# Regime Switching Model with Endogenous Autoregressive Latent Factor*

Yoosoon Chang† Yongok Choi‡ Joon Y. Park§

**Abstract**

This paper introduces a model with regime switching, which is driven by an autoregressive latent factor correlated with the innovation to the observed time series. In our model, the mean or volatility process is switched between two regimes, depending upon whether the underlying autoregressive latent factor takes values above or below some threshold level. If the latent factor becomes exogenous, our model reduces to the conventional markov switching model, and therefore, our model may be regarded as an extended markov switching model allowing for endogeneity in regime switching. Our model is estimated by the maximum likelihood method using a newly developed modified markov switching filter. For both mean and volatility models that are frequently analyzed in markov switching framework, we demonstrate that the presence of endogeneity in regime switching is indeed strong and ubiquitous.

This version: April 8, 2015

*JEL Classification*: C13, C32

*Key words and phrases*: regime switching model, latent factor, endogeneity, mean reversion, leverage effect, maximum likelihood estimation, markov chain

*We thank the co-editor, Oliver Linton, an associate editor and an anonymous referee for helpful comments. We are also grateful for many useful comments to James Hamilton, Chang-Jin Kim, Frank Schorfheide and the participants at 2013 Princeton-QUT-SMU Conference on Measuring Risk (Princeton), 2013 SETA (Seoul), Conference on Stochastic Dominance & Related Themes (Cambridge, UK), 2013 African Econometric Society Meeting (Accra), 2013 MEG Meeting (Bloomington), 2014 Hitotsubashi Conference on Econometrics for Macroeconomics and Finance, 2014 Workshop on Time Series Econometrics (Frankfurt), 2014 ESAM (Hobart), 2014 NSF-NBER Time Series Conference (St. Louis FED), and 2014 CEF (Oslo), and the seminar attendants at UC Riverside, University of Washington, VU Amsterdam-Tinbergen, Groningen, CORE, Humboldt, Chung-Ang, Keio, Bank of Japan, Bundesbank, ECB, Toulouse School of Economics, Bank of Portugal, Durham, Queen Mary University London, Yale, Atlanta FED, and Bank of Canada.

†Department of Economics, Indiana University.

‡Department of Financial Policy, Korea Development Institute.

§Department of Economics, Indiana University and Sungkyunkwan University.

# 1 Introduction

Regime switching models have been used extensively. In most of these models, two regimes, designated as high and low states of an economy, are introduced with a state process determining one of the regimes to take place in each period. The bi-valued state process is typically modeled as a markov chain. The autoregressive model with this type of markov switching in the mean was first considered by Hamilton (1989), which was further analyzed in Kim (1994). Subsequently, the markov switching has been introduced in more general class of models such as regression models and volatility models by numerous authors. Moreover, various statistical properties of the model have been studied by Hansen (1992), Hamilton (1996), Garcia (1998), Timmermann (2000) and Cho and White (2007), among others. For a nice overview and some extensions of the related literature, the reader is referred to the monograph by Kim and Nelson (1999). Markov-switching models with endogenous explanatory variables have also been considered recently by Kim (2004, 2009).

Though the markov switching models have been used and proven to be quite useful in a wide range of contexts, they have some drawbacks. Most importantly, with a very few exceptions including Diebold et al. (1994) and Kim et al. (2008),[1] they all assume that the markov chain choosing the state of regime is completely independent from all other parts of the model, which is extremely unrealistic in many cases.[2] Note that the exogenous regime switching implies in particular that the future transition between states is completely determined by the current state, and does not rely on the realizations of underlying time series. This is highly unlikely in many practical applications. Instead, we normally expect that the future transition depends critically on the realizations of underlying time series as well as the current and possibly past states. Furthermore, the markov chain determining the state of regime in virtually all of the existing switching models is assumed to be strictly stationary, and cannot accommodate the nonstationarity in the transition probability. This can be restrictive if the transition is strongly persistent.

In this paper, we propose a novel approach to modeling regime switching. In our approach, the mean or volatility process is switched between two regimes, depending upon whether the underlying autoregressive latent factor takes values above or below some threshold level. The latent factor, on the other hand, is assumed to be correlated with the previous

[1] Diebold et al. (1994) differs from our approach in that they consider a markov-switching driven by a set of observed variables. We discuss Kim et al. (2008) in more detail later. See also Kalliovirta et al. (2015) for a related approach.

[2] This is true only for the markov switching models analyzed by the classical approach. In the literature on Bayesian regime switching models, there are several works, including Chib (1996), Chib and Dueker (2004) and Bazzi et al. (2014), which allow for endogenous markov chains. See also Kang (2014), which extends Kim et al. (2008) to a general state space model using a Bayesian approach.

innovation in the model. A current shock to the observed time series therefore affects the regime switching in the next period. Moreover, we allow the autoregressive latent factor to have a unit root and accommodate a strongly persistent regime change. Consequently, our approach remedies both of the aforementioned shortcomings in the conventional markov switching model, and yields a broad class of models with endogenous and possibly non-stationary regime changes. Moreover, it also provides an extracted autoregressive latent factor, which can be used to investigate the dynamic interactions of the mean or volatility process of a given time series with the levels of other observed time series. Our model can be estimated by a modified markov switching filter that we develop in the paper.

If the autoregressive latent factor is exogenous, our model reduces to the conventional markov switching model. Indeed, we show in this case that the conventional two state markov switching model specified by two transition probabilities has the exact one-to-one correspondence with our model specified by the autoregressive coefficient of the latent factor and the threshold level. Therefore, we may always find our model with an exogenous autoregressive latent factor corresponding to a conventional two state markov switching model. They are observationally equivalent and have exactly the same likelihood. Consequently, our model may be regarded as a natural extension of the conventional markov switching model, with the extension made to relax some of its important restrictive features. In the presence of endogeneity, however, our model diverges sharply from the conventional markov switching model. In particular, we show in the paper that the state process in our model is given by a markov process jointly with the underlying time series, and the transition of state systematically interacts with the realizations of underlying time series.

Our paper is closely related to Kim et al. (2008), which considers a regime switching model driven by an endogenous i.i.d. latent factor with the threshold level determined by the previous state and possibly lagged values of the underlying time series.[3] Our model has some important advantages over their model. First, they require the state transition to be dependent only on its immediate past, and this is in contrast with our approach which allows for a high order markov structure.[4] Second, the innovation in our model is set to be correlated with the state variable in the next period, in contrast to their model where it is assumed to be contemporaneously correlated with the state variable in the current period. We believe that the endogeneity of regime switching is more appropriately structured in our approach. In fact, the presence of contemporaneous correlation between the state variable and the innovation of the error term makes their regression model seriously misspecified from

[3]In their model, as well as in our model, the threshold level is also allowed to be dependent upon other exogenous covariates.

[4]This is a serious restriction. For instance, their model cannot be used directly to fit a mean switching AR($p$) model with $p > 1$, including the original Hamilton's model.

the conventional point of view.[5] Finally, we may easily allow for nonstationary transition by letting our autoregressive latent factor have a unit root, whereas their model strictly requires stationarity in transition.

To evaluate the performance of our model and estimation procedure, we conduct an extensive set of simulations. Our simulation results can be summarized as follows. First, the endogeneity of regime switching, if ignored, has a significant deleterious effect on the estimates of model parameters and transition probabilities. This is more so for the mean model than the volatility model, and for the models with stationary latent factors relative to the models with nonstationary latent factors. Second, the presence of endogeneity, if taken into account properly, improves the efficiency of parameter estimates and the precision of estimated transition probabilities. This is because the presence of endogeneity helps to extract more information in the data on the latent states and their transitions. The efficiency gain and the precision enhancement are substantial in some cases, particularly when the latent factor is stationary and the endogeneity is strong. Finally, the likelihood ratio tests for endogeneity work reasonably well in all cases we consider. Though they tend to overreject the null hypothesis of no endogeneity for relatively small samples, they overall appear to be very powerful. In fact, their powers increase sharply up to unity as the degree of endogeneity increases.

For the empirical illustrations of our approach, we analyze the US GDP growth rates and the NYSE/AMEX index returns respectively for our mean and volatility models. For both models, the evidence for endogeneity is unambiguously strong. The estimated correlations between the current shock to the observed time series and the latent factor determining the state in the next period are all very significantly different from zero. For our volatility model, the correlation is estimated to be strongly negative with the values $-0.970$ and $-0.999$ for the two sample periods we consider. Such almost perfect negative correlation implies the presence of strong leverage effect on stock returns. On the other hand, the correlation in our mean model is estimated to be either strongly negative with the value $-0.923$ for the earlier sample period considered in Kim and Nelson (1999) or has nearly perfect positive correlation for the recent subsample. The negative correlation in our stationary mean model implies that the mean reversion of the observed time series occurs in two different levels. Not only does the observed time series revert to its state dependent mean, but also the state dependent mean itself moves to offset the effect of a shock to the observed time series. In the case with the perfect positive correlation in the recent sample, on the

[5] In their regression model, regime switching coefficients of regressors and regression errors are correlated, and regression errors are serially correlated. Consequently, as they point out themselves, their state-dependent mean and volatility no longer represent the conditional mean and volatility of the underlying time series. They consider such a model to analyze the volatility feedback effect of equity returns.

other hand, the movement of the state dependent mean at the second level would entail an unstabilizing effect on the observed time series. For both mean and volatility models, the inferred probabilities appear to be much more accurately predicting the true states of changing regimes if we allow for endogeneity in regime switching.

The rest of the paper is organized as follows. In Section 2, we introduce our model and compare it with the conventional markov switching model. In particular, we show that our model becomes observationally equivalent to the conventional markov switching model, if endogeneity is not present. Section 3 explains how to estimate our model using a modified markov switching filter. The markov property of the state process is also discussed in detail. Section 4 reports our simulation studies, which evaluate the performance of our model relative to the conventional markov switching model. The empirical illustrations in Section 5 consist of the analysis of the US GDP growth rates and the NYSE/AMEX index returns using respectively our mean and volatility models. Section 6 concludes the paper, and Appendix collects the proofs of theorems in the paper and additional figures.

A word on notation. We denote respectively by $\varphi$ and $\Phi$ the density and distribution function of standard normal distribution. The equality in distribution is written as $=_d$. Moreover, we use $p(\cdot)$ or $p(\cdot|\cdot)$ as the generic notation for density or conditional density function. Finally, $\mathbb{N}(a, b)$ signifies the density of normal distribution, or normal distribution itself, with mean $a$ and variance $b$. These notations and notational conventions will be used throughout the paper without further reference.

# 2 A New Approach to Modeling Regime Switching

In this section, we introduce a new approach to modeling regime switching and compare it with the approach used in the conventional markov switching model.

## 2.1 A New Regime Switching Model

In our model, we let a latent factor $(w_t)$ be generated as an autoregressive process

$$w_t = \alpha w_{t-1} + v_t \tag{1}$$

for $t = 1, 2, \ldots$, with parameter $\alpha \in (-1, 1]$ and i.i.d. standard normal innovations $(v_t)$. We use $(\pi_t)$ as a generic notation to denote a state dependent parameter taking values $\pi_t = \underline{\pi}$ or $\bar{\pi}$, $\underline{\pi} < \bar{\pi}$, depending upon whether we have $w_t < \tau$ or $w_t \geq \tau$ with $\tau$ being a threshold

level, or more compactly,

$$\pi_t = \pi(w_t) = \underline{\pi} 1\{w_t < \tau\} + \bar{\pi} 1\{w_t \geq \tau\}, \tag{2}$$

where $\tau$ and $(\underline{\pi}, \bar{\pi})$ are parameters, $\pi : \mathbb{R} \to \{\underline{\pi}, \bar{\pi}\}$, and $1\{\cdot\}$ is the indicator function. In subsequent discussions of our models, we interpret two events $\{w_t < \tau\}$ and $\{w_t \geq \tau\}$ as two regimes that are switched by the realized value of the latent factor $(w_t)$ and the level $\tau$ of threshold, and call $\pi$ the level function of state dependent parameter $(\pi_t)$.

To compare our model with the conventional markov switching model, we may set

$$s_t = 1\{w_t \geq \tau\}, \tag{3}$$

so that we have

$$\pi_t = \pi(s_t) = \underline{\pi}(1 - s_t) + \bar{\pi} s_t$$

exactly as in the conventional markov switching model. The state process $(s_t)$ represents low or high state depending upon whether it takes value 0 or 1. The conventional markov switching model simply assumes that $(s_t)$ is a markov chain taking value either 0 or 1, whereas our approach introduces an autoregressive latent factor $(w_t)$ to define the state process $(s_t)$. In the conventional markov switching model, $(s_t)$ is assumed to be completely independent of the observed time series. In contrast, it will be allowed in our approach to be endogenous, which appears to be much more realistic in a wide range of models used in practical applications.

For identification of the level function $\pi$ in (2), we need to assume that $\underline{\pi} < \bar{\pi}$. To see this, note that $(v_t)$ has the same distribution as $(-v_t)$, and that our level function is invariant with respect to the joint transformation $w \mapsto -w$, $\tau \mapsto -\tau$ and $(\underline{\pi}, \bar{\pi}) \mapsto (\bar{\pi}, \underline{\pi})$. Recall also that, to achieve identification of our level function, we must restrict the variance of the innovations $(v_t)$ to be unity. This is because, for any constant $c > 0$, $(cv_t)$ generates $(cw_t)$ and our level function remains unchanged under the joint transformation $w \mapsto cw$ and $\tau \mapsto c\tau$ in scale. If $\alpha = 1$ and the latent factor $(w_t)$ becomes a random walk, we have an additional issue of joint identification for the initial value $w_0$ of $(w_t)$ and the threshold level $\tau$. In this case, we have $w_t = w_0 + \sum_{i=1}^{t} v_i$ for all $t$ and the transformation $w_0 \mapsto w_0 + c$ for any constant $c$ yields $(w_t + c)$ in place of $(w_t)$. However, our level function does not change under the joint transformation $w \mapsto w + c$ and $\tau \mapsto \tau + c$ in location. Therefore, we set $w_0 = 0$ in this case. On the other hand, the identification problem of the initial value $w_0$ of $(w_t)$ does not arise if we assume $|\alpha| < 1$. Under this assumption, the latent factor $(w_t)$

becomes asymptotically stationary, and we set

$$w_0 =_d \mathbb{N}\left(0, \frac{1}{1-\alpha^2}\right)$$

to make it a strictly stationary process.

We specify our model as

$$\begin{aligned} y_t &= m(x_t, y_{t-1}, \ldots, y_{t-k}, w_t, \ldots, w_{t-k}) + \sigma(x_t, w_t, \ldots, w_{t-k})u_t \\ &= m(x_t, y_{t-1}, \ldots, y_{t-k}, s_t, \ldots, s_{t-k}) + \sigma(x_t, s_t, \ldots, s_{t-k})u_t \end{aligned} \tag{4}$$

with mean and volatility functions $m$ and $\sigma$ respectively, where $(x_t)$ is exogenous and $(u_t)$ and $(v_t)$ in (1) are jointly i.i.d. as

$$\begin{pmatrix} u_t \\ v_{t+1} \end{pmatrix} =_d \mathbb{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{5}$$

with unknown parameter $\rho$.[6] For the brevity of notation, we write

$$m_t = m(x_t, y_{t-1}, \ldots, y_{t-k}, w_t, \ldots, w_{t-k}) = m(x_t, y_{t-1}, \ldots, y_{t-k}, s_t, \ldots, s_{t-k}) \tag{6}$$

$$\sigma_t = \sigma(x_t, w_t, \ldots, w_{t-k}) = \sigma(x_t, s_t, \ldots, s_{t-k}), \tag{7}$$

subsequently. Note that $m_t$ and $\sigma_t$ are conditional mean and volatility of the state dependent variable $(y_t)$ given present and past values of latent factors $w_t, \ldots, w_{t-k}$, as well as the current values of exogenous variables $x_t$ and lagged endogenous variables $y_{t-1}, \ldots, y_{t-k}$.[7]

Our model (4) includes as special cases virtually all models considered in the literature. In our simulations and empirical illustrations, we mainly consider the model

$$\gamma(L)\big(y_t - \mu_t\big) = \sigma_t u_t, \tag{8}$$

where $\gamma(z) = 1 - \gamma_1 z - \cdots - \gamma_k z^k$ is a $k$-th order polynomial, $\mu_t = \mu(w_t) = \mu(s_t)$ and $\sigma_t = \sigma(w_t) = \sigma(s_t)$ are respectively the state dependent mean and volatility of $(y_t)$. We may easily see that the model introduced in (8) is a special case of our general model (4). The model describes an autoregressive process with conditional mean and volatility that are state dependent. It is exactly the same as the conventional markov switching model

[6] We may equivalently define the dynamics of $(w_t)$ as $w_t = \alpha w_{t-1} + \rho u_{t-1} + \sqrt{1-\rho^2}v_t$, where $(u_t)$ and $(v_t)$ are independent i.i.d. standard normals.

[7] The endogenous latent factor $(w_t)$ may naturally be regarded as an economic fundamental determining the regimes of an economy.

considered by Hamilton (1989) and many others, except that the states in our model (8) are determined by an endogenous latent autoregressive factor $(w_t)$ and the level function $\pi$ specified as in (1) and (2), respectively. In fact, it turns out that if we set $\rho = 0$, together with $|\alpha| < 1$, our model in (8) becomes observationally equivalent to the conventional markov switching model, and we may represent it as the standard regime switching model driven by an exogenous two state markov chain. This is shown below.

