonally-adjusted real GDP. For the period 1967:Q1 through 2006:Q4, Table 1 shows that OLS estimation of (27) generates a maximum \( R^2 \) of 0.16 at the two-quarter horizon. At horizons greater than two quarters, the \( R^2 \) steadily decreases, falling to 0.03 at the six-quarter horizon.

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; (iii) Nyblom’s (1989) test, which is locally optimal against a martingale \( \beta_t \); and (iv) Bai and Perron’s (1998) sequential \( F \)-test, which estimates the number of breaks. I report these statistics in Table 1. All tests find evidence of breaks at the two-quarter horizon, and some tests indicate breaks at the one and three-quarter horizons. There is little evidence of breaks in the relationship at horizons four quarters or longer, for which the predictive ability of the yield spread is weak. Bai and Perron’s sequential procedure estimates a single break in 1984 for the two- and three-quarter horizons, and three different break dates for the one-quarter horizon model.

Table 2 shows estimates of the parameters of an MB(20) model for all 6 horizons. The two-quarter horizon model exhibits the largest likelihood value, and Figure 3 plots filtered and
smoothed coefficients, error variance, and break probabilities for this case. The dramatic drop in $\beta_2t$ and $\sigma_t$ in 1984 stands out. This period marks the beginning of the so-called “great moderation” (see, for example, McConnell and Perez-Quiros 2000). Thus, in addition to a drop in volatility, the great moderation coincided with a drop in the predictive content of the yield spread for output. This result matches the finding of Atkeson and Ohanian (2001) that the great moderation coincided with a drop in the ability of the output gap to predict inflation. Figure 3 also shows that the great moderation returned $\beta_2t$ and $\sigma_t$ to their pre-1973 levels.

Figure 3 also shows that the yield spread did not predict the late 90s boom or the following recession well. The yield spread dropped steadily from 1993 until 2001, when reductions in the federal funds rate target steepened the yield curve. However, GDP growth did not drop steadily between 1993 and 2001. Rather, GDP growth averaged 4.4 percent per annum from 1995-99, before dropping to 2.2 percent in 2000 and 0.2 percent in 2001. To compensate for the lack of predictability from the yield spread during this period, Figure 3 shows that the intercept jumped from 1.5 to about 3.1 in 1995 before dropping back to 1.5 just as quickly in 2000.

The smoothed probabilities in Figure 3 reveal the large break in 1984 associated with the great moderation. The smoothed probability that a break occurred in 1984:Q3 is 0.53, and the estimated probability that a break occurred some time in the year from 1984:Q2 to 1985:Q1 equals 0.93. I obtain this latter probability estimate easily by summing the first four elements of the smoothed state variable $\xi_{tT}$ for $t=1985:Q1$. The filtered probabilities in Figure 3 show few spikes, which is not surprising because these probabilities are based on a comparison of a single observation to the predictive density. However, the MB model can learn about breaks quickly. For example, the estimated probability of a break in 1984:Q3 rises from the unconditional estimate of 0.05 to 0.09 upon realization of the 1984:Q3 data. After one more quarter it rises to 0.18; it rises to 0.27 after two quarters and to 0.37 after three quarters. This quick reaction is generated by the fact that GDP grew at an annual rate of 7.3% in the first half of 1984 and 3.6%
in the second half of the year, whereas the lagged yield spread changed little during 1983 and 1984. As a result, the filtered yield curve coefficient dropped from 1.53 in 1984:Q3 to 0.76 just four quarters later.

The largest filtered probability estimate is 0.23 in 2003:Q3. This observation is associated with a moderate jump in the intercept $\beta_1$. This relatively small break generates a large break probability because volatility is low at this point in the sample. In contrast, initial volatility is much larger in 1984, so a much larger break is needed before the model will move.

The log likelihood values for the MB model exceed their no-break counterparts by 15-20 at each horizon. Although these likelihood differences appear very large, they should be interpreted with some care because the LR test of the no break hypothesis has a nonstandard null distribution (see Remark 2). The likelihood improvement generated by the MB model emanates from the decline in the error variance associated with the great moderation.

