Specification Testing and Semiparametric Estimation of Regime Switching Models: An Examination of the US Short Term Interest Rate

Sean D. Campbell

Department of Economics

Brown University

Working Paper 2002-26

October 22, 2002

Abstract

This paper develops a method for quantitatively and qualitatively assessing the adequacy of the normality assumption in regime switching models. A formal test that extends Jarque and Bera's (1982) normality test to regime switching settings is proposed. Quasi maximum likelihood estimation of regime switching models is shown to be inconsistent. The feasibility of semiparametric identification of regime switching models is shown and a semiparametric estimator is proposed. Empirically, a two regime Gaussian model of the U.S. short term interest rate is shown to be misspecified. The semiparametric estimator reveals one low volatility regime that is well approximated by normality and one high volatility regime that is negatively skewed and leptokurtic relative to the normal distribution.

*Sean D. Campbell, 64 Waterman Street, Box B, Providence, RI 02912. Phone: 401.863.2087, Fax: 401.863.1970

I would like to thank Frank Diebold, Peter Reinhard Hansen, Yuichi Kitamura, Bobby Mariano and Frank Schorfheide for a series of useful discussions and comments. The usual disclaimer applies.

JEL Classification: C32, C14
Keywords: Regime Switching, Semiparametric Econometrics, Short Term Interest Rate

1 Introduction

Since Hamilton (1989) introduced regime switching models and demonstrated their ability to describe the salient business cycle features of aggregate output, their use has grown dramatically. While regime switching models are still used extensively to model aggregate macroeconomic time series such as aggregate output and industrial production that display radically different behavior during "boom" and "bust" periods, they have also been applied successfully in the modeling of many important financial time series ranging from stock market returns (Turner, Startz and Nelson (1989)), to foreign currencies (Engel and Hamilton (1990)) and interest rates (Gray (1996)).[1]

Aside from their widespread use in empirical work, models that incorporate regime switching in economic fundamentals (e.g., dividends, consumption or GDP) have recently shed light on some puzzling aspects of financial markets that are difficult to reconcile in single regime models. Cechetti, Lam and Mark (1990,2000) have shown that when economic fundamentals switch between persistent high and low growth regimes, long horizon stock returns exhibit the kind of negative serial autocorrelation that has been documented by Fama and French (1988) and others. Regime switching models have also been used to explain the weak correlation between changes in volatility and excess returns over time.[2] Whitelaw (2000) employs a model with high and low growth regimes in consumption to show that the simple monotonic relationship between expected returns and volatility in static models (CAPM) need not hold in dynamic, multi regime settings. In particular, Whiltelaw shows that accounting for regime switching in fundamentals results in excess returns that vary little with volatility, thus bridging the gap between theory and empirical examinations of the time series relationship between risk and return. In this way, regime switching models present an empirically relevant modeling framework that make a close connection with theoretical models.

In this paper we make several important contributions to the theoretical and empirical regime switching literature. First, building on the work of Hamilton (1996), we provide a set of diagnostics and a test that can be used as a means of qualitatively and quantitatively assessing the plausibility of the normality assumption in regime switching models. Second, we show that when normality is falsely imposed, the resulting quasi maximum likelihood (QMLE) estimator is inconsistent. Third, we propose a semiparametric extension of Hamilton's (1990) EM algorithm to estimate regime switching models in the presence of unspecified deviations from normality. In this sense our work is related to that of Engle and Gonzalez-Rivera (1991) who generalize traditional Gaussian ARCH models to more flexible semiparametric settings.

Empirically, we focus on a two regime model of the U.S. short term interest rate. The proposed test rejects the normality assumption at all conventional significance levels. In particualar, we find that interest rate changes are characterized by one low volatility regime that is well approximated

[1] A short list macroeconomic related references include: Diebold and Rudebusch (1996), Durland and McCurdy (1994), Filardo (1994).

[2] Glosten, Jagannathan and Runkle (1993) document the weak time series relationship between risk and return.

by normality and one high volatility regime that is negatively skewed and leptokurtic relative to a normal distribution.