The model given in (8) may therefore be viewed as an extension of the conventional autoregressive markov switching model, which allows in particular for endogeneity and nonstationarity in regime changes. The autoregressive parameter $\alpha$ of the latent factor $(w_t)$ in (1) controls the persistency of regime changes. In particular, if $\alpha = 1$, the regime change driven by $(w_t)$ becomes nonstationary, and such a specification may be useful in describing regime changes that are highly persistent. On the other hand, the parameter $\rho$ in the joint distribution (5) of the current model innovation $u_t$ and the next period shock $v_{t+1}$ to the latent factor determines the endogeneity of regime changes. As $\rho$ approaches to unity in modulus, the endogeneity of regime change driven by $(w_t)$ becomes stronger, i.e., the determination of the regime in time $t+1$ is more strongly influenced by the realization of innovation $(u_t)$ at time $t$. We observe that $\rho$ is significantly different from zero both in mean and volatility models for many economic and financial time series including the GDP growth rates and stock returns we analyze for our empirical illustrations in the paper.

The interpretation of endogeneity parameter $\rho$, especially its sign, is quite straightforward for our volatility model. If $\rho < 0$, the innovation $u_t$ in the level of $y_t$ at time $t$ becomes negatively correlated with the volatility $\sigma_{t+1}$ of $y_{t+1}$ at time $t+1$. This implies that a negative shock to $(y_t)$ in the current period would entail an increase in volatility in the next period. This is often referred to as the leverage effect, if $(y_t)$ is used to model returns from a financial asset. See, e.g., Yu (2005) for more discussions on the econometric modeling of leverage effect in volatility model for financial asset returns. Of course, $\rho > 0$ means that there is an anti-leverage effect in the model.

For the mean model, the sign of $\rho$ has a more subtle effect on the sample path of the observed time series $(y_t)$. If the lag polynomial $\gamma(z)$ satisfies the stationarity condition, $(y_t)$ becomes stationary. In this case, $(y_t)$ reverts to its state dependent mean $(\mu_t)$, as well as to its global mean $\mathbb{E}y_t$. This is true for both cases of $\rho < 0$ and $\rho > 0$. The mean reverting behavior of $(y_t)$, however, differs depending upon whether $\rho < 0$ or $\rho > 0$. If $\rho < 0$, a positive realization of $u_t$ at time $t$ increases the probability of having low regime in the state dependent mean $\mu_{t+1}$ of $y_{t+1}$ at time $t+1$, and in this sense, the state dependent mean $(\mu_t)$ of the observed time series $(y_t)$ is also reverting. Therefore, the mean reversion of $(y_t)$ takes place in two distinct levels: the reversion of $(y_t)$ to its state dependent mean

$(\mu_t)$, and the movement of $(\mu_t)$ to offset the effect of a shock to $(y_t)$. This would not be the case if $\rho > 0$. In this case, the movement of $(\mu_t)$ at the second level would entail an unstabilizing effect on $(y_t)$. Furthermore, the regime switching is more likely to happen for $\rho > 0$ if $(y_t)$ is between the two state dependent means, whereas it is so for $\rho < 0$ if $(y_t)$ is outside of the two state dependent means. Therefore, we may expect that regime switching becomes relatively more conspicuous if $\rho < 0$, compared with the case $\rho > 0$.

Kim et al. (2008) consider a regression model similar to ours in (4), yet they specify the state dependent regression coefficients $(\beta_t)$ in their model as being dependent only on the current state variable $(s_t)$ and their model is not directly applicable for model (8) with $k \geq 1$. In their model, the state process is defined as $s_t = 1\{v_t \geq \pi_{t-1}\}$, where $(v_t)$ is specified simply as a sequence of i.i.d. latent random variables that is contemporaneously correlated with innovation $(u_t)$ in regression error $(\sigma_t u_t)$.[8] Though their state process $(s_t)$ is endogenous, it is strictly restricted to be first order markovian and stationary as in the conventional markov switching model. Furthermore, in their approach, $(u_t)$ is jointly determined with $(s_t)$ for each time $t$. The presence of contemporaneous correlation between $(u_t)$ and $(s_t)$ entails undesirable consequences on their model: State dependent coefficients $(\beta_t)$ of regressors are contemporaneously correlated with regression errors $(\sigma_t u_t)$, in addition to that regression errors $(\sigma_t u_t)$ are serially correlated.[9] Their regression model is therefore seriously misspecified from the conventional point of view.

Our approach is different. In our model, the state process $(s_t)$ is driven by an endogenous autoregressive latent factor, instead of an independent and identically distributed sequence of random variables. One important consequence of modeling the latent factor as an endogenous autoregressive process is that $(s_t)$ alone is no longer markovian: It is markovian only jointly with the underlying time series $(y_t)$, and consequently, for our model in (8) the conditional distribution of $s_t$ is determined by the past observations of the state dependent variable $y_{t-i}$'s as well as the past states $s_{t-i}$'s for $1 \leq i \leq k+1$ at any time $t$. This is shown more explicitly in the next section. Furthermore, in our model, innovation $(u_t)$ affects the transition of $(s_t)$ only in the next period, and therefore, $(s_t)$ becomes pre-determined in this sense. Modeling endogeneity as in our model not only appears to be more realistic, but also yields a model that is correctly specified as a conventional regression model. Note that $(m_t)$ and $(\sigma_t)$ become respectively the mean and volatility of $(y_t)$ in our model (4).

[8] For an easier comparison, we present their model using our notation. Their model also includes other predetermined variables, which we ignore here to more effectively contrast their approach with ours. As we explain in more detail later, our model may also easily accommodate the presence of other covariates.

[9] Note also that $(\sigma_t)$ does not represent the conditional volatility of their error process $(\sigma_t u_t)$, since $(\sigma_t)$ is contemporaneously correlated with $(u_t)$.

## 2.2 Relationship with Conventional Markov Switching Model

Our model reduces to the conventional markov switching model when the underlying autoregressive latent factor is stationary and independent of the model innovation. This will be explored below. In what follows, we assume

$$\rho = 0$$

to make our models more directly comparable to the conventional markov switching models, and obtain the transition probabilities of the markovian state process $(s_t)$ defined in (3). In our approach, they are given as functions of the autoregressive coefficient $\alpha$ of the latent factor and the level $\tau$ of threshold. Note that

$$\mathbb{P}\big\{s_t = 0 \,\big|w_{t-1}\big\} = \mathbb{P}\big\{w_t < \tau \,\big|w_{t-1}\big\} = \Phi(\tau - \alpha w_{t-1}) \tag{9}$$

$$\mathbb{P}\big\{s_t = 1 \,\big|w_{t-1}\big\} = \mathbb{P}\big\{w_t \geq \tau \,\big|w_{t-1}\big\} = 1 - \Phi(\tau - \alpha w_{t-1}). \tag{10}$$

Therefore, if we let $|\alpha| < 1$ and denote the transition probabilities of the state process $(s_t)$ from low state to low state and from high state to high state by

$$a(\alpha, \tau) = \mathbb{P}\big\{s_t = 0 \,\big|s_{t-1} = 0\big\}\,, \quad b(\alpha, \tau) = \mathbb{P}\big\{s_t = 1 \,\big|s_{t-1} = 1\big\}\,, \tag{11}$$

then it follows that

**Lemma 2.1.** *For $|\alpha| < 1$, transition probabilities of state process $(s_t)$ defined in (3) from low state to low state and high state to high state are given by*

$$a(\alpha, \tau) = \frac{\displaystyle\int_{-\infty}^{\tau\sqrt{1-\alpha^2}} \Phi\left(\tau - \frac{\alpha x}{\sqrt{1-\alpha^2}}\right)\varphi(x)dx}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)}$$

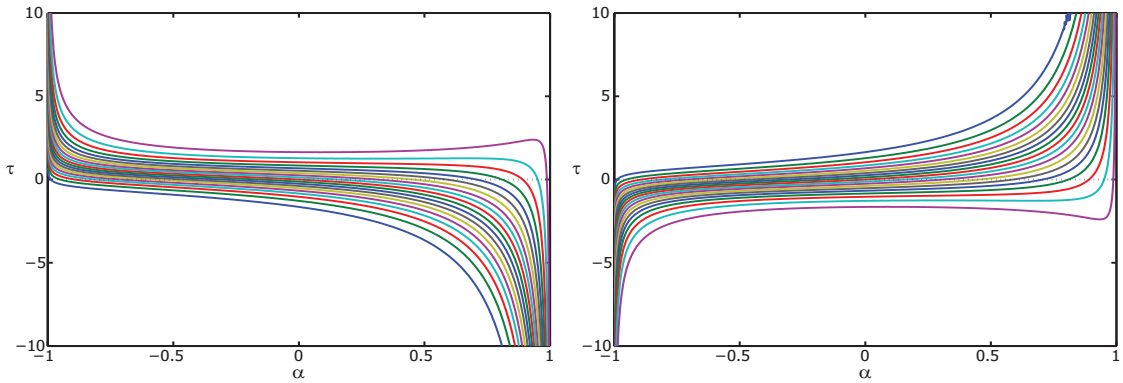
$$b(\alpha, \tau) = 1 - \frac{\displaystyle\int_{\tau\sqrt{1-\alpha^2}}^{\infty} \Phi\left(\tau - \frac{\alpha x}{\sqrt{1-\alpha^2}}\right)\varphi(x)dx}{1 - \Phi\left(\tau\sqrt{1-\alpha^2}\right)},$$

*where $a(\alpha, \tau)$ and $b(\alpha, \tau)$ are defined in (11).*

In particular, the state process $(s_t)$ defined in (3) is a markov chain on $\{0, 1\}$ with transition density

$$p(s_t|s_{t-1}) = (1 - s_t)\omega(s_{t-1}) + s_t\big[1 - \omega(s_{t-1})\big], \tag{12}$$

Figure 1: Contours of Transition Probabilities in $(\alpha, \tau)$-Plane




Notes: The contours of $a(\alpha, \tau)$ and $b(\alpha, \tau)$ are presented respectively in the left and right panels for the levels from 0.05 to 0.95 in the increment of 0.05, upward for $a(\alpha, \tau)$ and downward for $b(\alpha, \tau)$. Hence the top line in the left panel is the contour of $a(\alpha, \tau) = 0.05$, and the bottom line on the right panel represents the contour of $b(\alpha, \tau) = 0.05$.

where $\omega(s_{t-1})$ is transition probability to low state given by

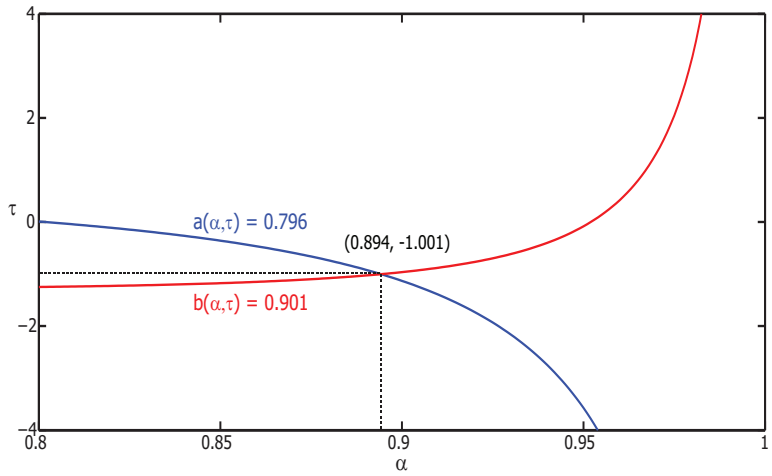
$$\omega(s_{t-1}) = \frac{\left[(1 - s_{t-1}) \int_{-\infty}^{\tau\sqrt{1-\alpha^2}} + s_{t-1} \int_{\tau\sqrt{1-\alpha^2}}^{\infty}\right] \Phi\left(\tau - \frac{\alpha x}{\sqrt{1-\alpha^2}}\right) \varphi(x) dx}{(1 - s_{t-1})\Phi\left(\tau\sqrt{1-\alpha^2}\right) + s_{t-1}\left[1 - \Phi\left(\tau\sqrt{1-\alpha^2}\right)\right]}$$

with respect to the counting measure on $\{0, 1\}$.

The contours of the transition probabilities $a(\alpha, \tau)$ and $b(\alpha, \tau)$ obtained in Lemma 2.1 are presented in Figure 1 for various levels of $0 < a(\alpha, \tau) < 1$ and $0 < b(\alpha, \tau) < 1$. Figure 1 provides the contours of $a(\alpha, \tau)$ and $b(\alpha, \tau)$ in the $(\alpha, \tau)$-plane with $-1 < \alpha < 1$ and $-\infty < \tau < \infty$ for the levels of $a(\alpha, \tau)$ and $b(\alpha, \tau)$ starting from 0.05 with the increment of 0.05 to 0.95. It is quite clear from Figure 1 that there exists a unique pair of $\alpha$ and $\tau$ values yielding any given levels of $a(\alpha, \tau)$ and $b(\alpha, \tau)$, since any contour of $a(\alpha, \tau)$ intersects with that of $b(\alpha, \tau)$ once and only once. For instance, the only pair of $\alpha$ and $\tau$ values that yields $a(\alpha, \tau) = b(\alpha, \tau) = 1/2$ is given by $\alpha = 0$ and $\tau = 0$, in which case we have entirely random switching from high state to low state and vice versa with equal probability.

To more clearly demonstrate the one-to-one correspondence between the pair $(\alpha, \tau)$ of autoregressive coefficient of latent factor and the threshold level and the pair $(a(\alpha, \tau), b(\alpha, \tau))$ of transition probabilities derived in Lemma 2.1, we show how we may find the corresponding values of $\alpha$ and $\tau$ when the values of $a(\alpha, \tau)$ and $b(\alpha, \tau)$ are given. In Figure 2, we set $a(\alpha, \tau) = 0.796$ and $b(\alpha, \tau) = 0.901$, the transition probabilities we obtain from our

Figure 2: Correspondence Between $(\alpha, \tau)$ and $\big(a(\alpha,\tau), b(\alpha,\tau)\big)$




Notes: The increasing and decreasing curves are, respectively, the contours of $a(\alpha,\tau) = 0.796$ and $b(\alpha,\tau) = 0.901$ in the $(\alpha,\tau)$-plane. Their intersection at $(\alpha,\tau) = (0.894, -1.001)$ provides the $(\alpha,\tau)$-pair that yields the transition probabilities $a(\alpha,\tau) = 0.796$ and $b(\alpha,\tau) = 0.901$.

estimates from the Hamilton's model for US GDP growth rates, and plot their contours in the $(\alpha,\tau)$-plane. It is shown that the two contours intersect at one and only one point, which is given by $\alpha = 0.894$ and $\tau = -1.001$.

If we set $\rho = 0$ in our model (4), the transition probabilities of the state process $(s_t)$ in (3) alone completely determine the regime switching without any interaction with other parts of the model. This implies that by setting $\rho = 0$ and obtaining the values of $\alpha$ and $\tau$ corresponding to the given values of $a(\alpha,\tau)$ and $b(\alpha,\tau)$, we may always find a regime switching model with an autoregressive latent factor that is observationally equivalent to any given conventional markov switching model. Our approach, however, produces an important by-product that is not available from the conventional approach: an extracted time series of the autoregressive latent factor driving the regime switching.

Now we let $\alpha = 1$. In this case, the state process $(s_t)$ defined in (3) becomes nonstationary and its transition evolves with time $t$. For $t \geq 1$, we subsequently define the transition probabilities explicitly as functions of time as

$$a_t(\tau) = \mathbb{P}\big\{s_t = 0 \,\big|\, s_{t-1} = 0\big\}\,, \quad b_t(\tau) = \mathbb{P}\big\{s_t = 1 \,\big|\, s_{t-1} = 1\big\}\,, \tag{13}$$

and show that

**Corollary 2.2.** *Let $\alpha = 1$, and let $a_t(\tau)$ and $b_t(\tau)$ be defined as in (13). For $t = 1$, $a_1(\tau) = \Phi(\tau)$ with $\mathbb{P}\{s_0 = 0\} = 1$ if $\tau > 0$, and , $b_1(\tau) = 1 - \Phi(\tau)$ with $\mathbb{P}\{s_0 = 1\} = 1$ if*

$\tau \leq 0$. *Moreover, we have*

$$a_t(\tau) = \frac{\int_{-\infty}^{\tau/\sqrt{t-1}} \Phi\left(\tau - x\sqrt{t-1}\right)\varphi(x)dx}{\Phi\left(\tau/\sqrt{t-1}\right)}$$
$$b_t(\tau) = 1 - \frac{\int_{\tau/\sqrt{t-1}}^{\infty} \Phi\left(\tau - x\sqrt{t-1}\right)\varphi(x)dx}{1 - \Phi\left(\tau/\sqrt{t-1}\right)}$$

*for* $t \geq 2$.