Based on the parameter estimates in Table 2, the estimated mean of the intercept term ($\beta_{10}$) is about 2.1 and the estimated mean of the slope term ($\beta_{20}$) is about 0.5 for all horizons. However, the associated variance terms ($V_{\beta_1}$ and $V_{\beta_2}$) decline towards zero as the horizon increases. These variance terms equal zero in a model with no breaks in the coefficients. To test for breaks in $\beta$, I conduct a likelihood ratio test of the null hypothesis $V_{\beta_1} = V_{\beta_2} = 0$ using the critical values in Remark 7, which are valid for this test as long as $\sigma_t$ is not constant. I reject the null hypothesis at 5 percent for only the two-horizon model, although I also reject at 10 percent for the one-horizon model. This result mirrors the break tests in Table 1, which found little evidence of breaks at the longer horizons.

The estimated transition probabilities show that breaks are rarer at longer horizons. They also show no evidence of Markov dependence in the breaks because the hypothesis that $p_{11} = 1 - p_{00}$ cannot be rejected at any horizon. The estimated unconditional break probability equals 0.09 for the one-quarter horizon, 0.05 at two and three quarters and 0.01 beyond three quarters. This
pattern derives from the presence of breaks in the coefficients $\beta_1$ and $\beta_2$ at short horizons, in addition to the great moderation in $\sigma$. Longer horizons display little evidence of breaks in $\beta_1$ and $\beta_2$, so the only break is the change in $\sigma$ associated with the great moderation.

6.2 Forecasting Comparison

In Table 3, I present an out-of-sample forecasting comparison of the MB model to several alternative breaks models for the two-quarter horizon model. I estimate (27) using data up to 1986:Q4, and use the data from 1987–2006 to evaluate post-sample forecasting performance. I compare the MB model to Markov switching models with recurring (MSR) and nonrecurring (MSNR) states, a time-varying parameter GARCH(1,1) model, and 10-year rolling OLS regressions. Table 3 shows the estimated Kullback-Leibler information loss from applying these alternative models rather than an MB(20) 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 (MSFE) of the alternative models relative to the MB(20) model for the out-of-sample period.

I do not re-estimate the parameters of the models as I iterate through the forecast period, with the exception of the MSNR model and the rolling OLS regressions. The MSNR model has no capacity to predict post-break values of $\beta_i$ and $\sigma$, so I re-estimate the parameters of this model every quarter, and I select the number of states using the Markov switching criterion (MSC) of Smith, Naik and Tsai (2006). I also use MSC to select the number of states in the MSR model. For all models, the forecasts should be interpreted as one-quarter ahead predictions because, when forecasting period $t+1$, the filtering algorithms use information up to period $t$ to infer the coefficient values even though the explanatory variable is measured at period $t-1$.

Table 3 shows that the MB model appears worse than the Markov switching models in-sample, but performs significantly better out of sample. The predictive likelihood values for the MB model exceed their Markov switching counterparts by 7.0 for the MSNR model and 21.2 for
the MSR model. Similarly, the MSNR model has a 12 percent worse MSFE and the MSR model has a 59 percent worse MSFE than the MB model. The MB model also performs significantly better than the TVP-GARCH model, both in terms of predictive likelihood and MSFE. Rolling OLS regressions get closest to the performance of the MB model, with a predictive log likelihood 1.6 worse and a MSFE 2 percent worse. Overall, the MB model outperforms the other models out of sample, although the difference is not statistically significant at 5 percent for the rolling OLS and MSNR models.

I perform the forecasting exercise without re-estimation to highlight the ability of the models to adapt to breaks, which is an ability that each model is designed to possess. Nonetheless, recursive re-estimation of MB model increases its performance advantage. Its RMSE decreases from 4.55 to 4.06, which is 25 percent better than recursively estimated MSR and MSNR models, 14 percent better than the rolling OLS estimates, and 8 percent better than recursively estimated TVP model. Recursive estimation produces little improvement in the predictive log likelihood of the MB model, but it remains substantially better than the recursively estimated MSR and MSNR models.

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 in an application to the predictive ability of the yield curve for GDP growth.
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 can perform 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:

\[ y_t = x_t' \beta_t + \varepsilon_t, \quad t = 1, 2, \ldots, T \]

where \( \varepsilon_t \mid (x_t, \beta_t, \sigma_t) \sim N(0, \sigma_t^2) \), \( (\beta_t, \sigma_t \mid s_t = 1) \perp (x_t, \mathcal{I}_{t-1}) \), \( \beta_t \mid \sigma_t \sim N(\beta_0, \sigma_t^2 \nu_0) \), 
\( \sigma_t^{-2} \sim G(\sigma_0^{-2}, \eta_0) \), and \( \mathcal{I}_{t-1} \) denotes the information in \( (y_1, y_2, \ldots, y_{t-1}, x_1, x_2, \ldots, x_{t-1}) \). Define the Markov random variable \( s_t \in \{0,1\} \) such that \( s_t = 1 \) implies a new draw of \( (\beta_t, \sigma_t) \) in period \( t \), and \( s_t = 0 \) implies \( \beta_t = \beta_{t-1} \) and \( \sigma_t = \sigma_{t-1} \).

