

# 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 breaks that are stochastic in both timing and magnitude. 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. In particular, the likelihood function can be computed analytically using a single iterative pass through the data and thereby avoids Monte Carlo integration. The model generates forecasts and conditional coefficient predictions 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 paper and Jeremy Piger for helpful comments. I also thank an anonymous reviewer for comments that improved the clarity of the paper.

## 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 posits a stochastic process that generates breaks by drawing new values for the coefficients and error variance in some periods but not others. I specify a two-state Markov process to generate the breaks; the first state designates a break and the second state indicates no break. This structure produces a globally stationary process, and thereby generates a stable probability model for the breaks. In addition, the likelihood function for this model can be computed analytically using a single iterative pass through the data. Similar models in the literature (e.g., McCulloch and Tsay 1993, Pesaran, Pettenuzzo, and Timmermann 2006, Giordani and Kohn 2008) require Monte Carlo integration and therefore demand much more computation. Moreover, each of the aforementioned authors employs a Bayesian approach to inference, whereas I present a frequentist approach.

The MB model has a hierarchical hidden Markov structure (see Fine, Singer and Tishby, 1998) with two layers of hierarchy in the hidden Markov process. The break arrival process comprises the first hidden layer. In the second hidden layer, the model draws random coefficients and error variances conditional on the breaks. Finally, the observed data are

generated in a third layer from a linear process conditional on draws from the two hidden Markov layers. The econometric literature on hierarchical regression contains many models with a single hidden layer but few with two layers; some models have random coefficients and fixed breakpoints (e.g., Hildreth and Houck 1968; Swamy 1970; Wooldridge 2005) whereas other models exhibit random breakpoints and fixed coefficients (e.g., Hamilton 1989; Chib 1998; Timmermann 2001).

Standard random coefficient models condition on knowing the breakpoints, i.e., which observations correspond to which random coefficient draw. For example, Wooldridge (2005) presents a model of heterogeneous treatment effects across individuals in a panel (see also Swamy 1970). In that context, the econometrician assumes that the coefficient takes the same value for each observation on an individual, but different values across individuals, i.e., the breakpoints are known. Hildreth and Houck (1968) specify a model in which a new regression coefficient is drawn independently for each observation. With only one observation per draw of the random coefficient, the data provide little information about any particular realization of the coefficient and the model is better conceptualized as a particular model of heteroskedasticity. Cooley and Prescott (1973) extend this model by specifying autoregressive dependence in the random coefficients. Although the coefficients exhibit breakpoints every period, serially correlated draws allow the model to learn about the path of the random coefficients over time.

The hidden Markov model literature tends to take the opposite perspective by treating coefficients as fixed parameters and modeling breakpoints using a discrete hidden Markov chain. In these models, each state in the Markov chain represents a regime, and each regime is characterized by a set of fixed coefficients. Hamilton's (1989) Markov switching model specifies an irreducible chain, which implies in an infinite sample that each regime has been visited before and will be visited again with probability one. When forecasting using this model, each

possible regime in the forecast horizon must have occurred with positive probability in the estimation sample. In contrast, Chib (1998) and Timmermann (2001) each specify a hidden Markov model with a reducible chain. Their model never reverts to previously observed regimes. It therefore has no capacity to forecast through breaks because a break implies a new fixed coefficient value that is unidentified by the pre-break data.

Using a Bayesian approach, Pesaran, Pettenuzzo, and Timmermann (2006) and Koop and Potter (2007) extend the work of Chib (1998) and Timmermann (2001) by adding a second hidden layer to the model. This additional layer allows the coefficients to be modeled as random draws from a stationary distribution. Although their hierarchical hidden Markov structure makes them appear similar to the MB model, they specify the breaks quite differently. Specifically, both Pesaran et al (2006) and Koop and Potter (2007) specify a discrete reducible hidden Markov process with dimension equal to the number of breaks. In contrast, 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 difference not only avoids conditioning on the number of breaks but also produces marked gains computational efficiency. The MB likelihood can be computed in seconds or less, and the likelihood can be maximized numerically in a few minutes. In contrast, the models of Pesaran et al (2006) and Koop and Potter (2007) require computationally expensive Monte Carlo integration.

Several authors have proposed models of permanent breaks that can be framed as double-hidden-layer models. As in the proposed MB model, these models specify a hidden state variable that defines the breaks, rather than the regimes. However, unlike the MB model, the parameters evolve according to random walk processes in the long run. Using Markov chain Monte Carlo simulations, McCulloch and Tsay (1993) developed a univariate version of this model that they applied to several economic time series. However, their algorithm demands

heavy computation. The computational demands imposed by this model motivated several approximations, including those in Harrison and Stevens (1976) and Engle and Smith (1999). Recently, Giordani and Kohn (2008) improved computational efficiency by applying the efficient sampler of Gerlach, Carter, and Kohn (2000). This sampler gains efficiency by integrating the random parameters out of the predictive distribution before drawing the breaks, and it is a Bayesian analog to the predictive likelihood approach (see Lauritzen 1974 and Hinkley 1979) that I employ in generating the likelihood function for the MB model. Like McCulloch and Tsay (1993), Giordani and Kohn (2008) focus on univariate models. Moreover, they incorporate additional non-Markovian and non-Gaussian components that generate the need for simulation methods and strong priors to make inference with their model.

Because most of the aforementioned models require Monte Carlo integration to simulate the likelihood function, they naturally lend themselves to analysis within a Bayesian framework. The likelihood for MB model can be computed analytically, and I analyze the model in a frequentist framework. I use maximum likelihood to estimate the deep parameters of the model, which describe the distribution from which the random coefficients are drawn. In this sense, the approach could be described as empirical Bayes (Robbins 1955). I use the fact that we take multiple draws from the meta distribution to learn about the parameters of that distribution. As noted by Efron (2008) among others, empirical Bayes fits naturally in a frequentist framework, in spite of its “Bayes” moniker.

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. In Section 4, 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 5 contains concluding remarks.

## 2. Markov Breaks Model

The MB model is

$$y_t = x_t' \beta_t + \sigma_t \varepsilon_t, \quad \varepsilon_t | (x_t, \beta_t, \sigma_t) \sim iidN(0,1) \quad (1)$$

$$\beta_t = (1 - \xi_t) \beta_{t-1} + \xi_t v_t, \quad v_t | \sigma_t \sim N(\beta_0, \sigma_t^2 V_0)$$

$$\sigma_t = (1 - \xi_t) \sigma_{t-1} + \xi_t \eta_t^2, \quad \eta_t^{-2} \sim iidG(\sigma_0^{-2}, \eta_0)$$

$$\xi_t \in \{0,1\} \quad \Pr(\xi_{t+1} = j | \xi_t = i) = p_{ij}$$

where  $x_t$  is a  $r \times 1$  vector of explanatory variables that may include lags of  $y_t$ ,  $G$  denotes a Gamma distribution, and  $\xi_t \in \{0,1\}$  evolves according to a first-order Markov process. The process  $\{\xi_t\}$  is independent of  $\{\eta_t^2, v_t, \varepsilon_t\}$  at all leads and lags. The regression error variance  $\sigma_t^2$  is a random variable drawn from an inverse Gamma distribution and the  $r \times 1$  vector  $\beta_t$  contains a set of random coefficients drawn from a stationary normal distribution conditional on  $\sigma_t$ . The model exhibits breaks because it draws new values for  $\beta_t$  and  $\sigma_t$  in some periods but not others. A break occurs in period  $t$  if the model makes a new draw of  $\beta_t$  and  $\sigma_t$  in that period; no break occurs if the model does not make a new draw of  $\beta_t$  and  $\sigma_t$  in period  $t$  and retains the values from the previous period. 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 and 3.

I model the break arrival process using the first-order Markov random variable  $\xi_t \in \{0,1\}$ , defined such that  $\xi_t = 1$  denotes a break in period  $t$  and  $\xi_t = 0$  denotes no break. I use the notation  $p_{ij} = \Pr(\xi_{t+1} = j | \xi_t = i)$  to denote the transition probabilities in this Markov process. When a break occurs, the new draws of  $\beta_t$  and  $\sigma_t$  are independent of  $x_t$  and the observed data up to period  $t-1$ , i.e.,  $(v_t, \eta_t^2) \perp (x_t, \mathfrak{F}_{t-1})$  where  $\mathfrak{F}_{t-1}$  denotes the information in  $(y_1, y_2, \dots, y_{t-1}, x_1, x_2, \dots, x_{t-1})$ . Markov dependence in the break arrival process allows for

clustering of breaks. If breaks take multiple periods to transpire, then the probability of a break next period may be greater if a break occurred this period than if no break occurred this period. The deep parameters of the MB model are  $\beta_0$ ,  $V_0$ ,  $\sigma_0^2$ ,  $\eta_0$ ,  $p_{00}$ , and  $p_{11}$ , which I estimate by maximum likelihood. The distributional assumptions in (1) enable me to form a likelihood function and to generate filtered and smoothed estimates of  $\beta_t$  and  $\sigma_t$  conditional on the data.