The state process $(s_t)$ is a markov chain with transition density $p(s_t|s_{t-1})$ in (12) which is defined now with the transition probability to low state $\omega(s_{t-1})$ given by

$$\omega(s_{t-1}) = \frac{\left[(1-s_{t-1})\int_{-\infty}^{\tau/\sqrt{t-1}} + s_{t-1}\int_{\tau/\sqrt{t-1}}^{\infty}\right]\Phi\left(\tau - x\sqrt{t-1}\right)\varphi(x)dx}{(1-s_{t-1})\Phi\left(\tau/\sqrt{t-1}\right) + s_{t-1}\left[1 - \Phi\left(\tau/\sqrt{t-1}\right)\right]}$$

with respect to the counting measure on $\{0,1\}$. We may easily see that

$$a_t(\tau), b_t(\tau) \approx 1 - \frac{1}{\pi\sqrt{t-1}}$$

for large $t$, where $\pi = 3.14159\ldots$, and therefore, the transition becomes more persistent in this case as $t$ increases. Moreover, the threshold parameter $\tau$ is unidentified asymptotically. For the asymptotic identifiability of the threshold parameter when $\alpha = 1$, we must set $\tau = \bar{\tau}\sqrt{n}$ for some fixed $\bar{\tau}$. This is obvious because in this case the latent factor $(w_t)$ increases stochastically at the rate $\sqrt{n}$.

To compute the integrals in Lemma 2.1 and Corollary 2.2, we need to obtain the values of

$$M(a,b,c) = \int_{-\infty}^{a} \Phi(b+cx)\varphi(x)dx \tag{14}$$

for all $a, b, c \in \mathbb{R}$. This can be readily done. In fact, upon noticing that $M(a,b,c) = \mathbb{P}\{Z_1 \leq a, Z_2 \leq b + cZ_1\}$, where $Z_1$ and $Z_2$ are independent standard normal random variates, we may easily deduce that

$$M(a,b,c) = \int_{-\infty}^{a}\int_{-\infty}^{b} p(x,y)dydx,$$

where

$$p(x,y)=\frac{1}{2\pi}\exp\left(-\frac{(1+c^2)x^2+2cxy+y^2}{2}\right)=\mathbb{N}\left(0,\begin{pmatrix}1 & -c\\ -c & 1+c^2\end{pmatrix}\right).$$

Therefore, the integrals can be solved, if bivariate normal distribution function is provided. Note that

$$\int_a^\infty \Phi(b+cx)\varphi(x)dx=M(-a,b,-c),$$

which can also be easily obtained, once we compute the integral in (14) for all $a,b,c\in\mathbb{R}$.

## 3 Estimation

Our endogenous regime switching model (4) can be estimated by the maximum likelihood method. For the maximum likelihood estimation of our model, we write the log-likelihood function as

$$\ell(y_1,\dots,y_n)=\log p(y_1)+\sum_{t=2}^{n}\log p(y_t|\mathcal{F}_{t-1}) \tag{15}$$

where $\mathcal{F}_t=\sigma\big(x_t,(y_s)_{s\le t}\big)$, i.e., the information given by $x_t,y_1,\dots,y_t$ for each $t=1,\dots,n$. Of course, the log-likelihood function includes a vector of unknown parameters $\theta\in\Theta$, say, which specifies $m_t=m(x_t,y_{t-1},\dots,y_{t-k},w_t,\dots,w_{t-k})=m(x_t,y_{t-1},\dots,y_{t-k},s_t,\dots,s_{t-k})$ and $\sigma_t=\sigma(x_t,w_t,\dots,w_{t-k})=\sigma(x_t,s_t,\dots,s_{t-k})$. It is, however, suppressed for the sake of notational brevity. The maximum likelihood estimator $\hat\theta$ of $\theta$ is given by

$$\hat\theta=\underset{\theta\in\Theta}{\operatorname{argmax}}\,\ell(y_1,\dots,y_n)$$

as usual. For the model we consider in (8), $\theta$ consists of state dependent mean and volatility parameters, $(\underline{\mu},\bar\mu)$ and $(\underline{\sigma},\bar\sigma)$, as well as the threshold $\tau$ level, the autoregressive coefficient $\alpha$ of the latent factor, the correlation coefficient $\rho$, and the autoregressive coefficients $(\gamma_1,\dots,\gamma_k)$.

To estimate our general switching model in (4) by the maximum likelihood method, we develop a modified markov switching filter. The conventional markov switching filter is no longer applicable, since the state process $(s_t)$ defined in (3) for our model is not a markov chain unless $\rho=0$. To develop the modified markov switching filter that can be used to estimate our model, we let

$$\Phi_\rho(x)=\Phi\big(x/\sqrt{1-\rho^2}\big) \tag{16}$$

for $|\rho|<1$. We have

**Theorem 3.1.** *Let $|\rho| < 1$. The bivariate process $(s_t, y_t)$ on $\{0,1\} \times \mathbb{R}$ is a $(k+1)$-st order markov process, whose transition density with respect to the product of the counting and Lebesgue measure is given by*

$$\begin{aligned} &p(s_t, y_t | s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) \\ &\quad = p(y_t | s_t, \ldots, s_{t-k}, y_{t-1}, \ldots, y_{t-k}) p(s_t | s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}), \end{aligned}$$

*where*

$$p(y_t | s_t, \ldots, s_{t-k}, y_{t-1}, \ldots, y_{t-k}) = \mathbb{N}\left(m_t, \sigma_t^2\right) \tag{17}$$

*and*

$$\begin{aligned} &p(s_t | s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) \\ &= (1 - s_t)\omega_\rho(s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) \\ &\quad + s_t\big[1 - \omega_\rho(s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1})\big] \end{aligned} \tag{18}$$

*with the transition probability $\omega_\rho$ of the endogenous state process $(s_t)$ to low state. If $|\alpha| < 1$, $\omega_\rho$ is given by*

$$\begin{aligned} &\omega_\rho(s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) \\ &= \frac{\left[(1-s_{t-1})\displaystyle\int_{-\infty}^{\tau\sqrt{1-\alpha^2}} + s_{t-1}\int_{\tau\sqrt{1-\alpha^2}}^{\infty}\right] \Phi_\rho\left(\tau - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}} - \frac{\alpha x}{\sqrt{1-\alpha^2}}\right)\varphi(x)dx}{(1-s_{t-1})\Phi(\tau\sqrt{1-\alpha^2}) + s_{t-1}\left[1 - \Phi(\tau\sqrt{1-\alpha^2})\right]}, \end{aligned}$$

*and, if $\alpha = 1$, for $t = 1$, $\omega_\rho(s_0) = \Phi(\tau)$ with $\mathbb{P}\{s_0 = 0\} = 1$ and $\mathbb{P}\{s_0 = 1\} = 1$ respectively when $\tau > 0$ and $\tau \leq 0$ and, for $t \geq 2$,*

$$\begin{aligned} &\omega_\rho(s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) \\ &= \frac{\left[(1-s_{t-1})\displaystyle\int_{-\infty}^{\tau/\sqrt{t-1}} + s_{t-1}\int_{\tau/\sqrt{t-1}}^{\infty}\right] \Phi_\rho\left(\tau - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}} - x\sqrt{t-1}\right)\varphi(x)dx}{(1-s_{t-1})\Phi(\tau/\sqrt{t-1}) + s_{t-1}\left[1 - \Phi(\tau/\sqrt{t-1})\right]}. \end{aligned}$$

Theorem 3.1 fully specifies the joint transition of $(s_t)$ and $(y_t)$ in case of $|\rho| < 1$.[10]

If $|\rho| = 1$, we have perfect endogeneity and $\Phi_\rho$ in (16) is not defined. In this case, the current shock to model innovation $u_t$ fully dictates the realization of latent factor $w_{t+1}$ determining the state in the next period. Consequently the transition of the state process

[10] It is clearly seen from Theorem 3.1 that the parameters $\alpha$, $\tau$ and $\rho$ in our model are all identified.

$(s_t)$ given by the density $p(s_t|s_{t-1},\ldots,s_{t-k-1},y_{t-1},\ldots,y_{t-k-1})$, which is derived above for $|\rho|<1$ in Theorem 3.1, is no longer applicable. When $|\rho|=1$, the transition probability to low state $\omega_\rho(s_{t-1},\ldots,s_{t-k-1},y_{t-1},\ldots,y_{t-k-1})$ behaves differently, which in turn implies that transition density of the state process needs to be modified accordingly. The transition probability to low state $\omega_\rho$ in this case is given explicitly below for various values of AR coefficient $\alpha$ of the latent factor $(w_t)$.

**Corollary 3.2.** *If $|\rho|=1$, the transition probability of the endogenous state process $(s_t)$ to low state $\omega_\rho(s_{t-1},\ldots,s_{t-k-1},y_{t-1},\ldots,y_{t-k-1})$ on the previous states and past observed time series, introduced in Theorem 3.1, is given as follows:*

*(a) If $\alpha=0$*

$$\omega_\rho(s_{t-1},\ldots,s_{t-k-1},y_{t-1},\ldots,y_{t-k-1})=\begin{cases}1, & \text{if } \rho\dfrac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}<\tau\\ 0, & \text{otherwise}\end{cases}$$

*(b) If $0<\alpha<1$,*

$$\begin{aligned}&\omega_\rho(s_{t-1},\ldots,s_{t-k-1},y_{t-1},\ldots,y_{t-k-1})\\&\quad=(1-s_{t-1})\min\left(1,\frac{\Phi\left(\left(\tau-\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)\frac{\sqrt{1-\alpha^2}}{\alpha}\right)}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)}\right)\\&\qquad+s_{t-1}\max\left(0,\frac{\Phi\left(\left(\tau-\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)\frac{\sqrt{1-\alpha^2}}{\alpha}\right)-\Phi\left(\tau\sqrt{1-\alpha^2}\right)}{1-\Phi\left(\tau\sqrt{1-\alpha^2}\right)}\right)\end{aligned}$$

*(c) If $-1<\alpha<0$,*

$$\begin{aligned}&\omega_\rho(s_{t-1},\ldots,s_{t-k-1},y_{t-1},\ldots,y_{t-k-1})\\&\quad=s_{t-1}\min\left(1,\frac{1-\Phi\left(\left(\tau-\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)\frac{\sqrt{1-\alpha^2}}{\alpha}\right)}{1-\Phi\left(\tau\sqrt{1-\alpha^2}\right)}\right)\\&\qquad+(1-s_{t-1})\max\left(0,\frac{\Phi\left(\tau\sqrt{1-\alpha^2}\right)-\Phi\left(\left(\tau-\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)\frac{\sqrt{1-\alpha^2}}{\alpha}\right)}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)}\right)\end{aligned}$$

*(d) If $\alpha=1$, for $t=1$, $\omega_\rho(s_0,y_0)=\Phi\left(\tau-\rho(y_0-m_0)/\sigma_0\right)$ with $\mathbb{P}\{s_0=0\}=1$ and*

$\mathbb{P}\{s_0 = 1\} = 1$ *respectively when* $\tau > 0$ *and* $\tau \leq 0$ *and, for* $t \geq 2$,

$$\omega_\rho(s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1})$$
$$= \begin{cases} 1 - s_{t-1}, & \text{if } \rho \dfrac{y_{t-1} - m_{t-1}}{\sigma_{t-1}} > 0 \\ \dfrac{\Phi\left(\left(\tau - \rho \dfrac{y_{t-1} - m_{t-1}}{\sigma_{t-1}}\right) \dfrac{1}{\sqrt{t-1}}\right) - s_{t-1}\Phi\left(\tau/\sqrt{t-1}\right)}{(1 - s_{t-1})\Phi\left(\tau/\sqrt{t-1}\right) + s_{t-1}\left[1 - \Phi\left(\tau/\sqrt{t-1}\right)\right]}, & \text{otherwise} \end{cases}$$

As shown in Theorem 3.1 and Corollary 3.2, the transition density of the state process $(s_t)$ at time $t$ from time $t-1$ depends upon $y_{t-1}, \ldots, y_{t-k-1}$ as well as $s_{t-1}, \ldots, s_{t-k-1}$. The state process $(s_t)$ alone is therefore not markovian. However, the state process augmented with the observed time series $(s_t, y_t)$ becomes a $(k+1)$-st order markov process. If $\rho = 0$, we have $\omega_\rho(s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) = \omega(s_{t-1})$. In this case, the state process $(s_t)$ reduces to a first order markov process independent of $(y_t)$ as in the conventional markov switching model, with the transition probabilities obtained in Lemma 2.1.

Our modified markov switching filter consists of the prediction and updating steps, which are entirely analogous to those in the usual Kalman filter. To develop the modified markov switching filter, we write

$$p(y_t|\mathcal{F}_{t-1}) = \sum_{s_t} \cdots \sum_{s_{t-k}} p(y_t|s_t, \ldots, s_{t-k}, \mathcal{F}_{t-1}) p(s_t, \ldots, s_{t-k}|\mathcal{F}_{t-1}). \tag{19}$$

Since $p(y_t|s_t, \ldots, s_{t-k}, \mathcal{F}_{t-1}) = p(y_t|s_t, \ldots, s_{t-k}, y_{t-1}, \ldots, y_{t-k})$ is given by (17), it suffices to have $p(s_t, \ldots, s_{t-k}|\mathcal{F}_{t-1})$ to compute the log-likelihood function in (15), which we obtain in the prediction step. For the prediction step, we note that

$$p(s_t, \ldots, s_{t-k}|\mathcal{F}_{t-1}) = \sum_{s_{t-k-1}} p(s_t|s_{t-1}, \ldots, s_{t-k-1}, \mathcal{F}_{t-1}) p(s_{t-1}, \ldots, s_{t-k-1}|\mathcal{F}_{t-1}), \tag{20}$$

and that $p(s_t|s_{t-1}, \ldots, s_{t-k-1}, \mathcal{F}_{t-1}) = p(s_t|s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1})$, which is given in (18). Consequently, $p(s_t, \ldots, s_{t-k}|\mathcal{F}_{t-1})$ can be readily computed from (20), once we obtain $p(s_{t-1}, \ldots, s_{t-k-1}|\mathcal{F}_{t-1})$ from the previous updating step. Finally, for the updating step, we have

$$\begin{aligned} p(s_t, \ldots, s_{t-k}|\mathcal{F}_t) &= p(s_t, \ldots, s_{t-k}|y_t, \mathcal{F}_{t-1}) \\ &= \frac{p(y_t|s_t, \ldots, s_{t-k}, \mathcal{F}_{t-1}) p(s_t, \ldots, s_{t-k}|\mathcal{F}_{t-1})}{p(y_t|\mathcal{F}_{t-1})}, \end{aligned} \tag{21}$$

where $p(y_t|s_t,\ldots,s_{t-k},\mathcal{F}_{t-1})$ is given by (17), and we may readily obtain $p(s_t,\ldots,s_{t-k}|\mathcal{F}_t)$ from $p(s_t,\ldots,s_{t-k}|\mathcal{F}_{t-1})$ and $p(y_t|\mathcal{F}_{t-1})$.

Using our modified markov switching filter based on the state process $(s_t)$, we can also easily extract the latent autoregressive factor $(w_t)$. This can be done through the prediction and updating steps described above in (20) and (21). In the prediction step, we note that

$$p(w_t,s_{t-1},\ldots,s_{t-k}|\mathcal{F}_{t-1})=\sum_{s_{t-k-1}}p(w_t|s_{t-1},...,s_{t-k-1},\mathcal{F}_{t-1})p(s_{t-1},...,s_{t-k-1}|\mathcal{F}_{t-1}). \quad (22)$$

Since $p(s_{t-1},...,s_{t-k-1}|\mathcal{F}_{t-1})$ is obtained from the previous updating step, we may readily compute $p(w_t,s_{t-1},\ldots,s_{t-k}|\mathcal{F}_{t-1})$ from (22) once we find $p(w_t|s_{t-1},...,s_{t-k-1},\mathcal{F}_{t-1})$, the conditional density of latent factor $(w_t)$ on previous states and past information on the observed time series, which is derived below for various values of AR coefficient $\alpha$ of latent factor and endogeneity parameter $\rho$.