Regime switching models are typically constructed using the following specification,[3]

$$y_t = x_t'\beta_r + \sigma(z_t, \gamma_r)\varepsilon_{r,t};\ r = \{0, 1\} \tag{1.1}$$

in which the parameters governing the conditional mean ($x_t'\beta_r$) and variance ($\sigma^2(z_t, \gamma_r)$) are allowed to differ across regimes. The regime is unobserved and is determined as the outcome of a markov chain process in which the current regime, r_t, only depends on the lagged value of the (unobserved) regime, r_{t-1}. Specifically, conditional on r_{t-1} the probability of transiting from one regime to another is specified as $\Pr(r_t|r_{t-1})$. These are often referred to as the transition probabilities of the process and in the two regime case they are denoted as P_{00} and P_{11}, respectively[4].

In nearly all empirical work, the distribution of $\varepsilon_{r,t}$ is assumed to be iid and Gaussian(0,1).[5] This assumption precludes the possibility of any excess skewness or kurtosis in the regime conditional distribution of y, denoted as $f_r(y_t|x_t, z_t)$. Consequently, any departure from normality in the unconditional distribution, $f(y_t|x_t, z_t)$, can solely be attributed to the effect of switching between different regimes. Allowing for the possibility of significant departures from normality in the regime conditional distribution is important for two reasons. First, from an econometric perspective, a more accurate description of the uncertainty within a regime will necessarily yield better informed and more accurate estimates of the unobserved regime of the economy. More accurate estimates of the latent regime imply superior parameter estimation as well as more forecasting power for future regimes. Second, and more importantly, departures from normality in the regime conditional distribution have important economic implications in their own right. Consider, for example, a two regime model for stock returns in which each distribution is specified as Gaussian with mean μ_r and regime specific variance σ_r^2. Suppose further that the more volatile regime is also characterized by significant negative skewness and excess kurtosis. This type of information would be crucial for assessing portfolio risk and would also be important from an asset pricing perspective. In particular, this pattern in the regime conditional distributions of stock returns would have interesting implications for option prices in the kinds of regime switching option pricing models that have been recently explored by Bollen, Gray and Whaley (2000) and Campbell and Li (2001).

Exploring the regime conditional distribution is related to the question of choosing the number of regimes. A number of authors such as Hansen (1992), Garcia (1998) have taken up the question of how many regimes are required to adequately model the distribution of y. This line of research conditions on within regime normality and then tests the null of K regimes against the alternative

[3] In this paper we only consider two regime models. Two regime models dominate the regime switching literature and restricting attention to the two regime case simplifies exposition and notation considerably. Extensions to multi regime settings are straightforward.

[4] Also note that $Pr(r_t = 1|r_{t-1}) \equiv P_{01} = 1 - P_{00}$ and $Pr(r_t = 0|r_{t-1} = 1) \equiv P_{10} = 1 - P_{11}$

[5] Two important exceptions are Hamilton and Susmel (1994) and Perez-Quiros and Timmerman (2000).

of $K+1$ regimes. Our approach takes the opposite tack. We condition on the number of regimes and ask to what extent the Gaussian assumption is consistent with the observed data. In each case, the goal is to provide a means for assessing how to improve the regime switching model's characterization of the data. In this sense, both lines of research are complementary.

The remainder of the paper is organized as follows. Section 2 discusses Gaussian regime switching models, their applications to interest rate processes and why it is necessary to consider more general alternatives. Section 3 develops the test of normality and applies it to the U.S. short term interest rate. Section 4 discusses quasi maximum likelihood estimation (QMLE) of regime switching models and shows that the resulting estimator is inconsistent. Section 5 discusses model identification and outlines a framework for estimating the regime conditional distribution in a distribution free context and applies it to the U.S. short term interest rate. Section 6 concludes. All proofs may be found in the appendix.

2 The Structure of Regime Switching Models

2.1 Gaussian Regime Switching Models

In this section we summarize the key points regarding the construction and estimation of Gaussian regime switching models.