Filtered state probabilities:

\[
\hat{\pi}_{t+1|t} = \begin{cases} P_t \hat{\pi}_{t|t} & t \leq k + 1 \\ P_t \hat{\pi}_{t|t} & t > k + 1 \end{cases},
\]

where

\[
P_t = \begin{bmatrix} P_{11} & P_{01} & \cdots & P_{01} \\ P_{10} & 0 & \cdots & 0 \\ 0 & P_{00} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_{00} \end{bmatrix},
\]

\[ p_{tm} = \Pr(s_t = m \mid s_{t-1} = l), \quad \hat{\pi}_{t+1|t}^k = 1, \text{ and } \ast \text{ denotes element-by-element multiplication. The matrix } P_t \text{ has dimension } t \times (t-1), \text{ and } \Gamma \text{ has dimension } (k+1) \times (k+1). \]

Conditional density for \( i=0,1,\ldots,k-1 \):

\[
f(y_t \mid x_t, \hat{\pi}_{t+1|t}^k = 1, \mathcal{I}_{t-1}) = \frac{\Gamma(1 + i + \eta_0) / 2}{\Gamma(i + \eta_0) / 2} \frac{(y_t - x_t' \hat{\beta}_{t-i-1})^2}{\left(1 + x_t' \hat{\beta}_{t-i-1} + \sum_j x_{t-j} x'_t (i + \eta_0) \right)^{(1+i+\eta_0)/2}},
\]

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

\[
\hat{V}_{t-i-1} = \left( V_0^{-1} + \sum_j x_{t-j} x'_t \right)^{-1},
\]
\[ \hat{\sigma}^2_{t-i\bar{y}-1} = \frac{\eta_0 \sigma_0^2 + \sum_{j=1}^{i} (y_{t-j} - x'_{t-j} \hat{\beta}_{t-i\bar{y}-1})^2 + (\hat{\beta}_{t-i\bar{y}-1} - \beta_0)' V_0^{-1} (\hat{\beta}_{t-i\bar{y}-1} - \beta_0)}{\eta_0 + i} , \]

where \( \xi_{it} \) denotes the \( i^{th} \) element of \( \xi \) and \( \Gamma() \) denotes the gamma function.

Approximate conditional density for \( i=k \):

\[
f(y_t \mid x_t, \xi_{k+1}, = 1, \mathcal{F}_{t-1}) = \frac{\Gamma \left( 1 + i + \bar{\eta}_{t-i\bar{y}-1} \right) \left( 1 + \frac{(y_t - x'_t \bar{\beta}_{t-i\bar{y}-1})^2}{\bar{\sigma}^2_{t-i\bar{y}-1}(1 + x'_t \bar{V}_{t-i\bar{y}-1} x_t)(i + \bar{\eta}_{t-i\bar{y}-1})} \right)^{-(1 + i + \bar{\eta}_{t-i\bar{y}-1})/2} \Gamma \left( (i + \bar{\eta}_{t-i\bar{y}-1})/2 \right) \left( \pi \sigma^2_{t-i\bar{y}-1} \right)^{1/2}}{\Gamma \left( i + \bar{\eta}_{t-i\bar{y}-1} \right)/2 \left( 1 + x'_t \bar{V}_{t-i\bar{y}-1} x_t \right)}
\]

\[
\bar{\beta}_{t-k\bar{y}} = \bar{s}_{t-k\bar{y}+1} \hat{\beta}_{t-k\bar{y}} + (1 - \bar{s}_{t-k\bar{y}+1}) \left( \bar{\beta}_{t-k\bar{y}} + \bar{V}_{t-k\bar{y}-1} x_t (1 + x'_t \bar{V}_{t-k\bar{y}-1} x_t)^{-1} \right) (y_t - x'_t \bar{\beta}_{t-k\bar{y}-1})
\]