**Corollary 3.3.** *The transition density of latent factor $(w_t)$ on previous states and past observed time series is given as follows:*
*(a) When $|\alpha|<1$ and $|\rho|<1$,*

$$\begin{aligned}
&p\left(w_t|s_{t-1}=1,s_{t-2},...,s_{t-k-1},\mathcal{F}_{t-1}\right)\\
&=\frac{\left(1-\Phi\left(\sqrt{\frac{1-\rho^2+\alpha^2\rho^2}{1-\rho^2}}\left(\tau-\frac{\alpha\left(w_t-\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)}{1-\rho^2+\alpha^2\rho^2}\right)\right)\right)}{1-\Phi\left(\tau\sqrt{1-\alpha^2}\right)}\mathbb{N}\left(\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}},\frac{1-\rho^2+\alpha^2\rho^2}{1-\alpha^2}\right),\\
&p\left(w_t|s_{t-1}=0,s_{t-2},...,s_{t-k-1},\mathcal{F}_{t-1}\right)\\
&=\frac{\Phi\left(\sqrt{\frac{1-\rho^2+\alpha^2\rho^2}{1-\rho^2}}\left(\tau-\frac{\alpha\left(w_t-\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)}{1-\rho^2+\alpha^2\rho^2}\right)\right)}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)}\mathbb{N}\left(\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}},\frac{1-\rho^2+\alpha^2\rho^2}{1-\alpha^2}\right).
\end{aligned}$$

*(b) When $|\alpha| < 1$ and $|\rho| = 1$,*

$$
\begin{aligned}
&p\left(w_t | s_{t-1}=1, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) \\
&= \begin{cases} \dfrac{\frac{\sqrt{1-\alpha^2}}{\alpha}\phi\left(\dfrac{w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}}{\alpha}\sqrt{1-\alpha^2}\right)}{1-\Phi\left(\tau\sqrt{1-\alpha^2}\right)}, & \text{if } w_t \geq \alpha\tau + \rho\dfrac{y_{t-1}-m_{t-1}}{\sigma_{t-1}} \\ 0, & \text{otherwise,} \end{cases} \\
&p\left(w_t | s_{t-1}=0, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) \\
&= \begin{cases} \dfrac{\frac{\sqrt{1-\alpha^2}}{\alpha}\phi\left(\dfrac{w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}}{\alpha}\sqrt{1-\alpha^2}\right)}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)}, & \text{if } w_t \leq \alpha\tau + \rho\dfrac{y_{t-1}-m_{t-1}}{\sigma_{t-1}} \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

*(c) When $\alpha = 1$ and $|\rho| < 1$,*

$$
\begin{aligned}
&p\left(w_t | s_{t-1}=1, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) \\
&= \frac{\left(1 - \Phi\left(\sqrt{\frac{t-t\rho^2+\rho^2}{1-\rho^2}}\left(\tau - \frac{w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}}{t-t\rho^2+\rho^2}\right)\right)\right)}{1-\Phi\left(\tau/\sqrt{t-1}\right)}\mathbb{N}\left(\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}, \frac{t-t\rho^2+\rho^2}{t-1}\right), \\
&p\left(w_t | s_{t-1}=0, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) \\
&= \frac{\Phi\left(\sqrt{\frac{t-t\rho^2+\rho^2}{1-\rho^2}}\left(\tau - \frac{w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}}{t-t\rho^2+\rho^2}\right)\right)}{\Phi\left(\tau/\sqrt{t-1}\right)}\mathbb{N}\left(\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}, \frac{t-t\rho^2+\rho^2}{t-1}\right).
\end{aligned}
$$

*(d) When $\alpha = 1$ and $|\rho| = 1$,*

$$
\begin{aligned}
&p\left(w_t|s_{t-1}=1, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) \\
&= \begin{cases} \dfrac{\dfrac{1}{\sqrt{t-1}}\phi\left(\dfrac{w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}}{t-1}\right)}{1-\Phi\left(\tau/\sqrt{t-1}\right)}, & \text{if } w_t \geq \tau + \rho\dfrac{y_{t-1}-m_{t-1}}{\sigma_{t-1}} \\ 0, & \text{otherwise,} \end{cases} \\
&p\left(w_t|s_{t-1}=0, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) \\
&= \begin{cases} \dfrac{\dfrac{1}{\sqrt{t-1}}\phi\left(\dfrac{w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}}{t-1}\right)}{1-\Phi\left(\tau/\sqrt{t-1}\right)}, & \text{if } w_t \leq \alpha\tau + \rho\dfrac{y_{t-1}-m_{t-1}}{\sigma_{t-1}} \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

We may then obtain

$$
p(w_t, s_{t-1}, ..., s_{t-k}|\mathcal{F}_t) = \frac{p(y_t|w_t, s_{t-1}, ..., s_{t-k}, \mathcal{F}_{t-1})p(w_t, s_{t-1}, ..., s_{t-k}|\mathcal{F}_{t-1})}{p(y_t|\mathcal{F}_{t-1})}, \tag{23}
$$

in the updating step. By marginalizing $p(w_t, s_{t-1}, ..., s_{t-k}|\mathcal{F}_t)$ in (23), we get

$$
p\left(w_t|\mathcal{F}_t\right) = \sum_{s_{t-1}} \cdots \sum_{s_{t-k}} p\left(w_t, s_{t-1}, ..., s_{t-k}|\mathcal{F}_t\right),
$$

which yields the inferred factor,

$$
\mathbb{E}\left(w_t|\mathcal{F}_t\right) = \int w_t p\left(w_t|\mathcal{F}_t\right) dw_t
$$

for all $t = 1, 2, \ldots$. Therefore, we may easily extract the inferred factor, once the maximum likelihood estimates of $p(w_t|\mathcal{F}_t)$, $1 \leq t \leq n$, are available.

We may generalize our model and allow for other covariates to affect the regime switching process. For instance, we may specify the state dependent parameter $\pi_t$ as

$$
\pi_t = \pi(w_t, x_t),
$$

where $(x_t)$ is a time series of covariates that are predetermined and observable, and accordingly the level function $\pi$ as

$$
\pi(w, x) = \underline{\pi} 1\{w < \tau_1 + \tau_2' x\} + \bar{\pi} 1\{w \geq \tau_1 + \tau_2' x\}
$$

with parameters $(\underline{\pi}, \bar{\pi})$ and $(\tau_1, \tau_2)$, in place of (2). The level of threshold for regime switching is therefore given as a linear function of some predetermined and observable covariates. All of our previous results extend to this more general model only with some trivial modifications. Since the required modifications are quite clear, we do not explain them in detail here. This model is more directly comparable to the one considered in Kim et al. (2008).

We may also easily extend our model to allow for a more general level function $\pi(w)$ than the one introduced in (2). One obvious possibility is to use the level function that allows for multiple regimes, more than two. The extended models with a more general level function allowing for multiple regimes can also be estimated using our modified markov switching filter similar to that with the simple two-regime level function that we discussed in detail in the previous sections. We may further extend our model to allow for a continuum of regimes. In this case, however, our modified markov switching filter is no longer applicable, and we need to use a density based filter to estimate the parameters.

# 4 Simulations

To evaluate the performance of our model and estimation procedure, we conduct an extensive set of simulations. In the sequel, we will present our simulation models and results.

## 4.1 Simulation Models

In our simulations, we consider both mean and volatility switching models. For the volatility model, we consider

$$y_t = \sigma(s_t)u_t, \quad \sigma(s_t) = \underline{\sigma}(1 - s_t) + \overline{\sigma}s_t. \tag{24}$$

The parameters $\underline{\sigma}$ and $\overline{\sigma}$ are set at $\underline{\sigma} = 0.04$ and $\overline{\sigma} = 0.12$, which are roughly the same as our estimates for the regime switching volatilities for the stock returns we analyze in the next section. The level of volatility in high regime is three times bigger than that in low regime. On the other hand, our simulations for the mean model rely on

$$y_t = \mu(s_t) + \gamma(y_{t-1} - \mu(s_{t-1})) + \sigma u_t, \quad \mu(s_t) = \underline{\mu}(1 - s_t) + \overline{\mu}s_t. \tag{25}$$

We set the parameter values at $\sigma = 0.8$, $\gamma = 0.5$, $\underline{\mu} = 0.6$ and $\overline{\mu} = 3$. They are approximately the same as the estimates that we obtain using the US real GDP growth rates analyzed in the next section.

For both mean and volatility models, $(s_t)$ and $(u_t)$ are generated as specified in (1), (3) and (5) for the samples of size 500, and iterated 1,000 times. The correlation coefficient

$\rho$ between the current model innovation $u_t$ and the next period innovation $v_{t+1}$ of the latent autoregressive factor is set to be negative for both mean and volatility models, as in most of our empirical results reported in the next section. To more thoroughly study the impact of endogeneity on the estimation of our model parameters, we allow $\rho$ to vary from 0 to $-1$ in the increment of 0.1. On the other hand, we consider three pairs of the autoregressive coefficient $\alpha$ of the latent factor and the threshold level $\tau$ given by $(\alpha, \tau) = (0.4, 0.5), (0.8, 0.7), (1, 9.63)$. The first two pairs with $|\alpha| < 1$ yield stationary latent factors, while the last pair with $\alpha = 1$ makes the latent factor a random walk.

As discussed earlier, if $\rho = 0$, there exists a one-to-one correspondence between the $(\alpha, \tau)$ pair and the pair $(a, b)$ of transition probabilities of state process, where $a$ and $b$ denote respectively the transition probabilities from low to low state and from high to high state. The first pair $(\alpha, \tau) = (0.4, 0.5)$ corresponds to $(a, b) = (0.75, 0.5)$, and the second pair $(\alpha, \tau) = (0.8, 0.7)$ to $(a, b) = (0.86, 0.72)$. The transitions of these two pairs have the same equilibrium distribution given by $(a^*, b^*) = (2/3, 1/3)$, which also becomes the common invariant distribution.[11] This, in particular, implies that the unconditional probabilities of the state being in low and high regimes are $2/3$ and $1/3$ respectively in every period. For the third pair with $\alpha = 1$, the state process is nonstationary and its transition varies over time with no existing invariant distribution. Our choice of $\tau = 9.63$ in the third pair yields the unconditional probabilities $(2/3, 1/3)$ of low and high regimes at the terminal period of our simulation, which makes it comparable to the first two pairs.[12]

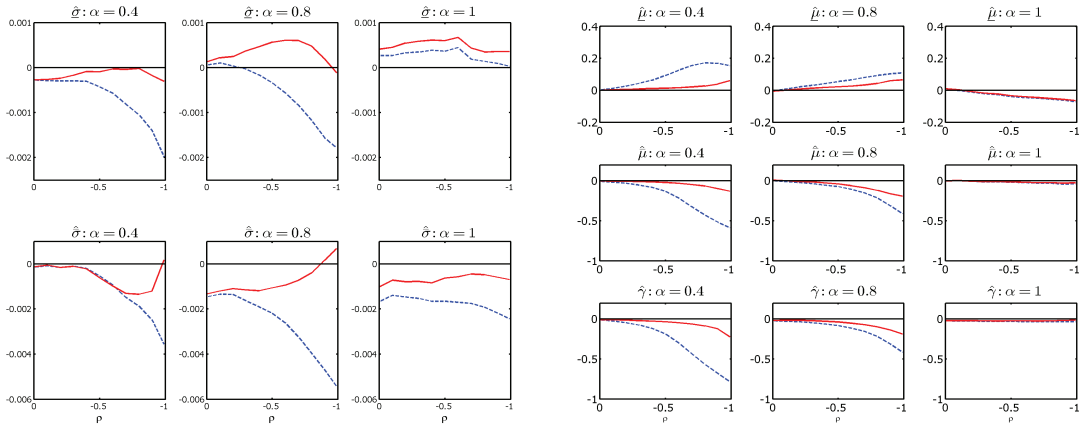
### 4.2 Simulation Results

In our simulations, we first examine the endogeneity bias. The estimators of parameters in our models are expected to be biased if the presence of endogeneity in regime switching is ignored. To see the magnitude of bias resulting from the neglected endogeneity in regime switching, we let $\rho = 0$ for the exogenous regime switching models. Our simulation results are summarized in Figure 3. On the left panel of Figure 3, the bias in the maximum likelihood estimates $\hat{\underline{\sigma}}$ and $\hat{\overline{\sigma}}$ of low and high volatility levels $\underline{\sigma}$ and $\overline{\sigma}$ in the volatility model are presented in the upper and lower parts of the panel for three different levels of $\alpha$ measuring persistency of latent factor in each of the three columns on the panel. Hence, there are 6 individual graphs covering the bias in the estimates $\hat{\underline{\sigma}}$ and $\hat{\overline{\sigma}}$ for three levels of $\alpha = 0.4, 0.8, 1$. Each graph plots the bias of the estimates from the endogenous (red solid line) and exogenous (blue dashed line) models across different levels of endogeneity $\rho$ on

[11] Recall that the invariant distribution of the two-state markov transition given by a $2 \times 2$ transition matrix $P$ is defined by $\pi^* = (a^*, b^*)$ such that $\pi^* = \pi^* P$.

[12] Note that $w_{500} =_d \mathbb{N}(0, 500)$ when $\alpha = 1$ and $\rho = 0$.

Figure 3: Endogeneity Bias

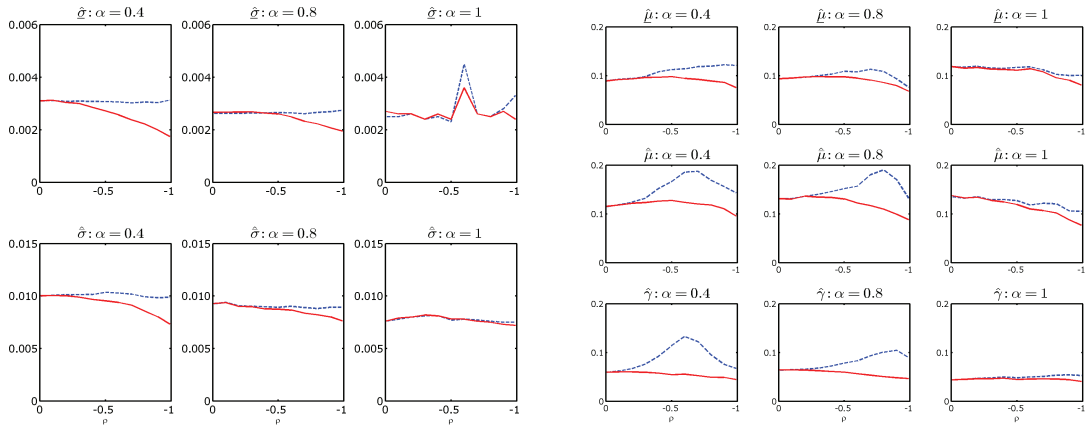



Notes: On the left panel, the bias in ML estimates $\underline{\hat{\sigma}}$ and $\hat{\overline{\sigma}}$ of low and high volatility levels $\underline{\sigma}$ and $\overline{\sigma}$ from the volatility model are presented respectively in the upper and lower parts, for three persistency levels of latent factor $\alpha = 0.4, 0.8, 1$, in each of its three columns. Each of the six individual graphs plots the bias from the endogenous (red solid line) and exogenous (blue dashed line) regime switching models across different levels of endogeneity parameter $\rho$ on the horizontal axis. Presented in the same manner on the right panel are the bias in the ML estimates $\underline{\hat{\mu}}, \hat{\overline{\mu}}$ and $\hat{\gamma}$ of low and high mean levels and AR coefficient of observed time series, $\underline{\mu}, \overline{\mu}$ and $\gamma$, estimated from the mean model. There are 9 individual graphs covering the bias in three estimates for three persistency levels of the latent factor.

the horizontal axis. Similarly the right panel of Figure 3 presents the bias in the maximum likelihood estimates $\underline{\hat{\mu}}, \hat{\overline{\mu}}$ and $\hat{\gamma}$ of low and high mean levels and AR coefficient of observed time series from the mean model. There are 9 individual graphs covering the bias in three estimates $\underline{\hat{\mu}}, \hat{\overline{\mu}}$ and $\hat{\gamma}$ for three persistency levels $\alpha = 0.4, 0.8, 1$ of the latent factor.

The endogeneity in regime switching, if ignored, may yield substantial bias in the estimates of model parameters. This turns out to be true for both mean and volatility models, though the deleterious effect of the neglected endogeneity is relatively bigger in the mean model. The magnitude of bias tends to be larger when $\alpha$ is away from unity and the latent autoregressive factor is more stationary. For example, when $\alpha = 0.4$ and $\rho = -0.7$, the bias of the estimates $\underline{\hat{\mu}}, \hat{\overline{\mu}}$, and $\hat{\gamma}$ in the mean model are respectively 38.7%, −10.8%, and −86.7%. If, however, $\alpha$ is close to unity, the neglected endogeneity does not appear to yield any substantial bias. In fact, when $\alpha = 1$ and the latent factor becomes a random walk, the effect of endogeneity on parameter estimates in both mean and volatility models becomes insignificant. In all cases, however, the magnitude of bias becomes larger as $|\rho|$ gets bigger and the degree of endogeneity increases. Though we do not report the details

Figure 4: Efficiency Gain from Endogeneity




Notes: Respectively presented in the left and right panels of Figure 4 are the standard errors of the ML estimates of the parameters in our endogenous volatility and mean switching models. The 6 graphs on the left and 9 graphs on the right panels present the standard errors of ML estimates from the volatility and mean models in the exactly the same manner as in Figure 3.

to save space, our simulations show that the inferred probabilities of latent states are also affected seriously if the endogeneity in regime switching is not properly taken care of.

Not only can the presence of endogeneity give us a pitfall leading to a substantial bias in parameter estimates, but also an opportunity to improve the precision of parameter estimates in markov switching models. In fact, in the endogenous regime switching model, additional information on the state process $(s_t)$ is provided by the observed time series $(y_t)$. Note that the transition of the state process $(s_t)$ in our models is determined by lags of $(y_t)$ as well as lags of $(s_t)$, and therefore we have an additional channel for the information in $(y_t)$ to be accumulated in the likelihood function. This is not the case if we let $\rho = 0$ as in the conventional markov switching model that does not allow for the presence of endogeneity. The simulation results in Figure 4 show that the presence of endogeneity in regime switching indeed improves the efficiency of parameter estimates, if accounted for properly as in our endogenous models. The standard errors of ML estimates of the parameters in our volatility and mean models are presented respectively in the left and right panels of Figure 4 in exactly the same manner as in Figure 3.

As shown in Figure 4, the efficiency gain from the presence of endogeneity in regime switching can be quite substantial. This is equally true for both mean and volatility models. For instance, if we set $\alpha = 0.8$, the standard deviations of the estimators $\hat{\bar{\mu}}$ and $\underline{\hat{\sigma}}$ from our

mean and volatility models having endogenous regime switching with $\rho = -0.9$ decrease by approximately 24% and 22%, respectively, if compared with the models having exogenous regime switching with $\rho = 0$. Of course, the presence of endogeneity yields efficiency gain, only when it is properly taken into account. If the conventional markov switching model is used, the presence of endogeneity in most cases has a negative effect on the standard deviations of parameter estimators.