Regime switching models are often specified as:

$$\begin{aligned} \Pr(r_t &= 0|r_{t-1}=0) = P_{00} \\ \Pr(r_t &= 1|r_{t-1}=1) = P_{11} \\ y_t &= x_t'\beta_r + \sigma_r\varepsilon_{r,t} \\ \varepsilon_{r,t} &\sim f_r(\cdot)\ (iid). \end{aligned} \tag{2.1}$$

In the case of Gaussian models, $f_0(\cdot) = f_1(\cdot) = N(0,1)$. Interesting and important variations on the standard model above are the inclusion of time varying transition probabilities, $P_{rr,t} = \Phi(z_t'\gamma_r)$, as in Diebold, Weinbach, and Lee (1994) and more complex volatility specifications that allow for volatility dynamics within each regime as in Gray (1996). While these are important extensions, we focus on the simple structure above to facilitate an intuitive exposition of the model while avoiding unnecessary complexity and notation.

Recalling that the regime is unobserved, the distribution of y_t conditional on the past information set, Ω_{t-1}, takes the form:

$$f(y_t|\Omega_{t-1};\theta) = \pi^1_{t|t-1}\phi\left(\frac{y_t - x_t'\beta_1}{\sigma_1}\right) + \pi^0_{t|t-1}\phi\left(\frac{y_t - x_t'\beta_0}{\sigma_0}\right) \tag{2.2}$$

where $\pi^1_{t|t-1} = 1 - \pi^0_{t|t-1} = \Pr(r_{t-1} = 1|\Omega_{t-1}; \theta)$ and $\theta = (\beta_0, \beta_1, \sigma_0, \sigma_1, P_{00}, P_{11})$ and ϕ represents the standard normal pdf. Expressions for $\pi^r_{t|t-1}$ are obtained by Bayes' Rule

$$\begin{aligned} \pi^1_{t|t-1} &= \pi^1_{t-1|t-1} P_{11} + \pi^0_{t-1|t-1}(1 - P_{00}), \\ \pi^1_{t|t} &= \frac{\pi^1_{t|t-1} \phi\left(\frac{y_t - x_t'\beta_1}{\sigma_1}\right)}{f(y_t|\Omega_{t-1})}, \end{aligned} \tag{2.3}$$

and, $\pi^r_{t|t+\tau}$ is similarly defined. Specifically, when τ is negative, $\pi^r_{t|t+\tau}$ represents a forecast of the probability that regime r will be realized in τ periods. When τ is positive, $\pi^r_{t|t+\tau}$ represents a smoothed or updated inference of the probability that regime r was in fact realized τ periods ago. Much of the analysis that follows will make use of these smoothed probabilities. At this point we stress that the formulae for constructing $\pi^r_{t|t+\tau}$ are sensitive to the choice of regime conditional distribution. A different choice of $f_r(\cdot)$ will lead to a different sequence of $\pi^r_{t|t+\tau}$. This simple observation will be a key ingredient in much of what follows in this paper.

The construction of the likelihood facilitates parameter estimation by MLE, using the log likelihood $\sum \ln(f(y_t|\Omega_{t-1}; \theta))$. The model's parameters may then be estimated by setting the score of the likelihood to zero using traditional optimization routines. Although setting the score to zero delivers parameter estimates, Hamilton (1990) shows how the MLE can be constructed using an application of the EM algorithm of Dempster, Laird and Rubin (1977). In the case of the model specified above, it can be shown that application of the EM algorithm results in the following set of recursive equations for $l = 1, 2...$,