## 2.1 Model Properties

If  $\{x_t\}$  is a covariance stationary process and  $p_{00} < 1$ , then  $\{y_t\}$  is also covariance stationary<sup>1</sup>. To see this point, I rewrite (1) in a variance components form as

$$\begin{aligned} y_t &= x_t' \beta_0 + x_t' (\beta_t - \beta_0) + \sigma_t \varepsilon_t \\ &= x_t' \beta_0 + u_t \end{aligned} \quad (2)$$

where  $u_t | (x_t, \sigma_t) \sim N(0, \sigma_t^2 (1 + x_t' V_0 x_t))$ . The error process  $u_t$  has two components: the model error  $\sigma_t \varepsilon_t$  and the deviation of  $x_t' \beta_t$  from its steady-state value. Next, I show that this second component induces autocorrelation and heteroskedasticity in  $u_t$ .

The unconditional autocovariance function for  $u_t$  is

$$\begin{aligned} E(u_t u_{t-k}) &= E(E(u_t u_{t-k} | x_t, x_{t-k}, \sigma_t, \sigma_{t-k})) \\ &= E(x_t' E((\beta_t - \beta_0)(\beta_{t-k} - \beta_0)' | \sigma_t, \sigma_{t-k}) x_{t-k}) \\ &= p_{00}^{k-1} \frac{E(\sigma_{t-k}^2)(1 - p_{11})}{2 - p_{00} - p_{11}} E(x_t' V_0 x_{t-k}) \\ &= p_{00}^{k-1} \frac{\eta_0 (1 - p_{11})}{(\eta_0 - 2)(2 - p_{00} - p_{11})} \sigma_0^2 E(x_t' V_0 x_{t-k}) \end{aligned}$$

where the last two lines follow from the autocovariance of a two-state Markov chain (Hamilton 1994, pg 684) and the first moment of an inverse Gamma distribution. In general, the error term  $u_t$  is autocorrelated, but if  $p_{00} < 1$ , then its autocovariance decays exponentially as  $k$  increases.

<sup>1</sup> If  $p_{00} = 1$ , then  $\{y_t\}$  would still be stationary if  $p_{11} = 0$ . In that case,  $\beta_t$  and  $\sigma_t$  would remain constant throughout any given sample although they may differ from  $\beta_0$  and  $\sigma_0$ . Alternately, if  $\beta_t = \beta_0$  and  $\sigma_t = \sigma_0$  with probability one, then  $\{y_t\}$  would be stationary and ergodic regardless of the value of  $p_{00}$ .

The autocorrelation equals zero when  $E(x_t'V_0x_{t-k})=0$ , which can occur when  $x_t$  is white noise and there is no intercept in the model, or when  $\rho_{11}=1$ , which occurs when there is a break every period. This latter case is the random coefficient model of Hildreth and Houck (1968).

Conditional on an absence of breaks between  $t-k$  and  $t$ , the autocovariance is

$$\begin{aligned} E(u_t u_{t-k} | \xi_t = \dots = \xi_{t-k+1} = 0) &= E(E(u_t u_{t-k} | x_t, \sigma_t, \xi_t = \dots = \xi_{t-k+1} = 0) | \xi_t = \dots = \xi_{t-k+1} = 0) \\ &= E(x_t' E((\beta_t - \beta_0)(\beta_{t-1} - \beta_0)' | \sigma_t, \xi_t = \dots = \xi_{t-k+1} = 0) x_{t-1}) \\ &= E(\sigma_t^2 x_t' V_0 x_{t-k}) \\ &= \frac{\eta_0}{(\eta_0 - 2)} \sigma_0^2 E(x_t' V_0 x_{t-k}) \end{aligned}$$

If a break occurs between  $t-k$  and  $t$ , then the autocovariance is

$$\begin{aligned} E(u_t u_{t-k} | \sum_{i=0}^{k-1} \xi_{t-i} > 0) &= E(E(u_t u_{t-k} | x_t, \sigma_t, \sum_{i=0}^{k-1} \xi_{t-i} > 0) | \sum_{i=0}^{k-1} \xi_{t-i} > 0) \\ &= E(x_t' E((\beta_t - \beta_0)(\beta_{t-k} - \beta_0)' | \sigma_t, \sum_{i=0}^{k-1} \xi_{t-i} > 0) x_{t-k}) \\ &= 0 \end{aligned}$$

Thus,  $u_t$  exhibits nonlinear dynamics because its autocorrelation differs depending on whether a break arrives between  $t-k$  and  $t$ . For example, suppose the model contains only an intercept and that  $E(y_t)=0$ , i.e.,  $x_t = 1$  and  $\beta_0 = 0$ . Suppose a break occurs in period  $t$  and the new draw is  $\beta_t = 1$ . Subsequent draws of  $y_t$  will be centered around one until a new break occurs. Thus, a model based on the unconditional mean (zero), would under-predict  $y_t$  until a new break occurs, i.e., the prediction errors would tend to be positive. A sequence of positive draws on a mean-zero random variable implies positive autocorrelation. However, these prediction errors are uncorrelated with those from before the period  $t$  break, which were centered around a different  $\beta$  value. This example illustrates how shocks in periods in which  $\xi_t=1$  have greater persistence than other shocks (Smith, 2005; Engle and Smith, 1999).

The representation in (2) illuminates the effect of breaks on regression models. Breaks generate heteroskedastic errors and the lack of a break in a particular period produces

autocorrelation. If the MB model exhibits very frequent breaks, including the limiting case of a break every period (Hildreth and Houck, 1968), we have a particular type of heteroskedasticity. Breusch and Pagan (1979) noted this connection when they proposed their popular LM test as a test for homoskedastic errors and fixed coefficients. As breaks become less frequent in the MB model, the errors exhibit stronger autocorrelation.

The preceding discussion is silent about the case in which  $x_t$  includes lags of  $y_t$ . To understand this case suppose  $x_t = y_{t-1}$  so that  $y_t$  follows the conditional AR(1) process defined by

$$y_t = \beta_t y_{t-1} + \sigma_t \varepsilon_t$$

From Brandt (1986),  $y_t$  is stationary if  $-\infty < E(\log |\beta_t|) < 0$ . This result extends in a straightforward manner to cover AR processes of a general order, i.e., the process is covariance stationary if the largest root has mean less than one.

## 2.2 Comparison to Hidden Markov Models

To model stochastic breaks in regression models, Hamilton (1989) specifies an  $N$ -state hidden Markov switching (MS) model with recurring states. Under this model

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

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

$$Q = \begin{bmatrix} q_{11} & q_{21} & q_{31} & \cdots & q_{N1} \\ q_{12} & q_{22} & q_{32} & \cdots & q_{N2} \\ q_{13} & q_{23} & q_{33} & \cdots & q_{N3} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ q_{1N} & q_{2N} & q_{3N} & \cdots & q_{NN} \end{bmatrix}.$$

This model shares with the MB model the property of stationarity, assuming ergodicity of the Markov chain. However, the MS 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 MS model a break generates a draw from a discrete distribution. Thus, the MS model only forecasts well after breaks when the process reverts to a previously observed state. Moreover, as the number of states increases, the MS 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.

Chib (1998) and Timmermann (2001) specify  $N$ -state hidden Markov models with nonrecurring states (MNR). Under this model, the Markov state variable  $\xi_t$  is an  $N$ -dimensional reducible Markov chain with transition probability matrix

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

Like the MS model, this 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.

The MB model has two additional advantages over the MNR model. First, by specifying a stationary distribution from which the parameters are drawn, it produces a globally stationary model regardless of the number of breaks. This feature allows the MB model to forecast

through possible breaks. Second, it keeps the dimension of the state space under control, and therefore it is computationally straightforward even when the number of breaks is large.