In general, the standard deviations of parameter estimators are greatly reduced in both mean and volatility models if we have endogeneity in regime switching, as long as $|\alpha| < 1$ and the latent factor is stationary. Naturally, the efficiency gain increases as $|\rho|$ gets large and the degree of endogeneity increases. On the other hand, when the latent factor is nonstationary with $\alpha = 1$, the standard errors of parameter estimators from the endogenous model remain more or less constant across $\rho$, showing little or no sign of efficiency gain. This may be due to the fact that switching occurs rarely when the latent factor is highly persistent, reducing the opportunity for additional information contained in the observed time series on the switching to play a positive role.[13]

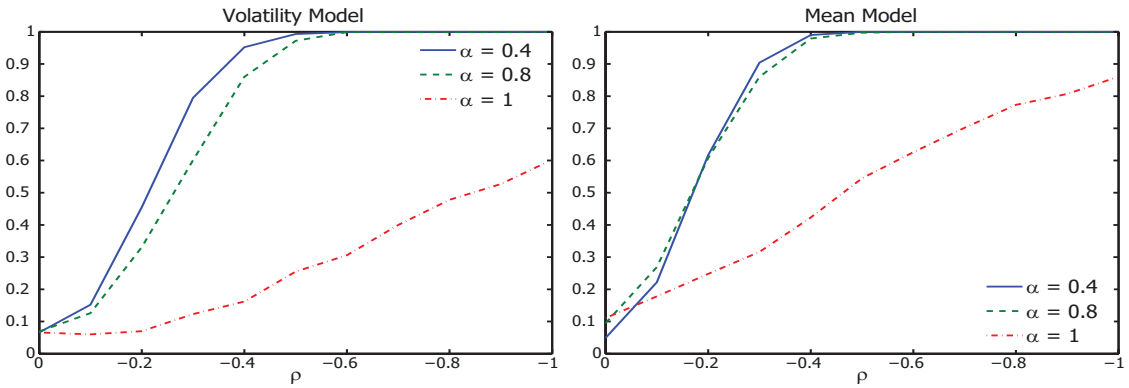
Finally, we consider testing for the presence of endogeneity in regime switching models based on the likelihood ratio test given by

$$2(\ell(\hat{\theta}) - \ell(\tilde{\theta})), \tag{26}$$

where $\ell$ signifies the log-likelihood function and the parameter $\theta$ with tilde and hat denote their maximum likelihood estimates with and without the no endogeneity restriction, $\rho = 0$. The likelihood ratio test has a chi-square limit distribution with one degrees of freedom. Presented in the left and right panels of Figure 5 are the power functions of the likelihood ratio test computed from the simulated volatility and mean switching models for three different levels of persistency in the latent factor measured by its AR coefficient $\alpha = 0.4, 0, 8, 1$. For the stationary regime switching model with $\alpha = 0.4$ or 0.8, the test is very powerful with the actual power increasing rapidly as the value of endogeneity parameter $|\rho|$ gets large. Under the null hypothesis of no endogeneity, the test has good size properties overall in the volatility model, but it tends to over-reject in the mean model when the sample size is only moderately large as the latent factor becomes more persistent. Though we do not report the details, the size distortion disappears as sample size increases. In contrast, the test does not work well when $\alpha = 1$ and the latent factor becomes nonstationary. In this

[13] On the average, the regime change occurs 160, 100, and 15 times out of 500, respectively, for three pairs of $(\alpha, \tau) = (0.4, 0.5), (0.8, 0.7)$, and $(1, 9.63)$ we consider in our simulations. This clearly shows a rapid decline of regime change frequency as the value of AR coefficient $\alpha$ gets closer to 1 and the latent factor becomes a random work.

Figure 5: Power Function of LR Test for Endogeneity




Notes: The left and right hand side graphs of Figure 5 present the power functions of the likelihood ratio test computed respectively from the volatility and mean switching models for three different levels of persistency in the latent factor $(w_t)$ measured by its AR coefficient $\alpha = 0.4, 0, 8, 1$.

case, the power function increases very slowly as $|\rho|$ gets large, and tends to over-reject in the mean model. The overall performance of the test in the nonstationary case is relatively much worse than the stationary cases.

# 5 Empirical Illustrations

To illustrate our approach empirically, we analyze the excess market returns using the volatility model with regime switching (24) studied in our earlier simulations, and the US GDP growth rates by a mean model with regime switching similar to (25) which is also examined in our simulations.

## 5.1 Stock Market Return Volatilities

For market returns, we consider the returns on NYSE/AMEX index from the Center for Research in Security Prices (CRSP). Specifically, we use the monthly series of value-weighted stock returns including dividend for the period from January 1926 to December 2012, along with the information on their quote date. For our analysis, we use the demeaned excess market returns $(y_t)$ to fit the volatility model in (24).[14] Table 1 reports the estimation results for the excess market return volatility model with regime switching for two sample

[14]To compute monthly excess return, we first obtain monthly risk free rate of return by continuously compounding daily risk free rate between the quotation dates. The number of days between quotation dates ranges from 28 to 33. CRSP provides monthly series of annualized yield to maturity (TMYTM), which is constructed from nominal price of three month treasury bill. Thus we obtain the yield to maturity at monthly frequency by first converting the annual yield to maturity to daily (TMYLD) by the conversion

periods: the full sample period (1926-2012) and the recent subsample period (1990-2012). We choose this subsample period to relate the extracted latent volatility factor obtained from our endogenous switching model with one of the most widely used volatility index VIX which is available only from 1990.

For the ML estimation of parameters and transition probabilities in the volatility switching model, we use our modified markov switching filter together with the numerical optimization method including the commonly used BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm. When likelihood function is highly nonlinear as in our switching models, it is well known that numerical optimization often fails to find global maximum and ends up with local maximum depending on the choice of initial values. However, we notice that in our switching models we can effectively profile out all the parameters but $\alpha$ and $\rho$, and once we fix $\alpha$ and $\rho$, our numerical maximization procedure successfully finds global maximum regardless of the choices on initial values.

For the exogenous model, we analyze profile likelihood in terms of $\alpha$ only, which is obtained by maximizing the likelihood for each value of $\alpha$ over all the other parameter values, and then maximize the profile likelihood over all $\alpha$ values considered. The left hand side graph of Figure 6 shows the profile likelihood function of $\alpha$ obtained from the exogenous switching model when we fix $\alpha$ at a value from 0.1 to 0.9 with 0.1 increment.[15] For the endogenous switching model, we consider profile likelihood in terms of both $\alpha$ and $\rho$, and maximize it over all pairs of $\alpha$ and $\rho$ considered. The graph on the right hand side of Figure 6 presents the profile likelihood surface plot obtained from our endogenous volatility switching model when we fix $\rho$ as well as $\alpha$. To obtain the surface plot, we consider 171 combinations of $\alpha$ and $\rho$ from setting $\alpha$ at a value from 0.1 to 0.9 and $\rho$ at a value from $-0.9$ to 0.9 with 0.1 increment.[16] We may see clearly from Figure 6 that the profile likelihood

formula CRSP provides as

$$TMYLD_t = \frac{1}{365}\frac{1}{100}TMYTM_t,$$

and then continuously compounding this daily yield to obtain the monthly yield as

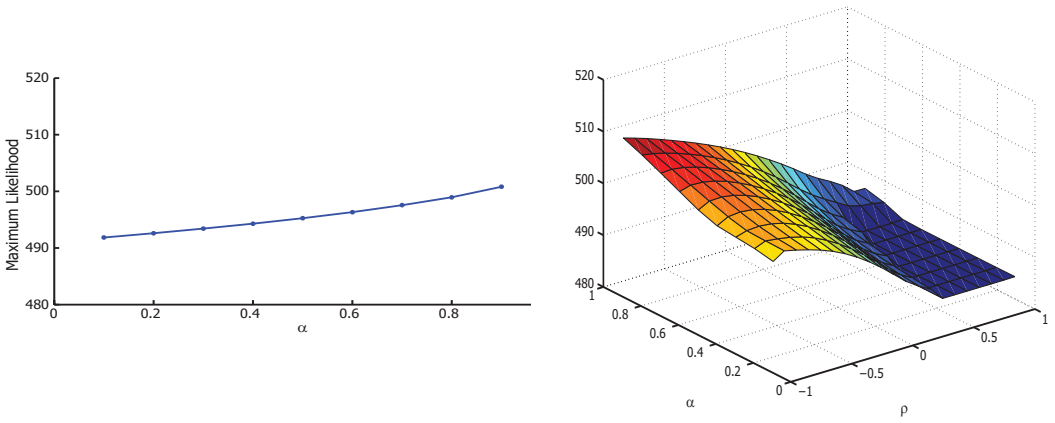
$$e^{TMYLD_{t-1}\times N_t} - 1,$$

where $N_t$ is the number of days between the quote date for the current month and the quote date $MCALDT_{t-1}$ for the previous month. Finally, we obtain monthly excess market return data by subtracting the above monthly risk free rate from the market return.

[15] For each fixed value of $\alpha$, we try 12 to 60 different initial value combinations for all the other parameters to find the maximum value for the profile likelihood function. However, regardless of the choices on initial values, the numerical maximization finds unique maximum once we fix $\alpha$ and the global maximum would be found around $\alpha = 0.9$.

[16] For each of these combinations of $\alpha$ and $\rho$, we try 7 to 35 different sets of initial values of all the other parameters to find the maximum. Our numerical method finds local maxima for some initial value combinations when $\rho$ is positive. Those local maxima correspond to the flat area in the surface plot on the right side of Figure 6, which is well below the global maximum. When $\rho$ is negative, on the other hand, our

Figure 6: Profile Likelihood for Volatility Switching Model




Notes: Figure 6 presents profile likelihood functions from volatility switching models for the period 1990 to 2012. The left hand side graph shows the profile likelihood of $\alpha$ from the exogenous regime switching model, which is obtained by concentrating out all the other parameters. Similarly, the right hand side graph shows the surface plot of joint profile likelihood of $\alpha$ and $\rho$ from our endogenous regime switching model obtained by profiling out all the other remaining parameters.

increases monotonically until $\alpha$ and $\rho$ reach 0.9 and $-0.9$ respectively.

As reported in Table 1, the volatility level estimates and their standard errors given underneath in parenthesis from our endogenous regime switching model are similar to those from the exogenous model in the full sample period, but they are bigger than those from their exogenous counterpart for the recent subsample. However, the estimates for the endogeneity parameter $\rho$ are quite substantial in both samples, $-0.970$ for the full sample and $-0.999$ for the subsample,[17] providing an ample evidence for the presence of endogeneity in regime switching in the market volatility. Table 1 also shows that the maximum value of the log-likelihood function from the endogenous switching model is larger than that from the exogenous switching model that ignores endogeneity. We formally test for the presence of endogeneity in regime switching using the usual likelihood ratio test given in (26). In both sample periods, we reject the null of no endogeneity at 1% significance level, as reported in the bottom line of Table 1. For the full sample period 1926-2012, the estimates of $\alpha$ and

numerical method seems to always find unique maximum for any $\alpha$ value.

[17] Here our estimate of $\rho$ for the subsample is virtually identical to $-1$, in which case the current shock to the stock returns would completely determine the level of latent factor and the state for their volatility regime in the next period. The true value of $\rho$ may be $-1$ or very close to $-1$. However, it may also be spuriously obtained by fitting a model with regime switching in volatility for a process generated without any changing regimes, whose probabilities we found by simulation to be around 30%. Indeed, when $\rho$ is restricted to $-1$, we have essentially the same estimates for $\underline{\sigma}$ and $\overline{\sigma}$ with almost identical maximum likelihood value.

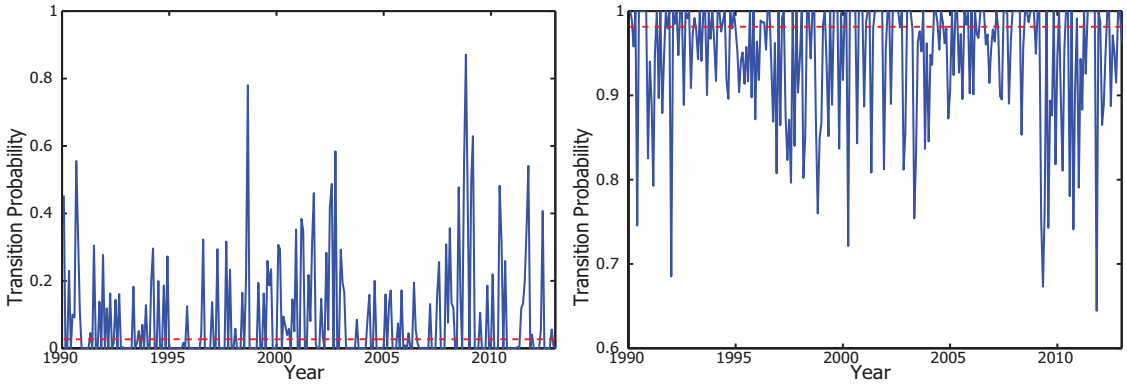
Table 1: Estimation Results for Volatility Model

| Sample Periods | 1926-2012 | | 1990-2012 | |
|---|---|---|---|---|
| Endogeneity | Ignored | Allowed | Ignored | Allowed |
| $\underline{\sigma}$ | 0.039 | 0.038 | 0.022 | 0.025 |
| | (0.001) | (0.001) | (0.002) | (0.004) |
| $\overline{\sigma}$ | 0.115 | 0.115 | 0.050 | 0.055 |
| | (0.009) | (0.009) | (0.003) | (0.008) |
| $\rho$ | | –0.970 | | –0.999 |
| | | (0.086) | | (0.010) |
| log-likelihood | 1742.536 | 1748.180 | 507.700 | 511.273 |
| $p$-value (LR test for $\rho = 0$) | | 0.001 | | 0.008 |

$\tau$ in the exogenous model are 0.994 (0.004) and 11.375 (4.472), from which we have 0.991 and 0.928 as the estimates of the low-to-low and high-to-high transition probabilities. In the endogenous model, we obtain 0.986 (0.009) and 7.385 (2.567) for $\alpha$ and $\tau$. On the other hand, for the recent subsample period 1990-2012, we find 0.997 (0.003) and $-3.212$ (7.055) as the estimates of $\alpha$ and $\tau$ in the exogenous model, which yield 0.973 and 0.981 for the low-to-low and high-to-high transition probabilities. For the endogenous model, we have 0.979 (0.044) and 0.685 (2.114) for $\alpha$ and $\tau$.

What is most clearly seen from Figure 7 is the striking difference in the time series plots of the transition probabilities estimated from the exogenous and endogenous volatility regime switching models. The transition probability estimated by the exogenous model is constant over the entire sample period, while the corresponding transition probabilities estimated by the endogenous model vary over time, and depend upon the lagged value of excess market return $y_{t-1}$ as well as the realized value of the previous state $s_{t-1}$. This point is clearly demonstrated in the left hand side graph of Figure 7 which presents the transition probability from low volatility regime at $t-1$ to high volatility regime at $t$ estimated by the exogenous and endogenous switching models. This low to high transition probability is estimated to be 2.7% throughout the entire sample period by the exogenous model, while in contrast the corresponding transition probabilities estimated by the endogenous model vary over time with the realized values of lagged excess market return. It shows in particular that the transition probabilities have been changing drastically, and reach as high as 87.1% at a time, which is 32 times bigger than its counterpart from the exogenous regime switching model. The right hand side graph of Figure 7 similarly illustrates the same point with the transition probability from high volatility state at $t-1$ to high volatility state at $t$ by the exogenous and endogenous switching models.

Figure 7: Estimated Transition Probability from Volatility Model

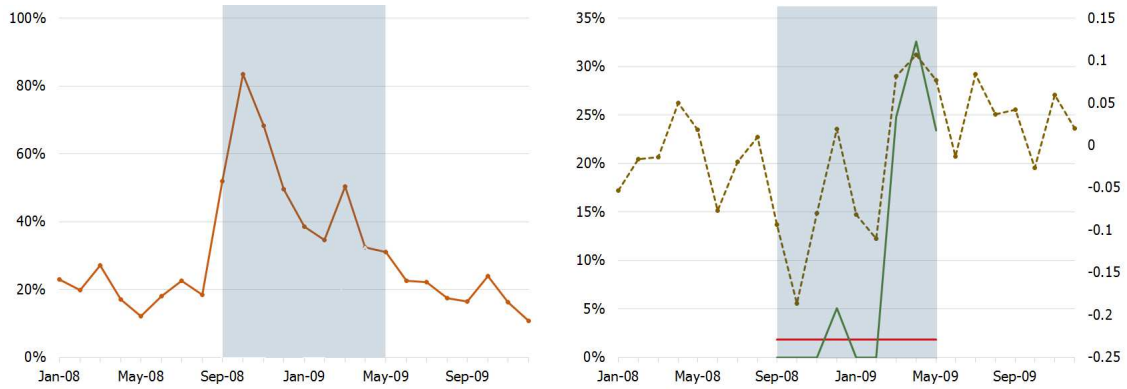



Notes: Figure 7 presents the transition probabilities from volatility model. The left hand side graph shows the transition probability from low to high volatility state: the blue solid line refers to $\mathbb{P}(s_t = 1|s_{t-1} = 0, y_{t-1})$ in our endogenous regime switching model, while the red dashed line corresponds to $\mathbb{P}(s_t = 1|s_{t-1} = 0)$ in the exogenous regime switching model. Similarly, the right hand side graph shows the transition probabilities staying at high volatility state.