Algorithm 2.1

$$\begin{aligned} b^l_r &= \left(X_r' X_r\right)^{-1} X_r' Y_r; \ r = \{0, 1\} \\ s^{2,l}_r &= \frac{1}{\sum \widehat{\pi}^r_{t|T}} \left(Y_r - X_r' b_r\right)' \left(Y_r - X_r' b_r\right); \ r = \{0, 1\} \\ P^l_{rr} &= \frac{\sum \Pr(r_t = r, r_{t-1} = r; \Omega_T, \theta^{l-1})}{\sum \Pr(r_t = r; \Omega_T, \theta^{l-1})}; \ r = \{0, 1\} \end{aligned} \tag{2.4}$$

where $X_r = \sqrt{\widehat{\pi}^r_{t|T}} \cdot X$, $Y_r = \sqrt{\widehat{\pi}^r_{t|T}} \cdot Y$, Y and X are defined in the usual way and $A \cdot B$ denotes element by element multiplication. The equations are iterated upon until the difference between successive values of θ^l is small. Lastly, note that the recursive nature of the estimation (EM) algorithm arises because of the need to construct $\widehat{\pi}^r_{t|T}$ (which depends on θ^{l-1}) before computing $\theta^l = \left(\beta^l_0, \beta^l_1, \sigma^l_0, \sigma^l_1, P^l_{00}, P^l_{11}\right)$.[6] Accordingly, the recursion begins by choosing an initial parameter vector θ^0 computing $\widehat{\pi}^r_{t|T}$, then computing θ^1 and so on until convergence is achieved.

[6] The formulae for $\pi^r_{t|T}$ and $\Pr(r_t = r, r_{t-1} = r; \Omega_T, \theta)$ are a set of highly nonlinear recursions. The exact form of these recursions may be found in Hamilton (1990).

2.2 Estimating Conditional Moments in a Regime Switching Setting

As summarized above, Hamilton (1990) shows that the MLE of the Gaussian regime switching model (2.4) is equivalent to a weighted regression where the weights are the smoothed state probabilities. Accordingly, given the smoothed probabilities, the MLE problem can be reduced to a GLS regression. The main difference between the weights used in this context and more traditional GLS weights, is that in a regime switching setting the weights, $\sqrt{\pi^r_{t|T}}$, do not only play an efficiency enhancing role but rather they guarantee consistency. Below we state a related result that will serve as a key insight underlying the remainder of the paper.

Lemma 2.2 *Let $z_t = [y_t, x_t]'$. Given a regime switching model of the form in (2.1) $E\left(\pi^r_{t|t+\tau} m(z_t)\right) = \pi^r E\left(m(z_t)|r\right)$ for any $\tau \geq 0$ and $r = \{0, 1\}$.*

Proof. See Appendix. ■

The lemma is a population statement. In population, weighting the data with the appropriate regime probabilities enables one to uncover the moments of the regime dependent distributions even though the regimes are unobserved regardless of the form of $f_r(\cdot)$. For example, setting $m(z_t) = x_t' x_t$ yields $E\left(\pi^r_{t|t+\tau} x_t' x_t\right) = \pi^r E\left(x_t' x_t | r\right)$ and setting $m(z_t) = x_t' y_t$ yields $E\left(\pi^r_{t|t+\tau} x_t' y_t\right) = \pi^r E\left(x_t' y_t | r\right)$ and their ratio reveals β_r. Also setting $m(z_t) = \left(y_t - x_t'\beta_r\right)^2$ yields $E\left(\pi^r_{t|t+\tau}\left(y_t - x_t'\beta_r\right)^2\right) = \pi^r E\left(\left(y_t - x_t'\beta_r\right)^2 | r\right) = \pi_r \sigma_r^2$. The lemma provides an intuitive basis for the EM algorithm when $f_0(\cdot) = f_1(\cdot) = N(0, 1)$. In a Gaussian, single regime, framework maximum likelihood estimation simply sets population moments (β, σ) to their sample counterparts. In the regime switching setting, MLE sets regime specific population moments (β_r, σ_r) to their regime weighted sample counterparts. Moreover, the only reason to iterate more than once in the EM algorithm is due to uncertainty over $\pi^r_{t|T}$. If $\pi^r_{t|T}$ were observed alongside the data, MLE could be carried out in a single iteration of the EM algorithm.