\[
\bar{V}_{t-k\bar{y}} = \bar{s}_{t-k\bar{y}+1} \bar{V}_{t-k\bar{y}} + (1 - \bar{s}_{t-k\bar{y}+1}) \left( \bar{V}_{t-k\bar{y}-1} - \bar{V}_{t-k\bar{y}-1} x_t (1 + x'_t \bar{V}_{t-k\bar{y}-1} x_t)^{-1} \right) x'_t \bar{V}_{t-k\bar{y}-1}
\]

\[
\bar{\sigma}^{-2}_{t-k\bar{y}} = \bar{s}_{t-k\bar{y}+1} \bar{\sigma}^{-2}_{t-k\bar{y}} + (1 - \bar{s}_{t-k\bar{y}+1}) \left( \bar{\eta}_{t-k\bar{y}-1} \sigma^2_{t-k\bar{y}-1} + (y_t - x'_t \bar{\beta}_{t-k\bar{y}})^2 + (\bar{\beta}_{t-k\bar{y}} - \bar{\beta}_{t-k\bar{y}-1})' \bar{V}_{t-k\bar{y}-1}^{-1} (\bar{\beta}_{t-k\bar{y}} - \bar{\beta}_{t-k\bar{y}-1}) \right)^{-2} \bar{\eta}_{t-k\bar{y}-1} + 1
\]

\[
\bar{s}_{t-k\bar{y}+1} = \frac{\xi_{k,\bar{y}t}}{\xi_{k,\bar{y}t} + \xi_{k+1,\bar{y}t}}
\]

Filtered coefficient estimates:

\[
\hat{\beta}_{t\bar{y}} = \sum_{i=0}^{k-1} \hat{\beta}_{t-i\bar{y}} \xi_{i+1,\bar{y}t} + \hat{\beta}_{t-k\bar{y}} \xi_{k+1,\bar{y}t} \equiv B^{k}_{t\bar{y}} \xi_{k,\bar{y}t}
\]

where \( B^{k}_{t\bar{y}} \equiv \begin{bmatrix} \hat{\beta}_{t\bar{y}} & \cdots & \hat{\beta}_{t-k\bar{y}} & \bar{\beta}_{t-k\bar{y}} \end{bmatrix} \).

Filtered error variance estimates:

\[
\bar{\sigma}^2_{t\bar{y}} = E \left( \sigma^2_{t\bar{y}} \mid \mathcal{F}_t \right) = \sum_{i=0}^{k-1} \frac{\eta_0 + i + 1}{\eta_0 + i - 1} \bar{\sigma}^2_{t-i\bar{y}} \xi_{i+1,\bar{y}t} + \frac{\bar{\eta}_{t-k\bar{y}}}{\bar{\eta}_{t-k\bar{y}} - 2} \bar{\sigma}^2_{t-k\bar{y}} \xi_{k+1,\bar{y}t} \equiv S^{k}_{t\bar{y}} \xi_{k,\bar{y}t}
\]

where \( S^{k}_{t\bar{y}} \equiv \begin{bmatrix} \frac{\eta_0 + k}{\eta_0} \bar{\sigma}^2_{t\bar{y}} & \cdots & \frac{\eta_0 + k-2}{\eta_0 + k-2} \bar{\sigma}^2_{t-k\bar{y}} & \frac{\eta_0 + k-1}{\eta_0 + k-1} \bar{\sigma}^2_{t-k\bar{y}} \end{bmatrix} \).
Smoothed state probabilities:
\[ \xi_{t \mid T}^{k} = \xi_{t}^{k} \ast \{ P_{t+1}^{k} \left( \xi_{t+1 \mid T}^{k} + \xi_{t+1}^{k} \right) \} \]
where \( \ast \) denotes element-by-element division.