The variation of the transition probabilities in the endogenous switching model is truly meaningful when the regime is about to change. Figure 8 illustrates convincingly that the time varying transition probabilities from the endogenous model can indeed produce more realistic assessment for the likelihood of moving into a low volatility regime from a high volatility regime. We first measure the monthly market volatility by sum of squared daily returns on NYSE/AMEX index during the recent 2008-2009 financial crisis to see if there is a period that can be unarguably defined as a high volatility regime. The left hand side graph of Figure 8 presents the time series of annualized monthly volatility. The time series of monthly stock returns is also presented on the graph on the right. The volatility increased dramatically in September 2008 when Lehman Brothers declared bankruptcy and stayed high until May 2009. We consider this period as a high volatility regime. The volatility level in each month during this high volatility regime from September 2008 to May 2009 is at least twice higher than the average volatility computed over the 32-month period ending at the start of this high volatility regime in September 2008. The shaded areas on both graphs of Figure 8 indicate this high volatility period.

The high to low transition probabilities during the high volatility regime are presented on the right hand side graph of Figure 8, where the green line signifies the time varying transition probability $\mathbb{P}(s_t = 0|s_{t-1} = 1, y_{t-1})$ estimated from our endogenous regime switching model, while the red line corresponds to the constant transition probability $\mathbb{P}(s_t = 0|s_{t-1} = 1)$ obtained from the exogenous switching model. Indeed the transition

Figure 8: Transition Probabilities for Recent Financial Crisis Period




Notes: The high to low transition probabilities during the high volatility regime from September 2008 to May 2009 are presented on the right hand side graph of Figure 8, where the green line signifies the time varying transition probability $\mathbb{P}(s_t = 0|s_{t-1} = 1, y_{t-1})$ estimated from our endogenous regime switching model, while the red line corresponds to the constant transition probability $\mathbb{P}(s_t = 0|s_{t-1} = 1)$ obtained from the exogenous switching model. The solid line on the left and the dashed line on the right graphs respectively present the time series of annualized monthly volatility and the monthly NYSE/AMEX index returns. The shaded areas on both graphs of Figure 8 indicate the high volatility regime.

probability estimated by the exogenous model stays constant for the entire duration of the high volatility regime, which is in sharp contrast to the substantially time varying transition probabilities obtained from the endogenous model. Notice that the high to low transition probability from our endogenous model is smaller than that from the exogenous model at the beginning of the high volatility regime; however, it goes up drastically toward the end of the high volatility regime, which coincides not surprisingly with the rapid recovery of the stock market in early spring of 2009. The time varying transition probability produced from our endogenous switching model therefore represents the reality we observe much better, which in turn explains the higher precision and efficiency of the parameter estimates obtained from our endogenous regime switching model observed in our earlier simulation studies.

To see how well our endogenous volatility switching model can explain the current state of market volatility, we compare the sample paths of the extracted latent factor with that of VIX, a popular measure for implied market volatility, over the subsample period 1990 to 2012 where VIX is available. See Figure 13 in Appendix, which presents the sample path of the extracted latent factor along with that of the CBOE (The Chicago Board Options Exchange) volatility index VIX for the period 1990-2012. VIX stayed relatively high during 1998-2004 and 2008 periods indicating that the volatility was high during those periods.

As shown in Figure 13, the extracted latent factor obtained from our endogenous volatility model also stays relatively high, moving closely with VIX during those high volatility periods. VIX has been used as a gauge for "fear factor" or an indicator for the overall risk level of market. Therefore the extracted latent factor from our volatility model may be considered as an alternative measure which can play the similar role played by VIX.

We also compute for each period the inferred probability of being in the high volatility regime using our endogenous volatility switching model as well as the conventional exogenous switching model. They are presented in Figure 14 in Appendix. Overall the time series of the high volatility probabilities computed from both endogenous and exogenous switching models are similar. There are however some noticeable differences. In general, those obtained from our endogenous model tend to fluctuate more over time compared to those from its exogenous counterpart. Moreover, it is interesting to note that the probability of being in high volatility regime computed from our endogenous model is exactly at 1 during the 2008 financial crisis, while that from the exogenous model does not quite reach 1, albeit close to 1, which demonstrates the potential of our model to more precisely estimate the probability of the financial market becoming unstable. Therefore we may also say that our endogenous switching model can be used as a warning system for detecting unusual unstability in financial market.

### 5.2 GDP Growth Rates

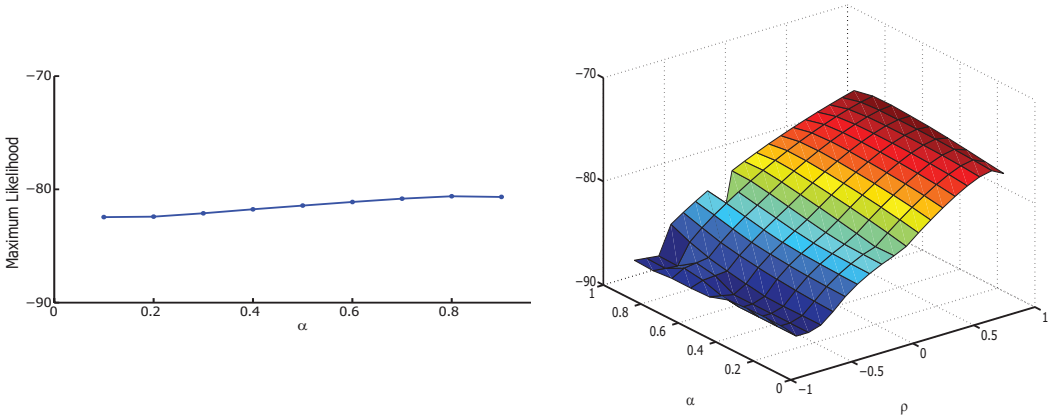
In this section, we investigate regime switching behavior of the US real GDP growth rates constructed from the seasonally adjusted quarterly real GDP series for the period 1952:Q1-2012:Q4.[18] As in Hamilton (1989), we model the real GDP growth rates ($y_t$) as an AR(4) process similarly as in (25). Since there seems to be a structural break in the postwar U.S. real GDP growth rates in 1984:Q1, as noted in Kim and Nelson (1999), we consider two sample periods: the earlier sample period covering 1952:Q1-1984:Q4, and the more recent sample period covering 1984:Q1-2012:Q4. Kim and Nelson (1999) provides an empirical study of the regime switching model for US growth rates considered in Hamilton (1989) for the earlier sample period. We use the same data used in their study and compare our results with theirs.[19]

We estimate the mean switching model for the GDP growth rates using our new modified markov switching filter along with BFGS method. To ensure that we find the global maximum, we look at the profile likelihood as done in the volatility model for the market

[18] Source: Bureau of Economic Analysis, US Department of Commerce. The growth rate of real GDP is calculated as the first difference of log real GDP.

[19] We use the data provided in the website for Kim and Nelson (1999).

Figure 9: Profile Likelihood for Mean Switching Model




Notes: Figure 9 presents the profile likelihood from mean switching model for the period 1984-2012. The left hand side graph shows the profile likelihood from the exogenous regime switching model which is obtained by profiling out all the parameters but $\alpha$ here. Similarly, the graph on the right presents the profile likelihood surface plot from our endogenous regime switching model obtained by profiling out all the parameters but $\alpha$ and $\rho$.

excess return considered earlier. Figure 9 shows the profile likelihood for the recent sample period from 1984:Q1 to 2012:Q4. The left hand side graph in Figure 9 shows the plot of the maximum values of the likelihood function from the exogenous model when we fix $\alpha$ at a value from 0.1 to 0.9 with 0.1 increment.[20] The profile likelihood function of $\alpha$ on the left side of Figure 9 clearly shows that the global maximum for the exogenous switching mean model is reached when $\alpha$ is around 0.8. The right hand side graph in Figure 9 presents the profile likelihood surface from our endogenous mean switching model when we fix $\alpha$ at a value from 0.1 to 0.9 and $\rho$ from $-0.9$ to 0.9 with 0.1 increment.[21] It is clearly demonstrated in Figure 9 that the global maximum is found around $(\alpha, \rho) = (0.8, 0.9)$ for the endogenous

[20] For each fixed value of $\alpha$, we try 60 to 300 different sets of initial values for all the other parameters to obtain maximized likelihood. For most of the cases, they converge to unique maximum once we fix $\alpha$. Unlike in the volatility model, we get local maxima for some of the initial values in the mean model. However, they do not cause any serious problem, because they are substantially away from the maximum of the profile likelihood function. For instance, when $\alpha = 0.7, 0.8$ or 0.9, we find the maximum and one other local maximum, which is far blow the maximum for each of these cases, and this happens exclusively when we set the initial value for the high to high transition probability at 0.5. We like to know that we always have unique local maximum whenever we have local maxima. We also have a local maximum when $\alpha = 0.2$, in which case the maximum likelihood value is close to the local maximum. Even in this case, however, our filter still manages to successfully identify the global maximum.

[21] For each combination of $\alpha$ and $\rho$, we try one to 15 different sets of initial values of the remaining parameters for profile likelihood maximization. Though our numerical methods find local maxima for profile likelihood for some initial value combinations, it always finds unique maximum when $\alpha$ is close to one and $\rho$ is close to negative one.

Table 2: Maximum Likelihood Estimates for Hamilton (1989) Model

| Sample Periods | 1952-1984 | | 1984-2012 | |
|---|---|---|---|---|
| Endogeneity | Ignored | Allowed | Ignored | Allowed |
| $\underline{\mu}$ | –0.165 | –0.083 | –0.854 | –0.758 |
| | (0.219) | (0.161) | (0.298) | (0.311) |
| $\overline{\mu}$ | 1.144 | 1.212 | 0.710 | 0.705 |
| | (0.113) | (0.095) | (0.092) | (0.085) |
| $\gamma_1$ | 0.068 | 0.147 | 0.154 | 0.169 |
| | (0.123) | (0.104) | (0.105) | (0.105) |
| $\gamma_2$ | –0.015 | 0.044 | 0.350 | 0.340 |
| | (0.112) | (0.096) | (0.105) | (0.103) |
| $\gamma_3$ | –0.175 | –0.260 | –0.036 | –0.076 |
| | (0.108) | (0.090) | (0.106) | (0.128) |
| $\gamma_4$ | –0.097 | –0.067 | 0.043 | 0.049 |
| | (0.104) | (0.095) | (0.103) | (0.112) |
| $\sigma$ | 0.794 | 0.784 | 0.455 | 0.452 |
| | (0.065) | (0.057) | (0.034) | (0.032) |
| $\rho$ | | –0.923 | | 0.999 |
| | | (0.151) | | (0.012) |
| log-likelihood | –173.420 | –169.824 | –80.584 | –76.447 |
| $p$-value | | 0.007 | | 0.004 |

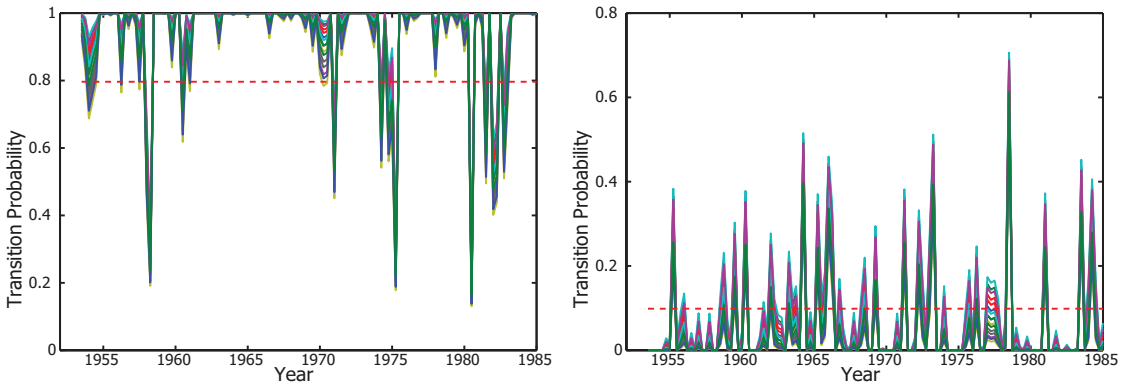
models.

Table 2 presents the estimation results for the two sample periods we consider.[22] The ML estimates obtained from the exogenous model with the constraint $\rho = 0$ imposed and those from the endogenous model with no constraint on $\rho$ are generally different as expected, though the estimates for some parameters such as $\overline{\mu}$ and $\sigma$ are similar. The difference between the estimates from two sample periods is bigger than that from two models, exogenous and endogenous. Especially, the estimates of $\underline{\mu}, \overline{\mu}$, and $\sigma$ from two sample periods are quite different, which may be used as a supporting evidence for presence of a structural break in the US GDP series. It is also interesting to note that the ML estimate for the correlation coefficient $\rho$ measuring the degree of endogeneity for the earlier sample period is quite large and negative, $-0.923$, which is drastically different from the value, 0.999, estimated from the recent sample period.[23] For the earlier sample period 1952-1984, the estimates

[22] Again, the standard errors are presented in parenthesis.

[23] Like one of our estimates for $\rho$ in our volatility model, here we also have an extreme case. The estimated $\rho$ for the recent sample period is very close to 1, which suggests that the GDP growth rates evolve with the mean regimes determined almost entirely by the shocks to themselves. As discussed, this may be spuriously obtained by fitting a model with regime switching in mean for a process generated without any changing

Figure 10: Estimated Transition Probabilities for Mean Model




Notes: Figure 10 presents the transition probabilities from the mean model for the US GDP growth rates. The left hand side graph shows the sample paths of the 17 transition probabilities of staying at low mean state: the 16 solid time varying lines represent transition probabilities obtained from our endogenous switching model by computing $\mathbb{P}(s_t = 0|s_{t-1} = 0, s_{t-2} = i, s_{t-3} = j, s_{t-4} = k, s_{t-5} = \ell, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5})$ for all 16 possible combinations of $i, j, k, \ell = 0, 1$, while the one red dashed straight line represents the probability of staying at low mean state $\mathbb{P}(s_t = 0|s_{t-1} = 0)$ obtained from the exogenous model. In the same way, the right hand side graph shows the transition probabilities from high to low mean state.

of $\alpha$ and $\tau$ in the exogenous model are 0.895 (0.077) and $-1.009$ (0.773), from which we obtain the estimates for the low-to-low and high-to-high transition probabilities 0.796 and 0.901. In the endogenous model, we have 0.927 (0.041) and $-0.758$ (0.883) for $\alpha$ and $\tau$. On the other hand, for the recent sample period 1984-2012, the estimates for $\alpha$ and $\rho$ in the exogenous model are 0.842 (0.162) and $-3.282$ (1.489), which yield 0.526 and 0.981 for the low-to-low and high-to-high transition probabilities. In the endogenous model, we have 0.809 (0.145) and $-2.782$ (1.144) for $\alpha$ and $\tau$. Moreover, the maximum value of the log-likelihood function from the unrestricted endogenous model is larger than that from the restricted exogenous model with $\rho = 0$ imposed, and consequently the null of no endogeneity is decisively rejected by the usual likelihood ratio test given in (26) at 1% significance level for both sample periods.

The estimated transition probabilities are presented in Figure 10. Since $(s_t, y_t)$ is jointly a fifth-order markov process, the transition probabilities at time $t$ depend on $s_{t-1}, ..., s_{t-5}$ as well as $y_{t-1}, ..., y_{t-5}$. The graph on the left hand side of Figure 10 shows the transition probabilities from low mean state at $t-1$ to low mean state at $t$. There are 17 lines in the graph. The 16 solid lines represent the sample paths of the 16 transition probabilities

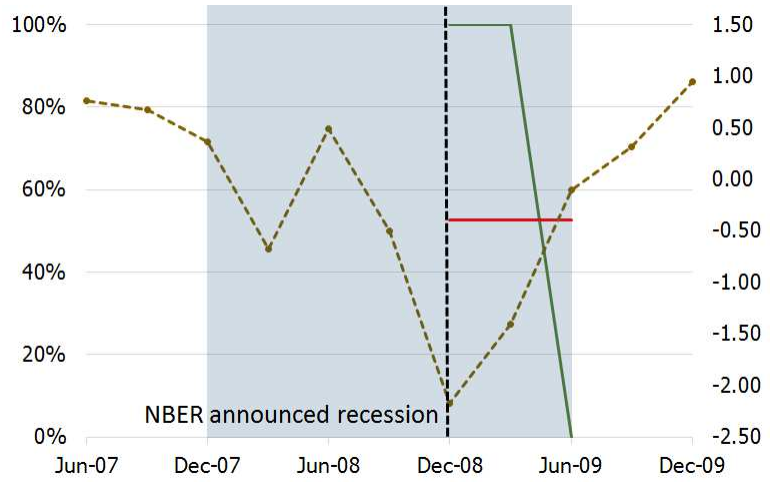
regimes, probabilities of this happening, we found by simulation, is to be around 10%. In fact, we obtain very similar estimates and likelihood values, when the model is re-estimated with the restriction $\rho = 1$ imposed.

obtained from our endogenous regime switching model for each of the 16 possible realizations of the four lagged state variables, $s_{t-2}, s_{t-3}, s_{t-4}$, and $s_{t-5}$. Note that each of the four lagged state variables $s_{t-2}, s_{t-3}, s_{t-4}$, and $s_{t-5}$ takes a value either 0 or 1, giving 16 possibilities for their joint realizations. We therefore calculate the transition probability from low state at $t-1$ to low state at $t$, i.e., $\mathbb{P}(s_t = 0 | s_{t-1} = 0, s_{t-2} = i, s_{t-3} = j, s_{t-4} = k, s_{t-5} = \ell, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5})$ for all 16 possible combinations of $i, j, k, \ell = 0, 1$. The one dashed line represents the corresponding probability of staying at low mean state obtained from the exogenous model. Similarly, the graph on the right hand side shows the sample paths of the 17 transition probabilities from high to low mean state, 16 solid time varying lines from the endogenous model and one dashed straight line from the exogenous model. The most salient feature from the two figures presented in Figure 10 is that the estimated transition probabilities estimated by our endogenous regime switching model are drastically different from the one obtained from its exogenous counterpart. The time varying transition probabilities estimated from our endogenous model are indeed much more sensible and realistic.