The intuition behind the lemma can be understood within the context of importance sampling. Consider the problem of computing $E_g[z] = \int z g(z) dz$ vis-a-vis Monte Carlo integration. In many cases, generating random draws from $g(\cdot)$ is too burdensome so we draw K random variates from a well chosen "importance sampler" $h(\cdot)$ and approximate $E_g[z]$ by $\frac{1}{K}\sum z \frac{g}{h}(z)$. Ultimately, as the law of large numbers takes hold, the finite summation converges to integration and we obtain: $\frac{1}{K}\sum z \frac{g}{h}(z) \rightarrow \int z \frac{g(z)}{h(z)} h(z) dz = E_g(z)$. In the current context, we would like to compute $\int m(z_t) f_r(z_t) dz_t$ but are precluded from doing so because we are unable to sample from $f_r(\cdot)$.[7] While we can't directly observe $f_r(\cdot)$, $f(z_t|\Omega_{t-1})$ is revealed to the econometrician through the sample $\{z_t\}_{t=1}^{t=T}$. Accordingly, $f(z_t|\Omega_{t-1})$ may be used as an importance sampler in conjunction with $\pi^r_{t|t+\tau}$ to recover $\int m(z_t) f_r(z_t) dz_t$. As $f(z_t|\Omega_{t-1})$ plays the role of the importance sampler, $\pi^r_{t|t+\tau}$ plays the role of the ratio between the density of interest, $f_r(z_t)$, and the importance sampler

[7] Note that since the distribution of x_t is independent of the regime in model (2.1), $f_r(z_t) = f_r(y_t|x_t) f(x_t)$

$f(z_t|\Omega_{t-1})$. This can be seen by using Bayes' Rule to re-write $\pi^r_{t|t+\tau}$ in a manner that is proportional to $\frac{f_r(z_t)}{f(z_t|\Omega_{t-1})}$. Heuristically, the lemma is related to importance sampling in the sense that we use $\pi^r_{t|t+\tau}$ to focus our attention on the observations that were most likely generated from the regime of interest.

The lemma provides a formal justification for using regime weighted averages to approximate moments of the regime dependent distributions. In particular, the within regime sample skewness ($\widehat{S}_r$) and kurtosis ($\widehat{K}_r$) are defined as follows:

$$\widehat{S}_r = \frac{1}{\widehat{\sigma}^3_r \sum \widehat{\pi}^r_{t|T}} \sum \widehat{\pi}^r_{t|T} (\widehat{\varepsilon}_{r,t})^3;\ r = \{0,1\} \tag{2.5}$$

$$\widehat{K}_r = \frac{1}{\widehat{\sigma}^4_r \sum \widehat{\pi}^r_{t|T}} \sum \widehat{\pi}^r_{t|T} (\widehat{\varepsilon}_{r,t})^4;\ r = \{0,1\}. \tag{2.6}$$

These expressions are the natural generalization of their single regime counterparts and can be easily computed once the model has been estimated assuming that $f_r(\cdot)$ is N(0,1). Namely, if $\pi^r_{t|T} = 1$ for all t then S_r and K_r are identical to the sample skewness and kurtosis. These sample statistics can be used in the same way that sample skewness and kurtosis are used in a single regime framework to gauge the size and nature of departures from normality. In particular, under the null hypothesis that the regime conditional distribution is Gaussian, we should expect to find values of S_r and $K_r - 3$ that are close to zero. Furthermore, the estimates of within regime skewness and kurtosis can be used to construct an informal Jarque-Bera test statistic

$$JB_r = \frac{\sum_{t=1}^T \widehat{\pi}^r_{t|T}}{6} \left(\widehat{S}^2_r + \frac{(\widehat{K}_r - 3)^2}{4} \right), \tag{2.7}$$

which may be used as a preliminary means of checking the adequacy of the within regime normality assumption before proceeding with the more formal testing procedure that we develop in the next section[8].