Several recent papers have extended the MNR model to allow forecasting when breaks may occur during the forecast period. Pesaran, Pettenuzzo, and Timmermann (2006) extend Chib (1998) and Timmermann (2001) specifying  $\beta_t$  and  $\sigma_t^{-2}$  in an identical manner to the MB model, i.e., as random draws from normal and Gamma distributions, respectively. However, their specification of the regimes using an  $N$ -dimensional reducible Markov chain means that inference using the model requires extensive computation. Koop and Potter (2007) point out some additional drawbacks of this MNR specification. Most notably, it must produce exactly  $N-1$  breaks, which implies a prior that breaks become increasingly more likely as the end of the sample approaches. Koop and Potter resolve this issue by dispensing with the Markov chain model for the regimes, and instead modeling regime durations using a Poisson distribution. This solution comes at the cost of increased computational demands because it induces non-Markovian dynamics in the regimes. Specifically, the probability of switching to the next regime depends on the time spent in the current regime.

Giordani and Kohn (2008) reduce the computational demands of Koop and Potter's approach by adopting a similar state-space representation to the MB model in (1). Although their paper presents several feasible models with varying degrees of flexibility, they focus on univariate models and apply the following specification to US inflation

$$y_t = x_t' \beta_t + \sigma_t \xi_{1t} \varepsilon_t$$

$$\beta_t = \beta_{t-1} + \xi_{2t} v_t$$

$$\ln \sigma_t = \ln \sigma_{t-1} + \xi_{3t} u_t$$

where the innovations  $\{\varepsilon_t, v_t, u_t\}$  are Gaussian,  $x_t = [1 \ y_{t-1}]$ , and  $\xi_{1t}$ ,  $\xi_{2t}$ , and  $\xi_{3t}$  are hidden binomial random variables that designate outliers ( $\xi_{1t}$ ), breaks in the coefficients ( $\xi_{2t}$ ), and

breaks in the variance ( $\xi_{3t}$ ). All breaks in this model are permanent, as implied by the random walk structure of  $\beta_t$  and  $\ln \sigma_t$ . In this sense, the model is similar to that of McCulloch and Tsay (1993). The non-Gaussianity induced by the assumed volatility process causes this model to require the use of simulation methods, and the flexibility imbued by the three  $\xi$  terms demands strong priors.

### 2.3 Comparison to Time-Varying Parameter Models

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

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

where  $\beta_t = A\beta_{t-1} + \omega v_t$ ,  $\varepsilon_t \sim N(0,1)$ ,  $v_t \sim N(0,1)$ , and  $E(\varepsilon_t v_s) = 0$  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, Giordani and Kohn 2008). Alternatively, a GARCH(1,1) model specifies

$$\sigma_t = \omega_0 + \omega_1 \sigma_{t-1} + \omega_2 E((y_{t-1} - x'_{t-1} \beta_{t-1})^2 | y_{t-1}, y_{t-2}, \dots, x_{t-1}, x_{t-2}, \dots)$$

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  $E((y_{t-1} - x'_{t-1} \beta_{t-1})^2 | y_{t-1}, y_{t-2}, \dots, x_{t-1}, x_{t-2}, \dots)$  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, MS, and MNR, 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.

### 3. Filtering, Forecasting, and Smoothing

#### 3.1 Likelihood Function

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

$$L(\theta) = \sum_{t=1}^T \log(f(y_t | x_t, \mathfrak{F}_{t-1})),$$

where  $f$  denotes the predictive density<sup>2</sup> of  $y_t$  conditional on  $x_t$  and  $\mathfrak{F}_{t-1}$ . To obtain the predictive density, I rewrite the model as

$$y_t = x'_t \beta_{t-j} + \sigma_{t-j} \varepsilon_t, \tag{3}$$

where  $t-j$  denotes the period of the most recent break. Then, defining the matrices

$B_t \equiv [\beta_t \ \beta_{t-1} \ \dots \ \beta_1]$  and  $S_t \equiv [\sigma_t \ \sigma_{t-1} \ \dots \ \sigma_1]$ , the model becomes

$$y_t = x'_t B_t \Xi_t + S_t \Xi_t \varepsilon_t, \tag{4}$$

---

<sup>2</sup> Throughout the manuscript, I use the notation  $f$  to denote a generic density function.

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.,

$$\Xi_t = \begin{bmatrix} \xi_t \\ (1-\xi_t)\xi_{t-1} \\ \vdots \\ (1-\xi_t)\dots(1-\xi_3)\xi_2 \\ (1-\xi_t)\dots(1-\xi_3)(1-\xi_2) \end{bmatrix}.$$

Using this nomenclature, the predictive density is

$$\begin{aligned} f(y_t | x_t, \mathfrak{I}_{t-1}) &= \sum_{i=1}^t f(y_t | \Xi_{i,t} = 1, x_t, \mathfrak{I}_{t-1}) \Pr(\Xi_{i,t} = 1 | x_t, \mathfrak{I}_{t-1}) \\ &\equiv \Theta_t' \Xi_{t|t-1}, \end{aligned}$$

where  $\Xi_{i,t}$  denotes the  $i^{\text{th}}$  element of  $\Xi_t$ ,  $\Xi_{t|t-1} \equiv \Pr(\Xi_{i,t} = 1 | x_t, \mathfrak{I}_{t-1}) = E(\Xi_t | x_t, \mathfrak{I}_{t-1}) = E(\Xi_t | \mathfrak{I}_{t-1})$  is the expected value of the state vector conditional on the past, and the term  $\Theta_t$  denotes a  $t$  dimensional vector whose  $i^{\text{th}}$  element denotes the density of  $y_t$ , conditional on  $x_t$ ,  $\mathfrak{I}_{t-1}$ , and the event  $\Xi_{i,t} = 1$ , i.e., conditional on the most recent break occurring in period  $t-i+1$ . 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.

To develop intuition for the construction of  $\Theta_t$ , I begin by conditioning on  $\sigma_t$ . In the first period, we have uncertainty from the error term  $\varepsilon_1$  and the coefficient  $\beta_1$ , so the predictive density for  $t=1$  is

$$y_1 | x_1, \sigma_1 \sim N(x_1' \beta_0, \sigma_1^2 (1 + x_1' V_0 x_1)).$$

In period 2 there may or may not be a break, so the state variable for  $t=2$  is

$$\Xi_2 = \begin{bmatrix} \xi_2 \\ 1 - \xi_2 \end{bmatrix}.$$

If a break occurs in period 2, then we draw a new value of  $\beta_t$  and we have

$$y_2 | x_2, \sigma_2, \Xi_{1,2} = 1, \mathfrak{S}_1 \sim N\left(x_2' \beta_0, \sigma_2^2 (1 + x_2' V_0 x_2)\right).$$

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<sup>3</sup>

$$y_2 | x_2, \Xi_{2,2} = 1, \mathfrak{S}_1 \sim N\left(x_1' \hat{\beta}_{1|1}, \sigma_2^2 (1 + x_2' \hat{V}_{1|1} x_2)\right).$$

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

$$\begin{bmatrix} y_1 | x_1, \sigma_1 \\ \beta_1 | x_1, \sigma_1 \end{bmatrix} \sim N\left(\begin{bmatrix} x_1' \beta_0 \\ \beta_0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 (1 + x_1' V_0 x_1) & \sigma_1^2 x_1' V_0 \\ \sigma_1^2 V_0 x_1 & \sigma_1^2 V_0 \end{bmatrix}\right),$$

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

$$y_t | x_t, \sigma_t, \Xi_{i+1,t} = 1, \mathfrak{S}_{t-1} \sim N\left(x_t' \hat{\beta}_{t-i|t-1}, \sigma_t^2 (1 + x_t' \hat{V}_{t-i|t-1} x_t)\right), \quad i = 0, 1, \dots, t-1, \quad (5)$$

where  $\hat{\beta}_{t-i|t-1} = \hat{V}_{t-1|t-1} (V_0^{-1} \beta_0 + \sum_{j=1}^i x_{t-j} y_{t-j})$  and  $\hat{V}_{t-i|t-1} = (V_0^{-1} + \sum_{j=1}^i x_{t-j} x_{t-j}')^{-1}$ . Because  $\beta_0$

and  $V_0$  can be viewed as prior moments of  $\beta_t$ , 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  $\sigma^2$  and normally distributed data and priors (Koop 2003, pg. 37).