Figure 11 presents the transition probabilities from low mean regime to low mean regime plotted along with the US GDP growth rates. NBER announced on December 1, 2008 that a recession began in December 2007 after the December 2007 peak which defines the turning point from expansion to recession, and it announced on October 21, 2010 that the recession ended in June 2009. By the time of the official announcement by NBER on December 1, 2008, everyone is presumed to know that we were in recession. Using this information, we calculate the transition probability from low to low mean regime. The red solid line is the corresponding transition probability from exogenous switching model. It is constant over time. However the green solid line signifying the transition probability from our endogenous model drastically changes over time. Note that the transition probability in our endogenous model is determined not only by previous states but also by the lagged values of the GDP growth rates. The endogenous switching model exploits the information from the past values of the observed time series to update the transition probability. Therefore, when the observed GDP growth rate is low, the transition probability from low to low regime is as high as 100%, but this transition probability sharply declines to virtually zero when we update our information with the high realized values of GDP growth rates.

We also extract the latent factor determining the states from our endogenous mean switching model, and compare it with the recession periods identified by NBER. See Figure 15 in Appendix, which presents the sample path of the extracted latent mean switching factor and NBER recession periods during the two sample periods we consider, 1952-1984 and 1984-2012. In both sample periods, we can see clearly that the trough times of the

Figure 11: Transition Probabilities During Recent Recession Period




Notes: Figure 11 presents the transition probabilities from the mean switching model for the most recent US recession period, 2007-2009. The shaded area indicates the recession period which started on 2007:Q4 and ended on 2009:Q2, and the dashed vertical line marks December 1, 2008 when NBER announced the recession began on December 2007. The solid green (red) line signifies the low to low transition probability estimated by the endogenous (exogenous) switching model. The dashed blue line plotted on the right vertical axis represents the US real GDP growth rates.

extracted latent factor coincide with NBER recession periods indicated by shaded areas in the graphs. It is not surprising and indeed well expected from our model, since a low value of the latent factor will likely result in a low mean state. Therefore we may use the extracted latent factor from our endogenous mean switching model as a potential indicator for business cycle.

Finally, we compute the inferred probability that we are in the recession regime, which are presented in Figures 16-17 for both sample periods, 1952-1984 and 1984-2012. Both endogenous and exogenous models produce high recession probabilities when the *levels* of growth rates become negative. However, our endogenous model has an additional channel via transition probabilities to reflect the *changes* in the observed growth rates, which its exogenous counter part does not. The extra ability of our endogenous model to bring in the information on the changes in the growth rates can indeed result in strikingly different recession probabilities, which is drastically demonstrated during the late 1970's shortly before the 1981 recession formally announced by NBER. The GDP growth rates sharply went up and peaked in the second quarter of 1978, and then quickly declined and became negative in the second quarter of 1980. The transition probabilities in our endogenous model depend on the past values of the observed growth rates, and therefore are responsive to the

movements of the growth rages. The is why we see much higher values of the recession probabilities computed from our endogenous model compared to those obtained from the exogenous model during the period prior to the 1981 recession. Our model could see the downward movements in the growth rates during this period, and accordingly increased the recession probabilities, even before the growth rates become negative. There is no way, however, for the exogenous model to update the recession probabilities upward even when it was evident that the growth rates were sinking, and it has to wait until after the levels of growth rates become negative to raise the recession probabilities.

# 6 Conclusions

In the paper, we propose a new regime switching model based on an autoregressive latent factor. As we demonstrate in the paper, our approach has several clear advantages over the conventional regime switching model. Most importantly, we may allow for endogeneity in regime switching. The endogeneity we introduce by using our approach is well structured. It models the effect of a shock to the observed time series in a very natural manner. In the mean model with regime switching, the presence of endogeneity implies that the mean reversion may occur in two different levels: one at the level of reversion of the observed time series to its state dependent mean, and the other at the level of movement of the state dependent mean to offset the effect of a shock. In the volatility model with regime switching, on the other hand, the presence of endogeneity means the leverage effect. Furthermore, our regime switching model becomes observationally equivalent to the conventional markov switching model, if the endogeneity of regime switching is not present. Finally, our approach allows the transition of the state process to be nonstationary and strongly persistent.

The empirical evidence for the presence of endogeneity in regime switching appears to be very strong and unambiguous. This implies, in particular, that it is worthwhile to refit any of the previously fitted conventional markov switching models, allowing for endogeneity in regime switching. Our extensive simulations make it clear that neglecting endogeneity in regime switching not only incurs a substantial bias in the estimates of model parameters, but also does it lead to significant information loss. If endogeneity in the regime switching is ignored, the variability of parameter estimates sharply increases and consequently the inferred probabilities of the latent states become much less precise. This is because the endogeneity in regime switching creates an important additional link between the latent states and observed time series, and therefore, the information that can be channeled through this link cannot be exploited if the endogeneity is ignored. The additional information that we may extract from this new link is certainly more valuable in a markov switching model,

since the state process playing such a critical role in the model is latent and must be inferred from a single observable time series.

## References

Bazzi, M., Blasques, F., Koopman, S. J., Lucas, A., 2014. Time varying transition probabilities for Markov regime switching models, tinbergen Institute Discussion Paper TI 2014-072/III.

Chib, S., 1996. Calculating posterior distributions and modal estimates in Markov mixture models. Journal of Econometrics 75, 79–97.

Chib, S., Dueker, M., 2004. Non-Markovian regime switching with endogenous states and time-varying state strengths, working Paper 2008-31, Federal Reserve Bank of St. Louis.

Cho, J. S., White, H., 2007. Testing for regime switching. Econometrica 75, 1671–1720.

Diebold, F., Lee, J.-H., Weinbach, G., 1994. Regime switching with time-varying transition probabilities. In: Hargreaves, C. (Ed.), Nonstationary Time Series Analysis and Cointegration. Oxford University Press, Oxford, UK, pp. 283–302.

Garcia, R., 1998. Asymptotic null distribution of the likelihood ratio test in markov switching models. International Economic Review 39, 763–788.

Hamilton, J., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–384.

Hamilton, J., 1996. Specification testing in markov-switching time-series models. Journal of Econometrics 70, 127–157.

Hansen, B. E., 1992. The likelihood ratio test under non-standard conditions. Journal of Applied Econometrics 7, S61–82.

Kalliovirta, L., Meitz, M., Saikkonen, P., 2015. A gaussian mixture autoregressive model. Journal of Time Series Analysis 36, 247–266.

Kang, K. H., 2014. State-space models with endogenous Markov regime switching parameters. Econometrics Journal 17, 56–82.

Kim, C.-J., 1994. Dynamic linear models with Markov-switching. Journal of Econometrics 60, 1–22.

Kim, C.-J., 2004. Markov-switching models with endogenous explanatory variables. Journal of Econometrics 122, 127–136.

Kim, C.-J., 2009. Markov-switching models with endogenous explanatory variables II: A two-step mle procedure. Journal of Econometrics 148, 46–55.

Kim, C.-J., Nelson, C., 1999. State-Space Models with Regime Switching. MIT Press, Cambridge, MA.

Kim, C.-J., Piger, J., Startz, R., 2008. Estimation of Markov regime-switching regression models with endogeneous switching. Journal of Econometrics 143, 263–273.

Timmermann, A., 2000. Moments of markov switching models. Journal of Econometrics 96, 75–111.

Yu, J., 2005. On leverage in a stochastic volatility model. Journal of Econometrics 127, 165–178.

# Appendix

## Appendix A: Mathematical Proofs

### Proof of Lemma 2.1

From (9), we may deduce that

$$\mathbb{P}\left\{s_t = 0 \,\middle|\, w_{t-1}\sqrt{1-\alpha^2} = x\right\} = \Phi\left(\tau - \frac{\alpha x}{\sqrt{1-\alpha^2}}\right),$$

from which it follows that

$$\begin{aligned}
\mathbb{P}\big\{s_t = 0 \,\big|\, w_{t-1} < \tau\big\} &= \mathbb{P}\left\{s_t = 0 \,\middle|\, w_{t-1}\sqrt{1-\alpha^2} < \tau\sqrt{1-\alpha^2}\right\} \\
&= \frac{\displaystyle\int_{-\infty}^{\tau\sqrt{1-\alpha^2}} \mathbb{P}\left\{s_t = 0 \,\middle|\, w_{t-1}\sqrt{1-\alpha^2} = x\right\} \varphi(x)dx}{\mathbb{P}\left\{w_{t-1}\sqrt{1-\alpha^2} < \tau\sqrt{1-\alpha^2}\right\}} \\
&= \frac{\displaystyle\int_{-\infty}^{\tau\sqrt{1-\alpha^2}} \Phi\left(\tau - \frac{\alpha x}{\sqrt{1-\alpha^2}}\right) \varphi(x)dx}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)},
\end{aligned}$$

upon noticing that $w_{t-1}\sqrt{1-\alpha^2} =_d \mathbb{N}(0,1)$. The stated result for $a(\alpha,\tau)$ can therefore be easily deduced from (11). Similarly, we have

$$\begin{aligned}\mathbb{P}\{s_t = 1 \,|w_{t-1} \geq \tau\} &= \mathbb{P}\left\{s_t = 1 \,\middle|\, w_{t-1}\sqrt{1-\alpha^2} \geq \tau\sqrt{1-\alpha^2}\right\} \\ &= \frac{\displaystyle\int_{\tau\sqrt{1-\alpha^2}}^{\infty} \mathbb{P}\left\{s_t = 1 \,\middle|\, w_{t-1}\sqrt{1-\alpha^2} = x\right\} \varphi(x)dx}{\mathbb{P}\left\{w_{t-1}\sqrt{1-\alpha^2} \geq \tau\sqrt{1-\alpha^2}\right\}} \\ &= \frac{\displaystyle\int_{\tau\sqrt{1-\alpha^2}}^{\infty} \left[1 - \Phi\left(\tau - \frac{\alpha x}{\sqrt{1-\alpha^2}}\right)\right] \varphi(x)dx}{1 - \Phi\left(\tau\sqrt{1-\alpha^2}\right)},\end{aligned}$$

since

$$\mathbb{P}\left\{s_t = 1 \,\middle|\, w_{t-1}\sqrt{1-\alpha^2} = x\right\} = 1 - \Phi\left(\tau - \frac{\alpha x}{\sqrt{1-\alpha^2}}\right),$$

due to (10), from which and (11) the stated result for $b(\alpha,\tau)$ follows readily as above. □

**Proof of Corollary 2.2**

The stated result for $t = 1$ is obvious, since $\mathbb{P}\{s_0 = 0\} = 1$ and $\mathbb{P}\{s_0 = 1\} = 1$ depending upon $\tau > 0$ and $\tau \leq 0$. Note that we set $w_0 = 0$ for identification when $\alpha = 1$. For $t \geq 2$, upon noticing that $w_{t-1}/\sqrt{t-1} =_d \mathbb{N}(0,1)$, the proof is entirely analogous to that of Lemma 2.1, and the details are omitted. □

**Proof of Theorem 3.1**

We only provide the proof for the case of $|\alpha| < 1$. The proof for the case of $\alpha = 1$ is virtually identical, except that we have $w_{t-1}/\sqrt{t-1} =_d \mathbb{N}(0,1)$ for $t \geq 2$ in this case, in place of $w_{t-1}\sqrt{1-\alpha^2} =_d \mathbb{N}(0,1)$ for the case of $|\alpha| < 1$. If we let

$$z_t = \frac{w_t - \alpha w_{t-1}}{\sqrt{1-\rho^2}} - \rho\frac{y_{t-1} - m_{t-1}}{\sigma_{t-1}\sqrt{1-\rho^2}},$$

we may easily deduce that

$$p(z_t|w_{t-1}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) = \mathbb{N}(0,1).$$

It follows that

$$\begin{aligned}
&\mathbb{P}\left\{w_t < \tau \middle| w_{t-1}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}\right\} \\
&= \mathbb{P}\left\{ z_t < \frac{\tau - \alpha w_{t-1}}{\sqrt{1-\rho^2}} - \rho \frac{y_{t-1} - m_{t-1}}{\sigma_{t-1}\sqrt{1-\rho^2}} \middle| w_{t-1}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1} \right\} \\
&= \Phi_\rho \left( \tau - \rho \frac{y_{t-1} - m_{t-1}}{\sigma_{t-1}} - \alpha w_{t-1} \right).
\end{aligned}$$

Note that

$$p(w_t | w_{t-1}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) = p(w_t | w_{t-1}, u_{t-1})$$

with $u_{t-1} = (y_{t-1} - m_{t-1})/\sigma_{t-1}$, and that $w_{t-1}$ is independent of $u_{t-1}$. Consequently, we have

$$\begin{aligned}
&\mathbb{P}\left\{w_t < \tau \middle| w_{t-1} < \tau, w_{t-2}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}\right\} \\
&= \mathbb{P}\left\{ w_t < \tau \middle| w_{t-1}\sqrt{1-\alpha^2} < \tau\sqrt{1-\alpha^2}, w_{t-2}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1} \right\} \\
&= \mathbb{P}\left\{ s_t = 0 \middle| s_{t-1} = 0, s_{t-2}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1} \right\} \\
&= \frac{\displaystyle\int_{-\infty}^{\tau\sqrt{1-\alpha^2}} \Phi_\rho \left( \tau - \rho \frac{y_{t-1} - m_{t-1}}{\sigma_{t-1}} - \frac{\alpha x}{\sqrt{1-\alpha^2}} \right) \varphi(x) dx}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left\{w_t < \tau \middle| w_{t-1} \geq \tau, w_{t-2}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}\right\} \\
&= \mathbb{P}\left\{ w_t < \tau \middle| w_{t-1}\sqrt{1-\alpha^2} \geq \tau\sqrt{1-\alpha^2}, w_{t-2}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1} \right\} \\
&= \mathbb{P}\left\{ s_t = 0 \middle| s_{t-1} = 1, s_{t-2}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1} \right\} \\
&= \frac{\displaystyle\int_{\tau\sqrt{1-\alpha^2}}^{\infty} \Phi_\rho \left( \tau - \rho \frac{y_{t-1} - m_{t-1}}{\sigma_{t-1}} - \frac{\alpha x}{\sqrt{1-\alpha^2}} \right) \varphi(x) dx}{1 - \Phi\left(\tau\sqrt{1-\alpha^2}\right)},
\end{aligned}$$

since in particular $w_{t-1}\sqrt{1-\alpha^2} =_d \mathbb{N}(0,1)$, from which the stated result for the transition density for $(s_t, y_t)$ may be readily obtained.

Now we write

$$\begin{aligned}
&p(s_t, y_t | s_{t-1}, \ldots, s_1, y_{t-1}, \ldots, y_1) \\
&= p(y_t | s_t, s_{t-1}, \ldots, s_1, y_{t-1}, \ldots, y_1) p(s_t | s_{t-1}, \ldots, s_1, y_{t-1}, \ldots, y_1).
\end{aligned}$$

It follows from (17) that

$$p(y_t|s_t, s_{t-1}, \ldots, s_1, y_{t-1}, \ldots, y_1) = p(y_t|s_t, \ldots s_{t-k}, y_{t-1}, \ldots, y_{t-k}).$$

Moreover, we have

$$p(s_t|s_{t-1}, \ldots, s_1, y_{t-1}, \ldots, y_1) = p(s_t|s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}),$$

as we have shown above. Therefore, it follows that

$$\begin{aligned}
&p(s_t, y_t|s_{t-1}, \ldots, s_1, y_{t-1}, \ldots, y_1) \\
&= p(y_t|s_t, \ldots s_{t-k}, y_{t-1}, \ldots, y_{t-k}) p(s_t|s_{t-1}, \ldots, s_{t-(k+1)}, y_{t-1}, \ldots, y_{t-k-1}) \\
&= p(s_t, y_t|s_{t-1}, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1})
\end{aligned}$$

and $(s_t, y_t)$ is a $(k+1)$-st order markov process. □

**Proof of Corollary 3.2**

We only provide the proof for the case of $0 < \alpha < 1$. The proof for the case of $\alpha = 0$ is trivial and the proof for the case of $-1 < \alpha < 0$ can be easily done with a simple modification of the case of $0 < \alpha < 1$. The proof for the case of $\alpha = 1$ is virtually identical, except that we have $w_{t-1}/\sqrt{t-1} =_d \mathbb{N}(0,1)$ for $t \geq 2$ in this case, in place of $w_{t-1}\sqrt{1-\alpha^2} =_d \mathbb{N}(0,1)$ for the case of $|\alpha| < 1$.