Smoothed coefficient estimates:
\[
E \left( \beta_{t} \mid \mathcal{F}_{T} \right) = E \left( \beta_{t} \mid n_{t} = t+1, \mathcal{F}_{T} \right) \Pr \left( n_{t} = t+1 \mid \mathcal{F}_{T} \right) + E \left( \beta_{t} \mid n_{t} > t+1, \mathcal{F}_{T} \right) \Pr \left( n_{t} > t+1 \mid \mathcal{F}_{T} \right)
\]
\[
= \sum_{m=t}^{t+k-2} E \left( \beta_{t} \mid n_{t} = m+1, \mathcal{F}_{m} \right) \Pr \left( n_{t} = m+1 \mid \mathcal{F}_{T} \right) + E \left( \beta_{t} \mid n_{t} \geq t+k, \mathcal{F}_{T} \right) \Pr \left( n_{t} \geq t+k \mid \mathcal{F}_{T} \right)
\]
\[
= \sum_{i=0}^{t+k-2} \sum_{m=t}^{t+k-2} E \left( \beta_{t} \mid s_{i} = i, n_{t} = m+1, \mathcal{F}_{m} \right) \Pr \left( s_{i} = i, n_{t} = m+1 \mid \mathcal{F}_{T} \right)
\]
\[+ \sum_{i=0}^{t+k-2} E \left( \beta_{t} \mid s_{i} = i, n_{t} \geq t+k, \mathcal{F}_{T} \right) \Pr \left( s_{i} = i, n_{t} \geq t+k \mid \mathcal{F}_{T} \right)
\]
where \( n_{t} \) denotes the date of the next break after period \( t \). The second equality uses the fact that \( E \left( \beta_{t} \mid s_{i} = i, n_{t} = m+1, \mathcal{F}_{m} \right) = E \left( \beta_{t} \mid s_{i} = i, n_{t} = m+1, \mathcal{F}_{m} \right) \) because of the break in period \( m+1 \). I approximate \( E \left( \beta_{t} \mid s_{i} = i, n_{t} \geq t+k, \mathcal{F}_{T} \right) \) by \( E \left( \beta_{t} \mid s_{i} = i, n_{t} \geq t+k, \mathcal{F}_{T+k-1} \right) \), which implies that the smoothed coefficient estimates are:
\[
\beta_{t \mid T} = \hat{B}_{t} \Pi_{t \mid T} + \bar{B}_{t} \Lambda_{t \mid T}
\]
where
\[
\hat{B}_{t} = \begin{bmatrix} \hat{\beta}_{t} & \hat{\beta}_{t+1} & \ldots & \hat{\beta}_{t+k-1} \end{bmatrix}, \quad \Pi_{t \mid T} = \begin{bmatrix} \pi_{t \mid T} & \pi_{t, t+k-2} & \ldots & \pi_{t, t+k-3} \end{bmatrix},
\]
\[
\hat{\beta}_{t \mid m} = E \left( \beta_{t} \mid s_{i} = 1, n_{t} = m+1, \mathcal{F}_{m} \right), \quad \pi_{t \mid m} = \Pr \left( s_{i} = 1, n_{t} = m+1 \mid \mathcal{F}_{T} \right),
\]
\[
\bar{B}_{t} = \begin{bmatrix} \bar{\beta}_{t} & \bar{\beta}_{t+1} & \ldots & \bar{\beta}_{t+k-1} \end{bmatrix}, \quad \Lambda_{t \mid T} = \begin{bmatrix} \lambda_{t \mid T} & \lambda_{t, t+k-2} & \ldots & \lambda_{t, t+k-3} \end{bmatrix},
\]
\[
\bar{\beta}_{t \mid m} = E \left( \beta_{t} \mid s_{i} = 0, n_{t} = m+1, \mathcal{F}_{m} \right), \quad \lambda_{t \mid m} = \Pr \left( s_{i} = 0, n_{t} = m+1 \mid \mathcal{F}_{T} \right).
\]

Notes:
(i) This smoother averages over various estimates of \( \beta \) conditional on whether a break occurred in period \( t \) and on when the next break occurs.
(ii) The dimension of the state variable \( \xi_{t}^{k} \) dictates the dimension of \( \hat{B}_{t} \), \( \bar{B}_{t} \), \( \Pi_{t} \), and \( \Lambda_{t} \).
(iii) Near the end of the sample, there are fewer than \( k-1 \) observations after \( t \), which reduces the dimension of \( \hat{B}_{t} \), \( \bar{B}_{t} \), \( \Pi_{t} \), and \( \Lambda_{t} \). Specifically, for \( t > T-k+1 \),
\[
\hat{B}_{t} = \begin{bmatrix} \hat{\beta}_{t} \ldots \hat{\beta}_{t \mid T} \end{bmatrix}, \quad \bar{B}_{t} = \begin{bmatrix} \bar{\beta}_{t} \ldots \bar{\beta}_{t \mid T} \end{bmatrix}, \quad \Pi_{t \mid T} = \begin{bmatrix} \pi_{t \mid T} \ldots \pi_{t, t-1 \mid T} \end{bmatrix}, \quad \Lambda_{t \mid T} = \begin{bmatrix} \lambda_{t \mid T} \ldots \lambda_{t, t-1 \mid T} \end{bmatrix}.
\]
Conditional coefficient estimates for smoothing:

\[ \hat{\beta}_{t,m} = \left( V_0^{-1} + \sum_{j=m}^{m} x_j' x_j \right)^{-1} \left( V_0^{-1} \beta_0 + \sum_{j=t}^{m} x_j' y_j \right) \]