2.3 A Jarque-Bera Test for Regime Switching Models

Hamilton (1996) introduces a framework for hypothesis testing in regime switching models that relies on the Lagrange Multiplier (LM) principle. In the current context we are interested in testing a variant of model (2.1) in which $f_r(\cdot)$ is assumed to be a member of the Pearson class of distributions. Below we formally state the null and alternative hypotheses.

[8] We note that as long as a constant is included in x_{t-1} then $\sum_{t=1}^T \widehat{\pi}^r_{t|T} \widehat{\varepsilon}_{r,t} = 0$. Accordingly, there is no need to demean $\widehat{\varepsilon}_{r,t}$

- H_0 : y_t is generated by model (2.1) with

$$\begin{aligned} \varepsilon_{r,t} &\sim f_p(z;\mathbf{a}_r)\ (iid) \\ \frac{d\ln f_p(z;\mathbf{a}_r)}{dz} &= \frac{a_{r,1}+z}{a_{r,2}+a_{r,3}z+a_{r,4}z^2} \end{aligned}$$

  and $a_{r,3}=a_{r,4}=0$ for $r=0,1$.

- H_A : At least one term $(a_{0,3},a_{0,4},a_{1,3},a_{1,4})$ is non-zero.

Following Hamilton (1996), we evaluate the restricted score of the above model and then compute

$$LM \equiv JB_{RS} = T\left[\frac{1}{T}\sum_{t=1}^{T} h_t(\widehat{\theta})\right]' \left[\widehat{I}(\widehat{\theta})\right]^{-1} \left[\frac{1}{T}\sum_{t=1}^{T} h_t(\widehat{\theta})\right] \tag{2.8}$$

where $h_t(\widehat{\theta})$ represents the restricted score vector and $\widehat{I}(\widehat{\theta})$ represents a consistent estimator of the model's information matrix. Under the maintained assumptions in Hamilton (1996) the test statistic is asymptotically distributed $\chi^2(4)$. We also note that while the test is developed for the null hypothesis of within regime normality across both regimes, the null hypothesis can easily be modified to test for within regime normality in one of the two regimes. The resulting tests are then asymptotically distributed $\chi^2(2)$. Below we present expressions for the score and discuss estimation of the information matrix.

Hamilton (1996) shows that the score of the likelihood, with respect to all parameters except the transition probabilities, can be represented as,

$$h^i_{t,r}(\theta) \equiv \frac{\partial \log f(y_t|\Omega_{t-1};\theta)}{\partial \theta^i_r} = \psi^i_{t,r}\pi^r_{t|t} + \sum_{\tau=1}^{t-1}\psi^i_{\tau,r}\left(\pi^r_{\tau|t}-\pi^r_{\tau|t-1}\right)$$

where $\psi^i_{t,r} \equiv \frac{\partial \log f(y_t|\Omega_{t-1},r_t=r;\theta)}{\partial\theta^i_r}$ represents the score of the regime conditional likelihood with respect to one of the parameters. Below we provide the score of the regime conditional likelihood. The derivation of these results can be found in Jarque and Bera (1982).

$$\begin{aligned} \psi^1_{t,r} &\equiv \frac{\partial \log f(y_t|\Omega_{t-1},r_t=r;\theta)}{\partial\beta_r} = \frac{(y_t-x'_{t-1}\beta_r)x'_t}{a_{r,2}} \\ \psi^2_{t,r} &\equiv \frac{\partial \log f(y_t|\Omega_{t-1},r_t=r;\theta)}{\partial a_{r,2}} = \frac{-1}{2a_{r,2}} + \frac{(y_t-x'_t\beta_r)^2}{2a^2_{r,2}} \\ \psi^3_{t,r} &\equiv \frac{\partial \log f(y_t|\Omega_{t-1},r_t=r;\theta)}{\partial a_{r,3}} = \frac{-(y_t-x'_t\beta_r)^3}{3a^2_{r,2}} \\ \psi^4_{t,r} &\equiv \frac{\partial \log f(y_t|\Omega_{t-1},r_t=r;\theta)}{\partial a_{r,4}} = \frac{(y_t-x'_t\beta_r)^4}{4a^2_{r,2}} - \frac{3}{4} \end{aligned}$$