It follows that

$$\begin{aligned}
&\mathbb{P}\left\{w_t < \tau \middle| w_{t-1}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}\right\} \\
&= \mathbb{P}\left\{\alpha w_{t-1} + v_t < \tau \middle| w_{t-1}, \ldots, w_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}\right\} \\
&= \mathbb{P}\left\{\alpha w_{t-1} + \rho u_{t-1} < \tau \middle| w_{t-1}, u_{t-1}\right\} \\
&= \begin{cases} 1, & \text{if } \alpha w_{t-1} + \rho u_{t-1} < \tau. \\ 0, & \text{if otherwise.} \end{cases}
\end{aligned}$$

We note that when $0 < \alpha < 1$,

$$\omega_\rho(s_{t-1} = 0, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) = \mathbb{P}\left\{\alpha w_{t-1} + \rho u_{t-1} < \tau \middle| w_{t-1} < \tau, u_{t-1}\right\}$$
$$= \mathbb{P}\left\{ \sqrt{1-\alpha^2} w_{t-1} < \frac{\sqrt{1-\alpha^2}}{\alpha}(\tau - \rho u_{t-1}) \middle| \sqrt{1-\alpha^2} w_{t-1} < \tau\sqrt{1-\alpha^2}, u_{t-1} \right\}$$
$$= \begin{cases} 1, & \text{if } \frac{1}{\alpha}\left(\tau - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right) < \tau, \\ \dfrac{\Phi\left(\left(\tau - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)\frac{\sqrt{1-\alpha^2}}{\alpha}\right)}{\Phi(\tau\sqrt{1-\alpha^2})}, & \text{otherwise,} \end{cases}$$

where $u_{t-1} = (y_{t-1} - m_{t-1})/\sigma_{t-1}$. Similarly, we have

$$\omega_\rho(s_{t-1} = 1, \ldots, s_{t-k-1}, y_{t-1}, \ldots, y_{t-k-1}) = \mathbb{P}\left\{\alpha w_{t-1} + \rho u_{t-1} < \tau \middle| w_{t-1} > \tau, u_{t-1}\right\}$$
$$= \mathbb{P}\left\{ \sqrt{1-\alpha^2} w_{t-1} < \frac{\sqrt{1-\alpha^2}}{\alpha}(\tau - \rho u_{t-1}) \middle| \sqrt{1-\alpha^2} w_{t-1} > \tau\sqrt{1-\alpha^2}, u_{t-1} \right\}$$
$$= \begin{cases} 0, & \text{if } \frac{1}{\alpha}\left(\tau - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right) < \tau. \\ \dfrac{\Phi\left(\left(\tau - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)\frac{\sqrt{1-\alpha^2}}{\alpha}\right) - \Phi(\tau\sqrt{1-\alpha^2})}{1 - \Phi(\tau\sqrt{1-\alpha^2})}, & \text{otherwise.} \end{cases}$$

**Proof of Corollary 3.3**

We note that

$$\begin{aligned} & p\left(w_t | s_{t-1} = 1, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) \\ &= p\left(w_t | s_{t-1} = 1, s_{t-2}, ..., s_{t-k-1}, y_{t-1}, ..., y_{t-k-1}\right) \\ &= p\left(w_t | w_{t-1} > \tau, u_{t-1}\right) \\ &= \frac{\int_\tau^\infty p\left(w_t, w_{t-1}, u_{t-1}\right) dw_{t-1}}{\int_\tau^\infty p\left(w_{t-1}, u_{t-1}\right) dw_{t-1}} \\ &= \frac{\int_\tau^\infty p\left(w_t | w_{t-1}, u_{t-1}\right) p\left(w_{t-1}\right) dw_{t-1}}{\int_\tau^\infty p\left(w_{t-1}\right) dw_{t-1}}. \end{aligned} \tag{27}$$

The last equality comes from the independence between $(w_t)$ and $(u_t)$. Likewise, we have

$$p\left(w_t | s_{t-1} = 0, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) = \frac{\int_{-\infty}^\tau p\left(w_t | w_{t-1}, u_{t-1}\right) p\left(w_{t-1}\right) dw_{t-1}}{\int_{-\infty}^\tau p\left(w_{t-1}\right) dw_{t-1}}.$$

When $|\alpha| < 1$ and $|\rho| < 1$, we have $w_t|w_{t-1}, u_{t-1} =_d \mathbb{N}\left(\alpha w_{t-1} + \rho u_{t-1}, 1-\rho^2\right)$. Since $w_t =_d w_{t-1} =_d \mathbb{N}\left(0, \frac{1}{1-\alpha^2}\right)$, we can easily deduce

$$
\begin{aligned}
&p\left(w_t|s_{t-1}=1, s_{t-2}, ..., s_{t-k-1}\mathcal{F}_{t-1}\right) \\
&= \frac{\left(1-\Phi\left(\sqrt{\frac{1-\rho^2+\alpha^2\rho^2}{1-\rho^2}}\left(\tau - \frac{\alpha\left(w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)}{1-\rho^2+\alpha^2\rho^2}\right)\right)\right)}{1-\Phi\left(\tau\sqrt{1-\alpha^2}\right)}\mathbb{N}\left(\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}, \frac{1-\rho^2+\alpha^2\rho^2}{1-\alpha^2}\right) \\
&p\left(w_t|s_{t-1}=0, s_{t-2}, ..., s_{t-k-1}\mathcal{F}_{t-1}\right) \\
&= \frac{\Phi\left(\sqrt{\frac{1-\rho^2+\alpha^2\rho^2}{1-\rho^2}}\left(\tau - \frac{\alpha\left(w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}\right)}{1-\rho^2+\alpha^2\rho^2}\right)\right)}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)}\mathbb{N}\left(\rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}, \frac{1-\rho^2+\alpha^2\rho^2}{1-\alpha^2}\right).
\end{aligned}
$$

When $|\alpha| < 1$ and $|\rho| = 1$, we note that

$$
\begin{aligned}
&p\left(w_t|s_{t-1}=1, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) \\
&= p\left(w_t|s_{t-1}=1, s_{t-2}, ..., s_{t-k-1}, y_{t-1}, ..., y_{t-k-1}\right) \\
&= p\left(w_t|w_{t-1} > \tau, u_{t-1}\right) \\
&= p\left(\alpha w_{t-1} + v_t|w_{t-1} > \tau, v_t\right). \qquad (28)
\end{aligned}
$$

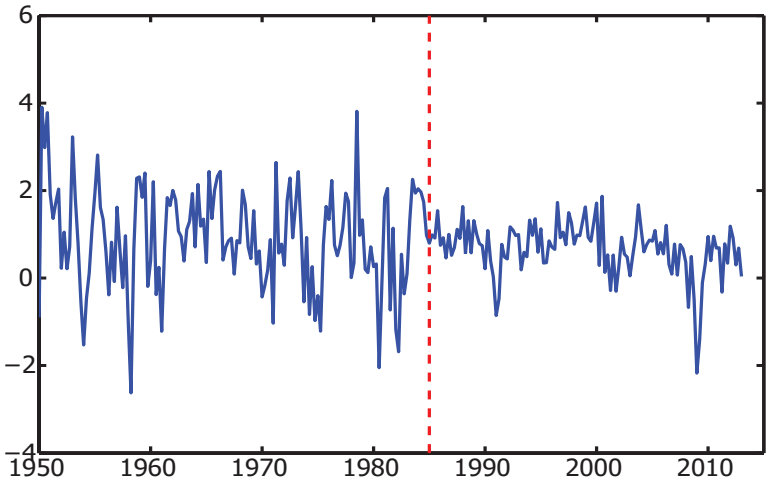
Since $p\left(w_{t-1}|w_{t-1} > \tau\right) = \frac{\sqrt{1-\alpha^2}\phi\left(w_{t-1}\sqrt{1-\alpha^2}\right)}{1-\Phi\left(\tau\sqrt{1-\alpha^2}\right)}$, due to (28), we can easily deduce

$$
\begin{aligned}
p\left(w_t|s_{t-1}=1, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) &= \frac{\frac{\sqrt{1-\alpha^2}}{\alpha}\phi\left(\frac{w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}}{\alpha}\sqrt{1-\alpha^2}\right)}{1-\Phi\left(\tau\sqrt{1-\alpha^2}\right)}, \\
p\left(w_t|s_{t-1}=0, s_{t-2}, ..., s_{t-k-1}, \mathcal{F}_{t-1}\right) &= \frac{\frac{\sqrt{1-\alpha^2}}{\alpha}\phi\left(\frac{w_t - \rho\frac{y_{t-1}-m_{t-1}}{\sigma_{t-1}}}{\alpha}\sqrt{1-\alpha^2}\right)}{\Phi\left(\tau\sqrt{1-\alpha^2}\right)}.
\end{aligned}
$$

When $\alpha = 1$, the proofs are analogues to those of above cases except the fact that $w_t =_d \mathbb{N}(0, t)$ and therefore omitted.

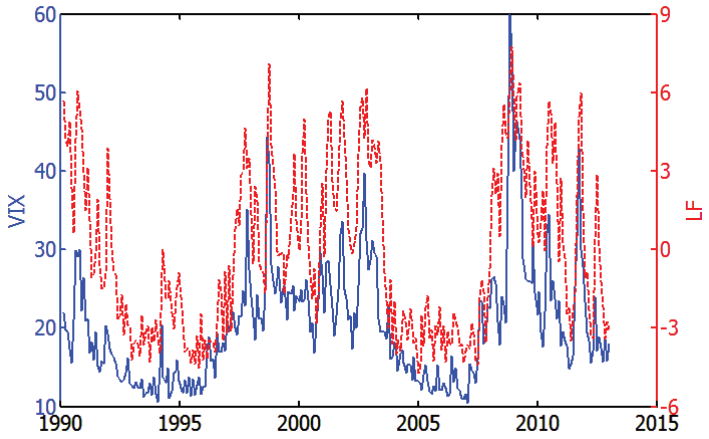
## Appendix B: Additional Figures

Figure 12: US Real GDP Growth Rates



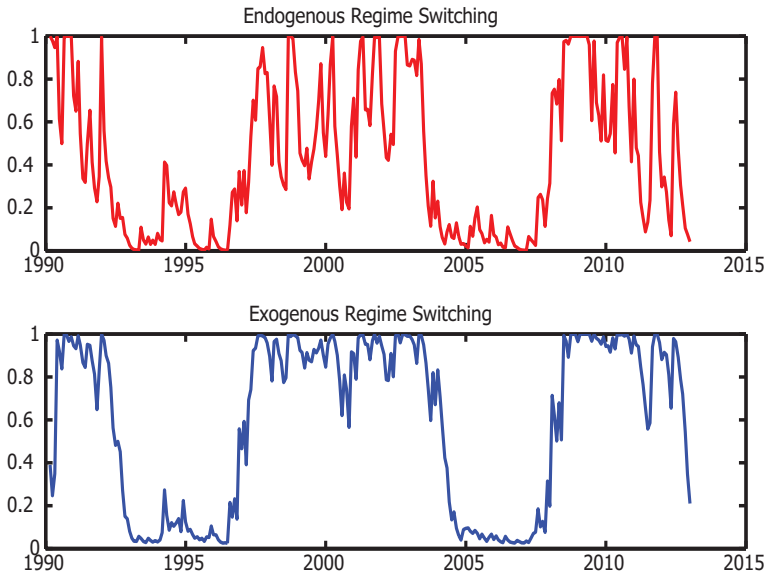

Notes: Figure 12 presents the US real GDP growth rates which is calculated as 100 times the change in the log of real GDP. It is seasonally adjusted, annualized, and collected at the quarterly frequency from 1952 to 2012. The vertical dashed red line indicates 1983:Q4.

Figure 13: Extracted Latent Factor and VIX



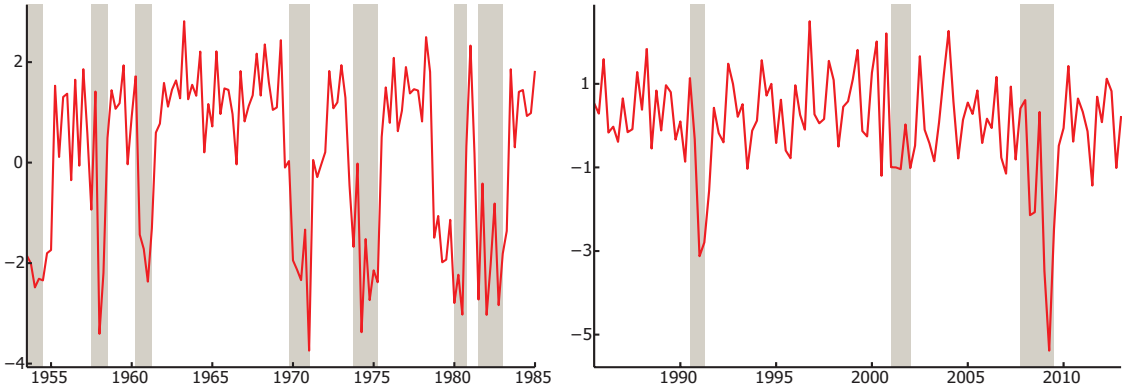

Notes: Figure 13 presents the sample path of the latent factor extracted from the endogenous volatility switching model (dashed red line) along with that of the CBOE (The Chicago Board Options Exchange) volatility index VIX (solid blue line) for the period 1990-2012, respectively, on the left and right vertical axis.

Figure 14: High Volatility Probabilities



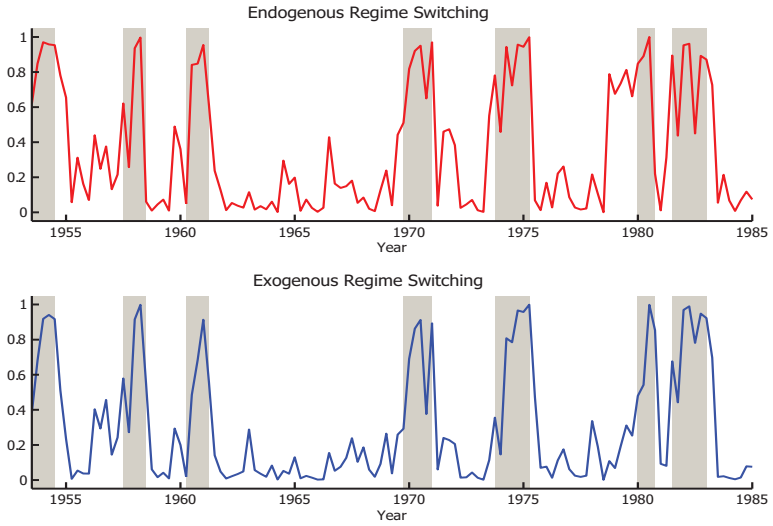

Notes: Figure 14 presents the time series of the probabilities of being in the high volatility regime. Top panel plots the high volatility probability series obtained from the endogenous volatility switching model with red line, while the bottom panel plots those from its conventional exogenous counterpart with blue line.

Figure 15: NBER Recession Periods and Extracted Latent Factor



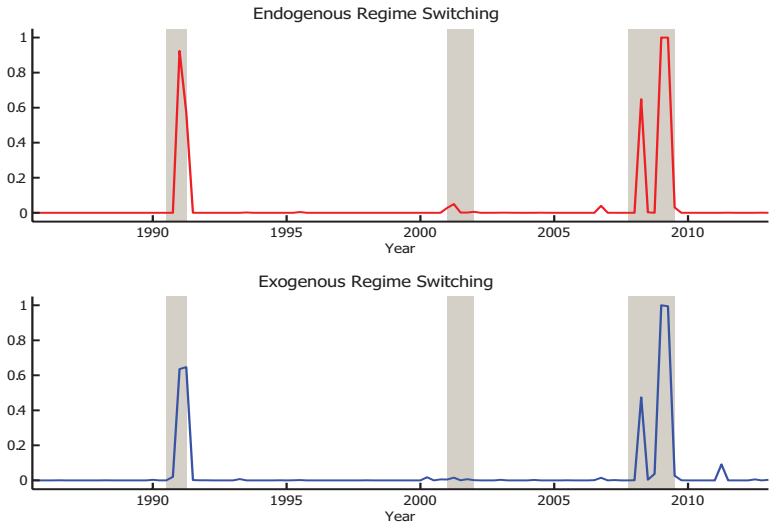

Notes: Figure 15 presents the latent factor determining the states extracted from the endogenous mean switching model, which is compared with the recession periods identified by NBER. The left hand side graph presents extracted latent factor plotted with solid red line and NBER recession periods displayed as shaded gray areas for the sample period 1952-1984, while the graph on the right presents those for the more recent sample period 1984-2012.

Figure 16: Recession Probabilities During 1952-1984




Notes: Figure 16 presents the recession probabilities for the earlier sample period, 1952-1984. Top panel plots the recession probabilities from the endogenous mean switching model with red line, while the bottom panel plots those from its conventional exogenous counterpart with blue line. Both panels also show the recession periods identified by NBER.

Figure 17: Recession Probabilities During 1984-2012




Notes: Figure 17 presents the recession probabilities for the recent sample period, 1984-2012. Top panel plots the recession probabilities from the endogenous mean switching model with red line, while the bottom panel plots those from its conventional exogenous counterpart with blue line. Both panels also show the recession periods identified by NBER.