---

<sup>3</sup> 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  $\beta_t = \beta_{t-5}$  conditional on data up to period  $t$  and knowledge that the most recent break occurred in period  $t-5$ .

To obtain  $\Theta_t$ , I must remove the conditioning on  $\sigma_t$ . Applying standard results that are most often used in Bayesian regression analysis with conjugate priors (Koop 2003, pg. 46), the  $(i+1)^{\text{th}}$  element of  $\Theta_t$  is a  $t$ -density with  $\eta_0 + i$  degrees of freedom. Specifically,

$$y_t | x_t, \xi_{i+1,t} = 1, \mathfrak{I}_{t-1} \sim t\left(x_t' \hat{\beta}_{t-i|t-1}, \hat{\sigma}_{t-i|t-1}^2 (1 + x_t' \hat{V}_{t-i|t-1} x_t), \eta_0 + i\right), \quad (6)$$

for all  $i = 0, 1, \dots, t-1$ , 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 (5). As is the case in Bayesian regression analysis, the term  $\hat{\sigma}_{t-i|t-1}^2$  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  $\beta_{t-j}$ . Specifically,

$$\hat{\sigma}_{t-i|t-1}^2 = \frac{\eta_0 \sigma_0^2 + \sum_{j=1}^i \left(y_{t-j} - x_{t-j}' \hat{\beta}_{t-i|t-1}\right)^2 + \left(\hat{\beta}_{t-i|t-1} - \beta_0\right)' V_0^{-1} \left(\hat{\beta}_{t-i|t-1} - \beta_0\right)}{\eta_0 + i}, \quad (7)$$

(see Koop 2003, pg. 37).

The break indicator  $\xi_t$  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  $\Xi_{t|t}$  (Hamilton 1989). The filter is

$$\Xi_{t|t} = \frac{\Theta_t * \Xi_{t|t-1}}{\Theta_t' \Xi_{t|t-1}} \quad (8)$$

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

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 numerically with respect to the unknown parameters of the model  $(\beta_0, V_0, \eta_0, \sigma_0^2, \rho_{00}, \rho_{11})$  and obtain standard errors through numerical differentiation of the log likelihood function. These parameters are not identified if the true predictive density can be obtained with probability one by evaluating at parameters other than the true values. If  $\rho_{00}=1$ , then  $V_0$ ,  $\eta_0$ , and  $\rho_{11}$  are unidentified because those parameters can only be identified by breaks. If  $\rho_{11}=1$ , then no periods without breaks would be observed so  $\rho_{00}$  would be unidentified. Otherwise, assuming the elements of  $x_t$  are not perfectly collinear, the model parameters are identified.

Consistency and asymptotic normality of the maximum likelihood estimator (MLE) for general hidden Markov models such as the MB model have proven difficult to establish formally, although models in this family have been applied extensively. Baum and Petrie (1966) proved consistency and asymptotic normality of the MLE when both the state and observed variables lie in a finite discrete space. Leroux (1992) first showed consistency of the MLE for the case with a general observation space and a finite state space under the assumption of exogenous right-hand-side variables. This result was extended by Douc, Moulines, and Rydén (2004) to include autoregressive models such as those proposed in Hamilton (1989) and those with a continuous but compact state space. None of these results covers models with a continuous non-compact state space, such as that of Cooley and Prescott (1973). Douc, Moulines, Olsson, and van Handel (2011) recently established consistency of MLE for a general class of hidden Markov models that includes both continuous and discrete state spaces and therefore includes the MB model. To my knowledge, asymptotic normality has not yet been proven for this general class of models.

My treatment of the regression coefficients and error variance as random variables evokes Bayesian imagery, partly because predicting the random coefficients conditional on the data

requires the same updating formulae used in Bayesian regression analysis with conjugate priors. However, as in the random coefficients literature (Harville 1977, Hildreth and Houck 1968), 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 models of Pesaran, Pettenuzzo, and Timmermann (2006), Koop and Potter (2007) and Giordani and Kohn (2008).

### 3.2 Filtering

Conditional on  $\mathfrak{S}_t$  and the most recent break having occurred in period  $t-i$ ,  $\beta_{t-i}$  has the multivariate  $t$  distribution

$$\beta_{t-i} | \Xi_{i+1,t} = 1, \mathfrak{S}_t \sim t\left(\hat{\beta}_{t-i|t}, \hat{\sigma}_{t-i|t}^2 \hat{V}_{t-i|t}, \eta_0 + i + 1\right), \quad (9)$$

for all  $i = 0, 1, \dots, t-1$  (Koop 2003, pg. 37). From (4), the regression coefficients can be written as

$\beta_t = B_t \Xi_t$ , which implies that the filtered coefficient predictions are

$$\beta_{t|t} \equiv E(\beta_t | \mathfrak{S}_t) = B_{t|t} \Xi_{t|t}, \quad (10)$$

where  $B_{t|t} \equiv \begin{bmatrix} \hat{\beta}_{t|t} & \cdots & \hat{\beta}_{2|t} & \hat{\beta}_{1|t} \end{bmatrix}$ . Thus, to predict the coefficients conditional on data up to  $t$ ,

the model takes a weighted average of predictions that include progressively more past data.

The prediction 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

$$\text{var}(\beta_t | \mathfrak{S}_t) = \sum_{i=0}^{t-1} \text{var}(\beta_{t-i} | \Xi_{i+1,t} = 1, \mathfrak{S}_t) \Xi_{i+1,t|t}$$

where

$$\text{var}(\beta_{t-i} | \Xi_{i+1,t} = 1, \mathfrak{S}_t) = \frac{\eta_0 + i + 1}{\eta_0 + i - 1} \hat{\sigma}_{t-i|t}^2 \hat{V}_{t-i|t}.$$

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 of  $\beta_t$  and  $\Xi_t$  to obtain

$$f(\beta_t | \mathfrak{S}_t) = \sum_{i=0}^{t-1} f(\beta_t | \Xi_{i+1,t} = 1, \mathfrak{S}_t) \Xi_{i+1,t|t},$$

where the densities  $f(\beta_t | \Xi_{i+1,t} = 1, \mathfrak{S}_t)$  come directly from the  $t$  distribution in (9). The distribution of  $\beta_t | \mathfrak{S}_t$  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  $\Xi_{i+1,t|t} \approx 1$  for some  $i$ , then the distribution of  $\beta_t$  conditional on  $\mathfrak{S}_t$  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

$$\sigma_{t-i}^{-2} | \Xi_{i+1,t} = 1, \mathfrak{S}_t \sim G(\hat{\sigma}_{t-i|t}^{-2}, \eta_0 + i + 1), \quad (11)$$

for all  $i = 0, 1, \dots, t-1$  (see Koop 2003, pg. 37). I obtain the distribution of  $\sigma_t^{-2}$  conditional only on the observed data by integrating  $\Xi_t$  out of the joint density  $f(\sigma_t^{-2}, \Xi_t | \mathfrak{S}_t)$ , which yields

$$f(\sigma_t^{-2} | \mathfrak{S}_t) = \sum_{i=0}^{t-1} f(\sigma_t^{-2} | \Xi_{i+1,t} = 1, \mathfrak{S}_t) \Xi_{i+1,t|t}, \quad (12)$$

where  $f(\sigma_t^{-2} | \Xi_{i+1,t} = 1, \mathfrak{S}_t)$  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

$$E\left(\sigma_{t-i}^2 \mid \Xi_{i+1,t} = 1, \mathfrak{S}_t\right) = \frac{\eta_0 + i + 1}{\eta_0 + i - 1} \hat{\sigma}_{t-i|t}^2. \quad (13)$$

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

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

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

The predictions 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

$$\begin{aligned} \beta_{t+l|t} &\equiv E(\beta_{t+l} \mid \mathfrak{S}_t) = B_{t+l|t} \left( \prod_{j=1}^l P_{t+j} \right) \Xi_{t|t}, \\ \sigma_{t+l|t}^2 &\equiv E(\sigma_{t+l}^2 \mid \mathfrak{S}_t) = S_{t+l|t} \left( \prod_{j=1}^l P_{t+j} \right) \Xi_{t|t}, \end{aligned}$$

where the conditional predictions are  $S_{t+l|t} \equiv \left[ \frac{\eta_0}{\eta_0-2} \sigma_0^2 \quad \dots \quad \frac{\eta_0}{\eta_0-2} \sigma_0^2 \quad \frac{\eta_0+1}{\eta_0-1} \hat{\sigma}_{t|t}^2 \quad \dots \quad \frac{\eta_0+l}{\eta_0+l-2} \hat{\sigma}_{1|t}^2 \right]$

and  $B_{t+l|t} \equiv \left[ \beta_0 \quad \dots \quad \beta_0 \quad \hat{\beta}_{t|t} \quad \dots \quad \hat{\beta}_{1|t} \right]$ . Because post-break values of  $\beta_t$  and  $\sigma_t^2$  are drawn from a stationary distribution, the long run forecasts are  $\lim_{l \rightarrow \infty} \beta_{t+l|t} = \beta_0$  and  $\lim_{l \rightarrow \infty} \sigma_{t+l|t}^2 = \frac{\eta_0}{\eta_0 - 2} \sigma_0^2$ .