**Forward recursions for elements of \( \bar{B} \):**

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

\[ \bar{\beta}_{t,m} = \frac{\hat{\beta}_{t-1,m|T} + \bar{\beta}_{t-1,m|T} \lambda_{t-1,m|T}}{\pi_{t-1,m|T} + \lambda_2^{-1,m|T}} \]

For \( \bar{\beta}_{t+k-1} \), 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 \( \bar{\beta}_{t,m} = \hat{\beta}_{t,m} \).

Conditional probabilities for smoothing:

From (16) and (17),

\[ \pi_{m|T} = \frac{\xi_m^k}{\pi_m^k + k_m^k} \]

\[ \hat{\pi}_{m|T} = \frac{\xi_m^k}{\pi_m^k + k_m^k} \]

Similarly,

\[ \lambda_{m|T} = \Pr(s_i = 0, s_{i+1} = 0, \ldots, s_m = 0, s_{m+1} = 1 | \mathcal{F}_T) \]

\[ = \Pr(s_i = 0, s_{i+1} = 0, \ldots, s_m = 0 | \mathcal{F}_T) - \Pr(s_i = 0, s_{i+1} = 0, \ldots, s_m = 0, s_{m+1} = 0 | \mathcal{F}_T) \]

\[ = \Pr(s_{i-1} = 1, s_i = 0, \ldots, s_m = 0 | \mathcal{F}_T) + \Pr(s_{i-1} = 0, s_i = 0, \ldots, s_m = 0 | \mathcal{F}_T) \]

\[ - \Pr(s_{i-1} = 1, s_i = 0, \ldots, s_{m+1} = 0 | \mathcal{F}_T) - \Pr(s_{i-1} = 0, s_i = 0, \ldots, s_{m+1} = 0 | \mathcal{F}_T) \]

\[ = \sum_{j=m-k+1}^{t-1} \Pr(s_j = 1, s_{j+1} = 0, \ldots, s_m = 0 | \mathcal{F}_T) + \Pr(s_{m-k+1} = 0, s_{m-k+2} = 0, \ldots, s_m = 0 | \mathcal{F}_T) \]

\[ - \sum_{j=m-k+2}^{t-1} \Pr(s_j = 1, s_{j+1} = 0, \ldots, s_{m+1} = 0 | \mathcal{F}_T) - \Pr(s_{m-k+2} = 0, s_{m-k+3} = 0, \ldots, s_{m+1} = 0 | \mathcal{F}_T) \]

\[ = \sum_{j=m-t+2}^{k+1} \frac{\xi_j^k}{\pi_{j,m|T}} - \sum_{j=m-t+3}^{k+1} \frac{\xi_j^k}{\pi_{j,m|T}} \]

and

\[ \tilde{\lambda}_{m|T} = \Pr(s_i = 0, s_{i+1} = 0, \ldots, s_m = 0 | \mathcal{F}_T) = \sum_{j=m-t+2}^{k+1} \frac{\xi_j^k}{\pi_{j,m|T}} . \]
Smoothed error variance:

\[ \sigma_{tT}^2 = \hat{S}_t \Pi_{tT} + \bar{S}_t \Lambda_{tT} \]

where

\[ \hat{S}_t = \left[ \frac{\eta_{t-1}}{\eta_t} \hat{\sigma}_{tT}^2 \frac{\eta_{t-1}^2}{\eta_t} \hat{\sigma}_{tT}^2 \ldots \frac{\eta_{t-k}}{\eta_t} \hat{\sigma}_{t+k-1T}^2 \right], \]
\[ \bar{S}_t = \left[ \frac{\bar{\eta}_{t-2}}{\eta_0} \hat{\sigma}_{tT}^2 \frac{\bar{\eta}_{t-2}}{\eta_0} \hat{\sigma}_{tT}^2 \ldots \frac{\bar{\eta}_{t+k-2}}{\eta_0} \hat{\sigma}_{t+k-1T}^2 \right] \]