Lastly, the expression for the score of the likelihood with respect to P_{00} ($h^5_{t,0}$) is given by,

$$\begin{aligned}\frac{\partial \log f(y_t|\Omega_{t-1};\theta)}{\partial P_{00}} &= P_{00}^{-1}\Pr(r_t=0,r_{t-1}=0|\Omega_t;\theta)-(1-P_{00})^{-1}\Pr(r_t=1,r_{t-1}=0|\Omega_t;\theta)\\ &+P_{00}^{-1}\sum_{\tau=2}^{t-1}[\Pr(r_\tau=0,r_{\tau-1}=0|\Omega_t;\theta)-\Pr(r_\tau=0,r_{\tau-1}=0|\Omega_{t-1};\theta)]-\\ &(1-P_{00})^{-1}\sum_{\tau=2}^{t-1}[\Pr(r_\tau=1,r_{\tau-1}=0|\Omega_t;\theta)-\Pr(r_\tau=1,r_{\tau-1}=0|\Omega_{t-1};\theta)]+\\ &\frac{\Pr(r_1=0|\Omega_t;\theta)-\Pr(r_1=0|\Omega_{t-1};\theta)}{(1-P_{00})}\end{aligned}$$

for $t=2,...,T$ and

$$\frac{\Pr(r_1=0|\Omega_1)-\left[\frac{(1-P_{11})}{(2-P_{11}-P_{00})}\right]}{(1-P_{00})}$$

for $t=1$. The expression for the score with respect to P_{11} ($h^5_{t,1}$) can be obtained analogously. Collecting these expressions together yields the score of the likelihood, $h_t(\theta)\equiv\left[h^1_{t,0},...,h^5_{t,0},h^1_{t,1},...,h^5_{t,1}\right]'$ from which the LM test can be constructed.

In nonlinear models LM tests are often difficult to interpret. Exactly, what is being tested by the LM test? In light of the previous lemma, the LM test can be easily interpreted. The LM test simply checks whether $E\left(h^i_{r,t}\right)=0$ by checking a standardized sample average of $h^i_{r,t}$. In the case of h^3_r, this amounts to checking whether

$$\underbrace{E\left(\pi^r_{t|t}\frac{(\varepsilon_{r,t})^3}{3\sigma_r^2}\right)}_{a}+\underbrace{\sum_{\tau=1}^{t-1}E\left(\frac{(\varepsilon_{r,\tau})^3}{3\sigma_r^2}\left(\pi^r_{\tau|t}-\pi_{\tau|t-1}\right)\right)}_{b}=0$$

or not. Under the null hypothesis, $a\propto\pi^r E\left(\varepsilon^3_{r,t}|r\right)$ which is zero. Secondly, $b\propto E\left(\varepsilon^3_{r,\tau}E\left(\pi^r_{\tau|t}-\pi^r_{\tau|t-1}|\Omega_{t-1}\right)\right)$ which is also zero under the null hypothesis since $E\left(\pi^r_{\tau|t}|\Omega_{t-1}\right)=\pi^r_{\tau|t-1}$ as long as $\tau\leq t-1$ by the law of iterated expectations. The first part of the score measures the degree of skewness within regime r while the second part of the score measures whether the dynamics of the regime switching model are misspecified. The intuition for $h^4_{t,r}$ is similar. Viewed in this light, the LM test emerges as a means of simultaneously measuring deviations from within regime normality (a) as well as departures in the dynamic structure of $\pi_{\tau|t}$ implied by the model (b).