### 3.3 Smoothing

In this section, I present an algorithm that uses future observations to smooth the filtered predictions in (10) and (14). I predict  $\beta_t$  and  $\sigma_t^2$  by averaging across various predictions 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 prediction is

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

where  $\pi_{jm|T} \equiv \Pr(\xi_j = 1, n_j = m+1 | \mathfrak{I}_T)$ ,  $\bar{\pi}_{jT|T} \equiv \Pr(\xi_j = 1, n_j \geq T+1 | \mathfrak{I}_T)$ , 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  $\{\xi_{j+1} = 0, \xi_{j+2} = 0, \xi_{j+3} = 1\}$ . The probability terms 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  $\div$  denotes element-by-element division (Hamilton 1994).

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

$$\begin{aligned} \pi_{jm|T} &= \Pr(\xi_j = 1, \xi_{j+1} = 0, \dots, \xi_m = 0, \xi_{m+1} = 1 | \mathfrak{I}_T) \\ &= \Pr(\xi_j = 1, \xi_{j+1} = 0, \dots, \xi_m = 0 | \mathfrak{I}_T) - \Pr(\xi_j = 1, \xi_{j+1} = 0, \dots, \xi_m = 0, \xi_{m+1} = 0 | \mathfrak{I}_T) \\ &= \Xi_{m-j+1, m|T} - \Xi_{m-j+2, m+1|T}. \end{aligned} \quad (16)$$

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,

$$\begin{aligned} \bar{\pi}_{jT|T} &= \Pr(\xi_j = 1, n_j \geq T+1 | \mathfrak{I}_T) \\ &= \Pr(\xi_j = 1, \xi_{j+1} = 0, \dots, \xi_{T-1} = 0, \xi_T = 0 | \mathfrak{I}_T) \\ &= \Xi_{T-j+1, T|T}. \end{aligned} \quad (17)$$

In sum, the smoothed predictions are

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

Similarly, for the error variance,

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

where  $\hat{\sigma}_{j|m}^2$  is as defined in (7).

The distribution of  $\beta_t$  conditional on  $\mathfrak{I}_T$  is non-Gaussian because of the possibility of

breaks. However, conditional on knowledge of the breaks, the coefficient predictions are  $t$ -distributed which implies that

$$f(\beta_t | \mathfrak{I}_T) = \sum_{m=t}^{T-1} \sum_{j=0}^t f(\beta_j | \xi_j = 1, n_j = m+1, \mathfrak{I}_T) (\Xi_{m-j+1, m|T} - \Xi_{m-j+2, m+1|T}) \\ + \sum_{j=0}^t f(\beta_j | \xi_j = 1, n_j \geq T+1, \mathfrak{I}_T) \Xi_{T-j+1, T|T}$$

where  $f(\beta_j | \xi_j = 1, n_j = m+1, \mathfrak{I}_T) = f(\beta_m | \Xi_{m-j+1, m} = 1, \mathfrak{I}_m)$  denotes the density of the distribution in (9). It follows that  $\beta_t$  conditional on  $\mathfrak{I}_T$  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,  $\beta_t | \mathfrak{I}_T$  departs from a  $t$ -distribution. Similarly, the distribution of  $\sigma_t^{-2}$  conditional on  $\mathfrak{I}_T$  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

$$f(\sigma_t^{-2} | \mathfrak{I}_T) = \sum_{m=t}^{T-1} \sum_{j=0}^t f(\sigma_j^{-2} | \xi_j = 1, n_j = m+1, \mathfrak{I}_T) (\Xi_{m-j+1, m|T} - \Xi_{m-j+2, m+1|T}) \\ + \sum_{j=0}^t f(\sigma_j^{-2} | \xi_j = 1, n_j \geq T+1, \mathfrak{I}_T) \Xi_{T-j+1, T|T}$$

where  $f(\sigma_j^{-2} | \xi_j = 1, n_j = m+1, \mathfrak{I}_T) = f(\sigma_m^{-2} | \Xi_{m-j+1, m} = 1, \mathfrak{I}_m)$  denotes the density of the distribution in (11).

### 3.5 Remarks

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

*Remark 1.* 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  $p_{11} = \Pr(\xi_t = 1 | \xi_{t-1} = 1)$  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 2.* In many applications, the value of the MB model may be determined by its ability to forecast out of sample. Within an estimation sample, 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.

*Remark 3.* Conditional on the deep parameters, the likelihood function is 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  $\hat{\beta}_{t-i|t-1}$ ,  $\hat{V}_{t-i|t-1}$ ,  $\hat{\sigma}_{t-i|t-1}^2$ , and  $\Xi_{t|t-1}$  to remove the unknown  $\beta_{t-i}$ ,  $\sigma_{t-i}^2$ , and  $\Xi_t$  from the likelihood. The idea of integrating the unknown  $\beta_{t-i}$  and  $\sigma_{t-i}^2$  out of the likelihood also underlies the sampler of Gerlach et al (2000), which enables efficient Markov chain Monte Carlo simulation of Gaussian mixture models (Giordani and Kohn 2008).

*Remark 4.* 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 5.* 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  $H_0: V_{10}=\dots=V_{q0}=0$  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  $x_1, \dots, x_q$  do not belong in the model (i.e.,  $H_0: \beta_{10}=\dots=\beta_{q0}=0, V_{10}=\dots=V_{q0}=0$ ), 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.

#### 4. 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.

#### 4.1 Model Specification

The model is

$$y_t = \beta_{1t} + \beta_{2t}x_{t-h} + \varepsilon_t, \quad t = 1, 2, \dots, T \quad (18)$$

where  $\varepsilon_t | \sigma_t \sim iidN(0, \sigma_t^2)$ ,  $h$  denotes the forecast horizon,  $x_t$  denotes the 10-year Treasury bond yield minus the 3-month Treasury bill yield, and  $y_t = 400 * \ln(GDP_t / GDP_{t-1})$  denotes the annualized quarterly change in the logarithm of seasonally-adjusted real GDP. For the period 1967:Q1 through 2009:Q4, Table 1 shows that OLS estimation of (18) generates a maximum  $R^2$  of 0.12 at the two-quarter horizon. At horizons greater than two quarters, the  $R^2$  steadily decreases, falling to 0.05 at the six-quarter horizon.

To assess whether the coefficients in (18) change over time, I apply three test statistics: (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 horizons up to three quarters. 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 break in 1984 and another in 2000 for the two- and three-quarter horizons; it finds four different break dates for the one-quarter horizon model.

Table 2 shows estimates of the parameters of an MB model for all 6 horizons. The two-quarter horizon model exhibits the largest likelihood value, and Figure 1 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 1 also shows that the great moderation returned  $\beta_{2t}$  and  $\sigma_t$  to their pre-1973 levels. The 2009 recession is characterized by a jump in  $\sigma_t$  to pre-1984 levels and a brief reversal in sign of the slope coefficient  $\beta_{2t}$ . I analyze this period in more depth in Section 4.2.

Figure 1 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 1 shows that the yield spread climbed from 1.5 in 1993 to 3.8 in 1999 before dropping back to 1.5 just as quickly in 2000.

The smoothed probabilities in Figure 1 reveal the large break in 1984 associated with the great moderation. The smoothed probability that a break occurred in 1984:Q3 is 0.54, and the estimated probability that a break occurred sometime 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  $\Xi_{t|T}$  for  $t=1985:Q1$ . The filtered probabilities in Figure 1 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 0.10 upon realization of the 1984:Q3 data. After one more quarter it rises to 0.20; it rises to 0.28 after two quarters and to 0.38 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 1.41 in 1984:Q3 to 0.68 just three quarters later.