Conditional error variance estimates for smoothing:

Elements of \( \hat{S}_t \):

\[ \hat{\sigma}_{tjm}^2 = \frac{\eta_0 \sigma_0^2 + \sum_{j=1}^{m} \left( y_j - x_j' \hat{\beta}_{jm} \right)^2 + \left( \hat{\beta}_{jm} - \beta_0 \right) V_0^{-1} \left( \hat{\beta}_{jm} - \beta_0 \right)}{\eta_0 + m - t + 1} \]

Forward recursions for elements of \( \bar{S}_t \):

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

\[ \hat{\sigma}_{tjm}^2 = \hat{\sigma}_{t-1jmT}^2 + \frac{\hat{\sigma}_{t-1jmT}^2 \lambda_{t-1,mT}}{\lambda_{t-1,mT} + \hat{\lambda}_{t-1,mT}} \]
\[ \bar{\eta}_{tjm} = \frac{(\eta_0 + m - t + 2) \pi_{t-1,mT} + \bar{\eta}_{t-1jmT} \hat{\lambda}_{t-1,mT}}{\pi_{t-1,mT} + \hat{\lambda}_{t-1,mT}} \]

For \( \hat{\sigma}_{t+k-1}^2 \) and \( \bar{\eta}_{t+k-1} \), 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 \( \hat{\sigma}_{tjm}^2 = \hat{\sigma}_{tjm}^2 \) and \( \bar{\eta}_{tjm} = (\eta_0 + m) \).
References


<table>
<thead>
<tr>
<th></th>
<th>Horizon</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1Q</td>
<td>2Q</td>
<td>3Q</td>
<td>4Q</td>
<td>5Q</td>
<td>6Q</td>
</tr>
<tr>
<td><strong>Full-Sample OLS Coefficient Estimates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ($\hat{\beta}_1$)</td>
<td>1.74*</td>
<td>1.39*</td>
<td>1.60*</td>
<td>1.69*</td>
<td>1.91*</td>
<td>2.28*</td>
</tr>
<tr>
<td>Slope ($\hat{\beta}_2$)</td>
<td>0.82*</td>
<td>1.04*</td>
<td>0.90*</td>
<td>0.85*</td>
<td>0.71*</td>
<td>0.47*</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.10</td>
<td>0.16</td>
<td>0.12</td>
<td>0.10</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Break Tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andrews-Ploberger (exp)</td>
<td>2.75</td>
<td>4.82*</td>
<td>3.29*</td>
<td>2.32</td>
<td>1.59</td>
<td>0.57</td>
</tr>
<tr>
<td>Nyblom</td>
<td>0.79*</td>
<td>1.18*</td>
<td>0.90*</td>
<td>0.66</td>
<td>0.46</td>
<td>0.18</td>
</tr>
<tr>
<td>Bai-Perron ($F$ statistics)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UDMax</td>
<td>38.88*</td>
<td>40.87*</td>
<td>14.82*</td>
<td>10.77</td>
<td>3.66</td>
<td>4.62</td>
</tr>
<tr>
<td>sup $F(1</td>
<td>0)$</td>
<td>38.88*</td>
<td>40.87*</td>
<td>14.82*</td>
<td>10.77</td>
<td>6.95</td>
</tr>
<tr>
<td>sup $F(2</td>
<td>1)$</td>
<td>18.31*</td>
<td>7.69</td>
<td>2.60</td>
<td>3.57</td>
<td>3.05</td>
</tr>
<tr>
<td>sup $F(3</td>
<td>2)$</td>
<td>18.31*</td>
<td>3.23</td>
<td>2.60</td>
<td>127.18*</td>
<td>8.92</td>
</tr>
<tr>
<td>sup $F(4</td>
<td>3)$</td>
<td>7.44</td>
<td>1.66</td>
<td>2.16</td>
<td>0.80</td>
<td>9.60</td>
</tr>
<tr>
<td>sup $F(5</td>
<td>4)$</td>
<td>10.55</td>
<td>3.23</td>
<td>1.34</td>
<td>8.72</td>
<td>3.91</td>
</tr>
</tbody>
</table>


**Note:** The tests apply to the intercept and slope in the regression model in (27) using quarterly data from 1968:Q1-2006:Q4. The minimum regime length in the Bai-Perron test is set to 10% of the sample and the maximum number of breaks equals 5. A * superscript denotes significance at 5%.
<table>
<thead>
<tr>
<th>Horizon</th>
<th>1Q</th>
<th>2Q</th>
<th>3Q</th>
<th>4Q</th>
<th>5Q</th>
<th>6Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{10}$</td>
<td>2.14</td>
<td>1.98</td>
<td>2.13</td>
<td>2.06</td>
<td>2.08</td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td>(0.56)</td>
<td>(0.73)</td>
<td>(0.97)</td>
<td>(0.53)</td>
<td>(0.52)</td>
<td>(0.50)</td>
</tr>
<tr>
<td>$V_{\beta_{10}}$</td>
<td>0.29</td>
<td>0.26</td>
<td>0.22</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.42)</td>
<td>(0.40)</td>
<td>(0.48)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>0.58</td>
<td>0.59</td>
<td>0.53</td>
<td>0.55</td>
<td>0.51</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(0.29)</td>
<td>(0.38)</td>
<td>(0.27)</td>
<td>(0.22)</td>
<td>(0.19)</td>
</tr>
<tr>
<td>$V_{\beta_{20}}$</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td>1.86</td>
<td>1.79</td>
<td>1.87</td>
<td>2.28</td>
<td>2.25</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>(0.34)</td>
<td>(0.32)</td>
<td>(0.39)</td>
<td>(1.14)</td>
<td>(1.07)</td>
<td>(0.92)</td>
</tr>
<tr>
<td>$\eta_0$</td>
<td>3.52</td>
<td>5.03</td>
<td>4.75</td>
<td>3.89</td>
<td>3.53</td>
<td>3.38</td>
</tr>
<tr>
<td></td>
<td>(1.85)</td>
<td>(2.93)</td>
<td>(2.63)</td>
<td>(2.24)</td>
<td>(2.63)</td>
<td>(2.25)</td>
</tr>
<tr>
<td>$p_{11}$</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td>(0.20)</td>
<td>(0.32)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>$p_{00}$</td>
<td>0.90</td>
<td>0.95</td>
<td>0.95</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

**Note:** I estimate the MB(20) model in (27) using the sample 1968:Q1-2006:Q4. I list robust standard errors in parentheses below the parameter estimates. The 5 percent critical value for the LR test of $H_0: V_{\beta_{10}} = V_{\beta_{20}} = 0$ is 4.23 and the 10 percent critical value is 2.95.
<table>
<thead>
<tr>
<th></th>
<th>MB(20)</th>
<th>MSNR (MSC)</th>
<th>MSR (MSC)</th>
<th>TVP</th>
<th>Rolling OLS(40)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In Sample</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC difference</td>
<td>−447.2</td>
<td>−12.2</td>
<td>−12.2</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td><strong>Out of sample: No Re-estimation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KL divergence</td>
<td>−176.5</td>
<td>7.0</td>
<td>21.2*</td>
<td>20.4*</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.3)</td>
<td>(2.4)</td>
<td>(3.5)</td>
<td>(0.3)</td>
</tr>
<tr>
<td>Relative MSFE</td>
<td>4.55</td>
<td>1.12</td>
<td>1.59*</td>
<td>1.53*</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.10)</td>
<td>(3.44)</td>
<td>(3.28)</td>
<td>(0.19)</td>
</tr>
</tbody>
</table>

**Note:** I estimate the model in (27) for $h=2$ using the sample 1967:Q1-1986:Q4, and forecast over the period 1987:Q1-2006:Q4. The MB(20) column shows AIC (in sample), predictive log likelihood (out of sample), and mean squared forecast error (out of sample). Defining $K$ as the number of estimated parameters, $AIC = 2L(\hat{\theta}) - 2K$. Below the out-of-sample statistics in parentheses are $t$-statistics for testing a zero difference between the MB(20) model and the alternative model; a * superscript denotes significance at 5 percent.
Figure 1: Percentage Expected Kullback-Leibler Loss from Truncating State Space

Panel A: $p_{00} = 0.995$

Panel B: $p_{00} = 0.95$

Note: Curves show percentage KL divergence between MB($k$) model and nontruncated 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 $k$

Panel A: Break Probability

Panel B: Slope Coefficient

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: Filtered and Smoothed Estimates

Intercept: $\beta_0$

Slope: $\beta_1$

Error Variance: $\sigma_t^2$

Break Probability

Note: The graphs show filtered (-----) and smoothed (— —) coefficient, error variance, and break probability estimates for the MB(20) model in (27) for $h=2$. 