The largest filtered probability estimate is 0.21 in 2003:Q3. This observation is associated with a moderate jump in the intercept  $\beta_{1t}$ . 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 between 16 and 21 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 1). Much of 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. The associated variance term for the slope coefficient ( $V_{\beta_2}$ ) declines towards zero as the horizon increases, and the variance term for the intercept ( $V_{\beta_1}$ ) declines slightly. These variance terms equal zero in a model with no breaks in the coefficients. To test for breaks in  $\beta_t$ , I conduct a likelihood ratio test of the null hypothesis  $V_{\beta_1} = V_{\beta_2} = 0$  using the critical values in Remark 5, which are valid for this test as long as  $\sigma_t$  is not constant. I reject the null hypothesis at 5 percent at all horizons.

The estimated transition probabilities show that breaks decrease in frequency as the horizon increases to four quarters. The five- and six-quarter horizon models show a greater break frequency, but the transition probabilities for these models have wide confidence intervals. There is no evidence of Markov dependence in the breaks because the hypothesis that  $p_{11} = 1 - p_{00}$  cannot be rejected at any horizon.

#### 4.2 Predictive Distribution

The likelihood framework of the MB model naturally produces a forecast distribution rather than a point forecast. To illustrate, I present in Figure 2 the evolution of the coefficients and forecasts through the beginning of the 2008-09 recession. The first row of the figure begins in 2008:Q1 and shows the conditional densities of the intercept term ( $\beta_{1t} | \mathfrak{F}_{t-1}$ ), the slope coefficient ( $\beta_{2t} | \mathfrak{F}_{t-1}$ ), the mean forecast ( $x'_{t-2}\beta_t | \mathfrak{F}_{t-1}$ ) and GDP growth ( $y_t | \mathfrak{F}_{t-1}$ ). The square on the horizontal axis in each predictive distribution plot denotes the realization of  $y_t$  in that quarter. The graph on the far right shows the conditional probability that a break occurred in each of the past 20 quarters (i.e., the first 20 elements of  $\Xi_{t|t-1}$ ).

In 2008:Q1, the intercept was centered around 1.8 and the slope around 0.4, which produced a one-step-ahead growth forecast tightly distributed around 1.9 (annualized). The 95 percent prediction interval for GDP growth in that quarter was (-1.3, 5.4) and actual annualized GDP growth came in near the lower end of the range at -0.73. The set of backward looking break probabilities show no quarter with greater than a 7 percent probability, but the sum of these 20 probabilities equals 0.6, which indicates that the regression model had been somewhat unstable over the previous 20 quarters.

The distributions for the next two quarters look quite similar. In 2008:Q2, GDP growth again came in at the low end of the 95 prediction interval, but in the third quarter realized GDP

growth was -4.1 percent. This low value caused the model to react in the following quarter. There are four notable features of this reaction. First, all predictive distributions for 2008:Q4 became quite wide reflecting uncertainty about the new regime. Second, the intercept and slope distributions both shifted to the left reflecting the fact that GDP growth had declined in the previous year in spite of a steepening yield curve. Third, the model pinpoints 2007:Q4 as the most likely break date. This quarter was also designated by the NBER dating committee as the beginning of the recession. On their own, the low growth observations in the first half of 2008 were not enough to move the model, but taken together with the large negative realization in the third quarter, the model recognizes the first half of 2008 as likely being part of the new regime. Fourth, the distributions are skewed to the left, which represents the fact that they are a mixture of the tight distributions with high means that would be obtained under the assumption of no break and the wide distributions with low means that arise under the assumption of a recent break.

The remaining rows in Figure 2 show how the model learns about the new regime and gradually tightens its predictive distributions, although substantial uncertainty remains at the end of 2009. The 95 percent prediction interval for GDP growth in 2009:Q4 was (-9.3, 7.8), which reflects both that high volatility of GDP growth and the lack of information that the yield curve had for growth at that time.

### **4.3 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 (18) using data up to 1986:Q4, and use the data from 1987–2009 to evaluate post-sample forecasting performance. I compare the MB model to Markov switching models with recurring (MS) and nonrecurring (MNR) 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 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 model for the out-of-sample period.

I conduct the forecasting experiment using both fixed and expanding estimation samples, with the exception of the MNR model and the rolling OLS regressions. The MNR model has no capacity to predict post-break values of  $\beta_t$  and  $\sigma_t$ , 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 MS 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  $t-1$ .

I perform the forecasting exercise with a fixed estimation sample to highlight the ability of the models to adapt to breaks, which is an ability that each model is designed to possess. Table 3 shows that the MS and TVP-GARCH models perform markedly worse in such a comparison, with predictive log likelihood values 25.0 and 23.2 points worse than the MB model. Similarly, their MSFE values are 62 percent and 55 percent worse than the MB model. Even the MNR and rolling OLS models, which require re-estimation to generate forecasts, perform worse than the fixed-sample MB model. The MNR model is 21.9 points worse in predictive log likelihood and 16 percent worse by MFSE. The rolling OLS model gets closest, with predictive log likelihood 5.5 points worse and MSFE 4 percent worse than the MB model.

Recursive re-estimation of MB model increases its performance advantage. Its RMSE decreases from 6.60 to 5.68, which is 35 percent better than the recursively estimated MNR and

47 percent better than the recursively estimated MS models. The MB model improves to 21 percent better than the rolling OLS estimates, and it is 4 percent better than the 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 MS and MNR models. 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 models.

## 5. Conclusion

In this article I develop the MB 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 and to keep the state space of low dimension. 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, which makes the model computationally straightforward. The resulting MB 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. See Pesaran and Pick (2011) for recent work on a similar topic.

## 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.
- Atkeson, A. and L.E. Ohanian, 2001, "Are Phillips Curves Useful for Forecasting Inflation?" *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1): 2-11.
- Bai, J. and P. Perron, 1998, "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66(1): 47-78.
- Chib, S., 1998, "Estimation and Comparison of Multiple Change-Point Models," *Journal of Econometrics*, 86(2): 221-241.
- Chou, R., R.F. Engle, and A. Kane, 1992, "Measuring risk aversion from excess returns on a stock index," *Journal of Econometrics*, 52(2): 201-24.
- 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.
- Cogley, T. and T.J. Sargent, 2005, "Drifts and Volatilities: Monetary Policies and Outcomes in the Post World War II U.S.," *Review of Economic Dynamics*, 8(2): 262-302.
- Cooley, T.F., and W.R. Parke, 1990, "Asymptotic Likelihood-Based Prediction Functions," *Econometrica*, 58(5): 1215-34.
- Cooley, T.F., and E.C. Prescott, 1973, "An Adaptive Regression Model," *International Economic Review*, 14(2): 364-371.
- Davies, R.B., 1977, "Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternative," *Biometrika*, 64(2): 247-54.
- Douc, R. E. Moulines, J. Olsson, and R. van Handel, 2011, "Consistency of the maximum likelihood estimator for general hidden Markov models," *The Annals of Statistics*, 39(1): 474-513.
- Douc, R., E. Moulines and T. Rydén, 2004, "Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime," *The Annals of Statistics*, 32(5): 2254-2304.
- Efron, B., 2008, "Microarrays, Empirical Bayes and the Two-Groups Model," *Statistical Science*, 23(1): 1-22.
- Elliott, G. and U.K. Müller, 2006, "Optimally Testing General Breaking Processes in Linear Time Series Models," *Review of Economic Studies*, 73(4): 907-940.
- Engle, R.F. and A.D. Smith, 1999, "Stochastic Permanent Breaks," *Review of Economics and Statistics*, 81(4): 553-574.

- Estrella A., A.P. Rodrigues, and S. Schich, 2003, "How Stable is the Predictive Power of the Yield Curve? Evidence from Germany and the United States," *The Review of Economics and Statistics*, 85(3): 629–644.
- Fine, S., Y. Singer and N. Tishby, 1998, "The Hierarchical Hidden Markov Model: Analysis and Applications", *Machine Learning*, 32: 41–62.
- Giacomini, R. and B. Rossi, 2006, "How Stable is the Forecasting Performance of the Yield Curve for Output Growth?," *Oxford Bulletin of Economics and Statistics*, 68(S): 783-795.
- Gerlach, R., C. Carter, and R. Kohn, 2000, "Efficient Bayesian Inference for Dynamic Mixture Models," *Journal of the American Statistical Association*, 95(451): 819-828.
- Giordani, P., and R. Kohn, 2008, "Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models," *Journal of Business and Economic Statistics*, 26(1): 66-77.
- 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.
- Harrison, P.J. and C.F. Stevens, 1976, "Bayesian Forecasting," *Journal of the Royal Statistical Society, Series B*, 38(3): 205-247.
- Harville, D.A., 1977, "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72(358): 320-338.
- 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.
- Kim S., N. Shephard and S. Chib, 1998, "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *Review of Economic Studies*, 65(3): 361-393.
- Koop, G., 2003, *Bayesian Econometrics*, John Wiley & Sons: West Sussex.
- Koop, G., and S. Potter, 2007, "Forecasting and Estimating Multiple Change-Point Models," *Review of Economic Studies*, 74(3): 763-789.
- Lauritzen, S.L., 1974, "Sufficiency, Prediction, and Extreme Models," *Scandinavian Journal of Statistics*, 1: 128-134.
- Leroux, B.G., 1992, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Processes and their Applications*, 40(1): 127–143.
- McConnell, M.M. and G. Perez-Quiros, 2000, "Output Fluctuations in the United States: What Has Changed since the 1980s?" *American Economic Review*, 90(5): 1464-1476.
- McCulloch, R.E., and Tsay, R.S., 1993, "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series," *Journal of the American Statistical Association*, 88(423): 968-978.
- Pesaran, M.H, D. Pettenuzzo , and A. Timmermann, 2006, "Forecasting Time Series Subject to Multiple Structural Breaks," *Review of Economic Studies*, 73(4): 1057-1084.
- Pesaran, M.H. and A. Pick, 2011, "Forecast Combination Across Estimation Windows," *Journal of Business and Economic Statistics*, in press.

- Pesaran, M.H. and A. Timmermann, 2007, "Selection of Estimation Window in the Presence of Breaks," *Journal of Econometrics*, 137(1): 134-161.
- Robbins, H., 1955, "An Empirical Bayes Approach to Statistics," *Proceedings of the Third Berkeley Symposium on Mathematics, Statistics and Probability*, 1: 157-164.
- 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.
- Smith, A., 2005, "Level Shifts and the Illusion of Long Memory in Economic Time Series," *Journal of Business and Economic Statistics*, 23(3): 321-335.
- Smith, A., P.A. Naik, and C.L. Tsai, 2006, "Markov-Switching Model Selection using Kullback-Leibler Divergence," *Journal of Econometrics*, 134(2):553-577.
- Stock, J.H. and M.W. Watson, 1989, "New Indexes of Coincident and Leading Indicators," *NBER Macroeconomic Annual*, 4: 351-394.
- Stock, J.H., and M.W., Watson, 1996, "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics*, 14(1): 11-30.
- 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.
- Wooldridge, J.M., 2005, "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *Review of Economics and Statistics*, 87(2): 385-390.

**Table 1: Tests for Changing Coefficients in GDP Growth Regression**

	Horizon					
	1Q	2Q	3Q	4Q	5Q	6Q
<i>Full-Sample OLS Coefficient Estimates</i>						
Intercept ( $\hat{\beta}_1$ )	1.76*	1.39*	1.51*	1.50*	1.63*	1.91*
Slope ( $\hat{\beta}_2$ )	0.68*	0.93*	0.86*	0.88*	0.80*	0.62*
R <sup>2</sup>	0.06	0.12	0.10	0.10	0.09	0.05
<i>Break Tests</i>						
Elliott-Müller ( <i>J</i> )	-15.06*	-20.87*	-15.80*	-12.48	-10.77	-8.57
Andrews-Ploberger (exp)	5.96*	5.55*	3.93*	2.70	1.82	1.48
Nyblom	1.06*	1.50*	1.14*	0.80*	0.53	0.32
Bai-Perron ( <i>F</i> statistics)						
UDMax	38.44*	39.89*	18.12*	11.41	5.82	3.27
sup F(1 0)	31.69*	39.89*	18.12*	11.41	5.82	2.57
sup F(2 1)	80.72*	25.86*	27.12*	39.06*	2.38	4.83
sup F(3 2)	24.21*	7.37	3.67	5.38	3.68	5.26
sup F(4 3)	24.21*	3.39	4.88	0.93	1.97	5.55
sup F(5 4)	7.25	1.38	9.27	6.03	6.72	5.99
Bai-Perron est. break dates	1980:Q4 1996:Q1 2000:Q2 2005:Q4	1984:Q2	1984:Q2 2000:Q2	-	-	-

**Note:** The tests apply to the intercept and slope in the regression model in (\*\*) using quarterly data from 1968:Q1-2009: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 2: MB Estimates for GDP Growth Regressions (1968:Q1-2009:Q4)**

	Horizon					
	1Q	2Q	3Q	4Q	5Q	6Q
$\beta_{10}$	2.10 (0.52)	2.06 (0.71)	2.19 (0.89)	1.88 (0.88)	1.99 (0.71)	2.14 (0.71)
$V_{\beta_1}$	0.29 (0.44)	0.39 (0.53)	0.40 (0.63)	0.37 (0.57)	0.53 (0.34)	0.58 (0.37)
$\beta_{20}$	0.56 (0.25)	0.46 (0.33)	0.45 (0.37)	0.61 (0.34)	0.53 (0.34)	0.46 (0.26)
$V_{\beta_2}$	0.06 (0.13)	0.06 (0.08)	0.04 (0.07)	0.02 (0.02)	0.00 (0.00)	0.00 (0.00)
$\sigma_0$	1.97 (0.32)	1.92 (0.29)	2.00 (0.33)	2.04 (0.36)	1.95 (0.34)	1.95 (0.29)
$\eta_0$	3.36 (1.37)	4.24 (2.19)	3.98 (1.85)	3.96 (1.89)	3.40 (1.12)	3.42 (1.09)
$\rho_{11}$	0.04 (0.18)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.03 (0.08)	0.03 (0.08)
$\rho_{00}$	0.87 (0.13)	0.94 (0.03)	0.93 (0.05)	0.94 (0.05)	0.75 (0.17)	0.77 (0.12)
$\Pr(s_t=1)$	0.12	0.06	0.07	0.02	0.20	0.19
$t$ -stat $H_0: \rho_{11}=1-\rho_{00}$	-0.31	-2.44	-0.53	-1.62	-0.98	-0.98
LR stat $H_0: V_{\beta_1}=V_{\beta_2}=0$	9.24*	12.60*	7.20*	5.20*	5.74*	6.72*
LLF	-418.9	-415.4	-419.6	-420.0	-420.0	-420.6
LLF of no breaks model	-440.3	-435.5	-437.2	-436.8	-438.4	-441.4

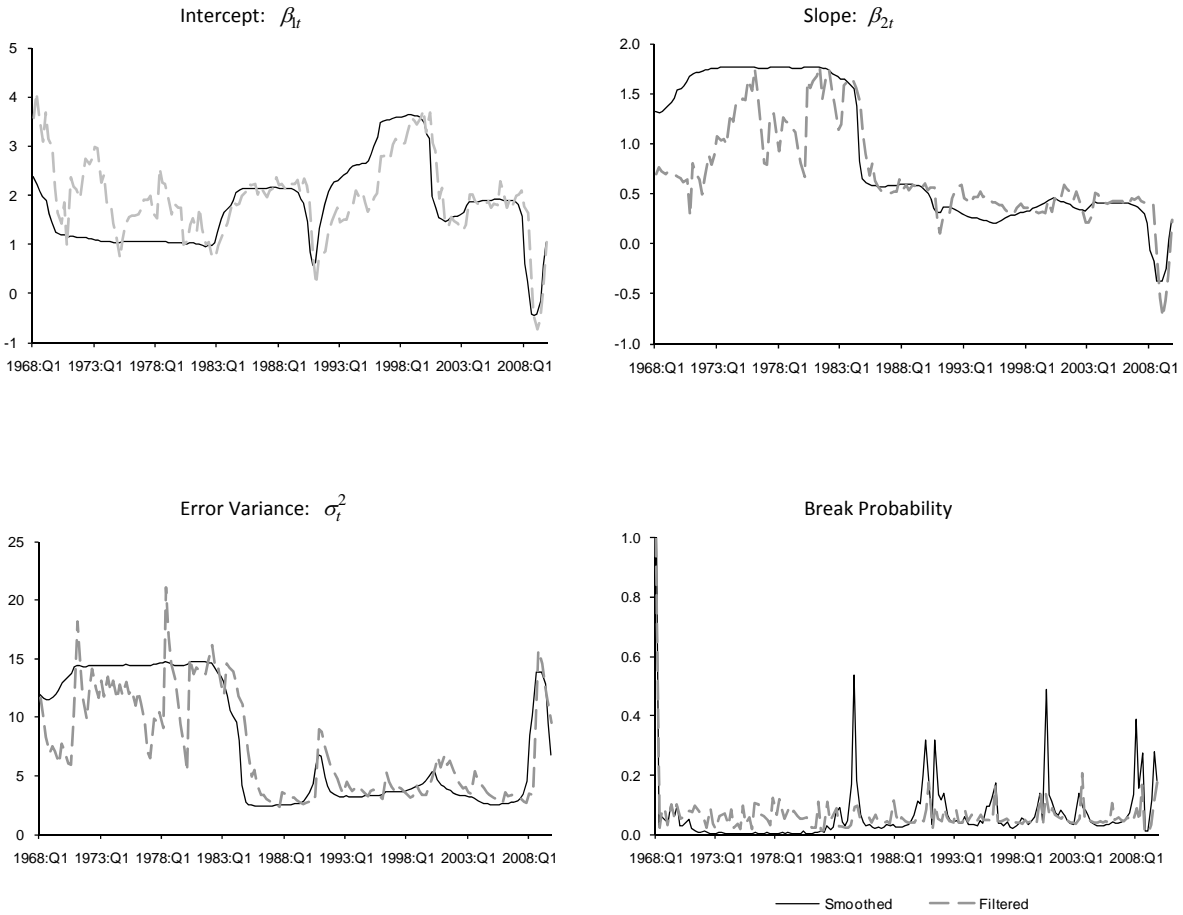
**Note:** I estimate the MB model in (\*\*) using the sample 1968:Q1-2008:Q4. I list robust standard errors in parentheses below the parameter estimates. The 5 percent critical value for the LR test of  $H_0: V_{\beta_1}=V_{\beta_2}=0$  is 4.23 and the 10 percent critical value is 2.95.

**Table 3: Forecasting Comparison**

	MB	Performance Relative to MB			
		MNR (MSC)	MS (MSC)	TVP	Rolling OLS(40)
<i>In Sample</i>					
AIC difference	-447.8	-12.2	-12.2	0.3	
<i>Out of sample: No Re-estimation</i>					
KL divergence	-215.3	21.9 <sup>*</sup> (3.4)	25.0 <sup>*</sup> (2.4)	23.2 <sup>*</sup> (3.3)	5.5 (0.8)
Relative MSFE	6.60	1.16 (1.73)	1.62 <sup>*</sup> (3.69)	1.55 <sup>*</sup> (3.62)	1.04 (0.50)
<i>Out of sample: Recursive Estimation</i>					
KL divergence	-214.9	22.2 <sup>*</sup> (2.0)	15.5 <sup>*</sup> (2.0)	4.1 (1.0)	5.8 (0.8)
Relative MSFE	5.68	1.35 (1.8)	1.47 <sup>*</sup> (2.1)	1.04 (0.6)	1.21 (1.3)

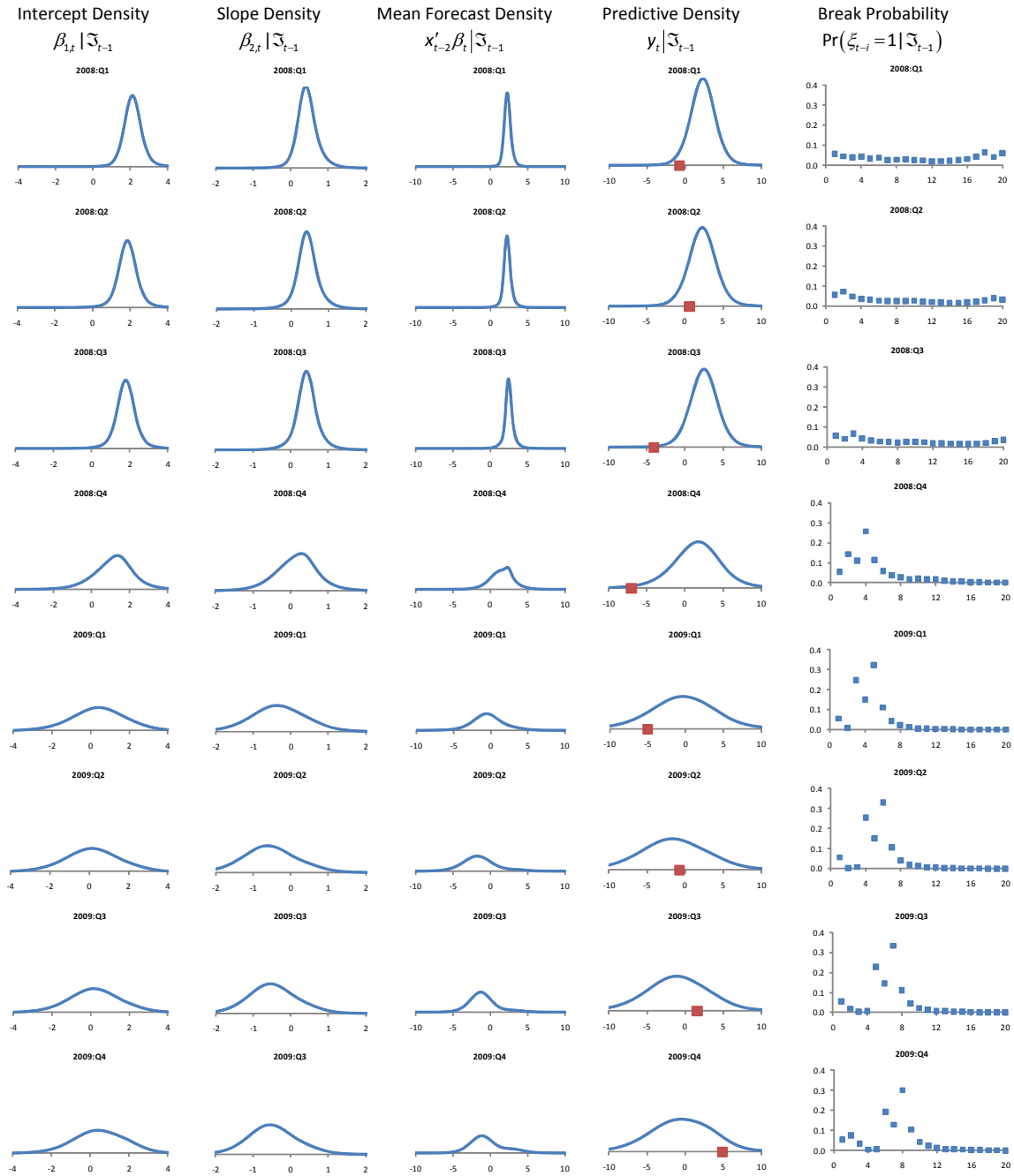
**Note:** I estimate the model in (\*\*) for  $h=2$  using the sample 1967:Q1-1986:Q4, and forecast over the period 1987:Q1-2009:Q4. The MB column shows AIC (in sample), predictive log likelihood (out of sample), and mean squared forecast error (out of sample). Defining  $K$  as the number of estimated parameters,  $AIC = 2L(\theta) - 2K$ . Below the out-of-sample statistics in parentheses are  $t$ -statistics for testing a zero difference between the MB model and the alternative model; a \* superscript denotes significance at 5 percent using standard normal critical values (West 1996).

**Figure 1: Filtered and Smoothed Estimates**



**Note:** The graphs show filtered (— —) and smoothed (—) coefficient, error variance, and break probability estimates for the MB model in (\*\*) for  $h=2$ . The associated parameter estimates are shown in the second column of Table 2.

**Figure 2: Forecasting Through the 2008-09 Recession**



**Note:** All figures generated from a MB model estimated using the second lag of the yield spread as a predictor and data from 1967:Q1-2007:Q4. The densities are calculated using the formulas in equations (10)-(14). The square (■) on the predictive density plots shows the realization of  $y_t$ . In the break probabilities plots, the horizontal axis represents the number of periods into the past, so the plots show estimated breaks probabilities for each quarter in the past years. The values of the yield spread ( $x_{t-2}$ ) for these eight observations were 0.43, 0.87, 1.62, 2.26, 2.37, 2.95, 2.53, and 3.14.