3 Quasi Maximum Likelihood Estimation of Regime Switching Models

As noted earlier, nearly all empirical regime switching models assume Gaussianity of the regime conditional distribution. In the presence of misspecification, the validity of this procedure rests on the consistency properties of the QML estimator. Even if misspecification yields inappropriate interval and density forecasts, the QMLE procedure could still be useful if it is consistent for the mean and variance parameters (β_r, σ_r) of a regime switching model. Unfortunately, the QMLE procedure is not consistent. The intuition behind this result is as follows. Consider a simple model in which only the mean differs across both regimes. The QMLE estimator for the regime dependent mean can be represented as:

$$\widehat{\mu}_r \propto \frac{1}{T} \sum \widehat{\pi}^r_{t|T} y_t;\ r = \{0, 1\} \tag{3.1}$$

where $\widehat{\pi}^r_{t|T}$ has been constructed assuming $\varepsilon_{r,t}$ is N(0,1). The estimate of μ_r is the result of a GLS regression of y on a constant using the weights $\left(\widehat{\pi}^r_{t|T}\right)^{\frac{1}{2}}$. Since $\varepsilon_{r,t}$ is not normally distributed the weights used in the GLS regression are misspecified. In this context, the weights guarantee both efficiency and consistency. As a result their misspecification creates the potential for an inconsistent QMLE. We may think of $\widehat{\pi}^r_{t|T}$ as $\widehat{\pi}^{r,*}_{t|T}+\eta_t$ where $\widehat{\pi}^{r,*}_{t|T}$ is the smoothed probability one would calculate under correct specification and η_t reflects the specification error. Accordingly, the estimator for μ_r takes the form:

$$\widehat{\mu}_r \propto \frac{1}{T} \sum \widehat{\pi}^{r,*}_{t|T} y_t + \frac{1}{T} \sum \eta_t y_t;\ r = \{0, 1\}. \tag{3.2}$$

Using the consistency properties of a correctly specified MLE, the first summation converges to a constant proportional to μ_r. Hence, whenever y and η are correlated, μ_r will be inconsistently estimated. In this way, the inconsistency of the QMLE can be understood in terms of the classic econometric problem that arises whenever a regressor is correlated with the residual. We formalize this notion in the following proposition.

Proposition 3.1 *QMLE estimation of regime switching models as specified in (2.1) leads to inconsistent estimates of the model parameters* $(\beta_0, \beta_1, \sigma_0, \sigma_1, P_{00}, P_{11})$.

Proof. See Appendix. ■

The formal proof proceeds by counterexample. A single regime switching process is defined such that the QMLE is inconsistent. It is important to stress that while this result formally shows that assuming within regime normality is not an innocuous assumption there may well be cases where the QMLE is consistent. Establishing consistency will hinge on demonstrating that there is no correlation between the specification error, η, and y.

4 Semiparametric Estimation

Below we discuss semiparametric identification of regime switching models and then we propose a method for estimating regime conditional distributions that allow for unspecified departures from normality[9].

4.1 Model Identification

In what follows we consider a specific version of model (2.1) in which the distribution of $\varepsilon_{r,t}$ is allowed to exhibit unspecified departures from normality and x_t only contains a constant. While this is a restrictive assumption many of the following results can be tediously (though easily) extended to cases where the conditional mean depends on a set of covariates. Additionally, the main focus of this paper is how to identify and estimate the shape of the regime conditional distribution $f_r(\varepsilon_{r,t})$ and so we focus on a model in which this is the only object to be estimated. Before discussing model estimation it is important to consider model identification. Given the generality of the model under consideration it is not immediately clear that one model can always be distinguished from another "false" model. In fact it is easy to construct an example in which the true model is indistinguishable from a second false model. Consider the following simple model (M).

$$\begin{aligned} y_t &= \varepsilon_{r,t} \\ \varepsilon_{0,t} &\sim f_0(\cdot)\ (iid). \\ \varepsilon_{1,t} &\sim f_1(\cdot)\ (iid). \\ P_{00} &= 1-P \\ P_{11} &= P \end{aligned}$$

A straightforward consequence of the fact that $P_{11} = 1 - P_{00}$ is that the distribution of y_t is simply a static mixture of $f_0(y_t)$ and $f_1(y_t)$ with weights $1-P$ and P,

$$f(y_t|\Omega_{t-1}) = f(y_t) = (1-P)f_0(y_t) + Pf_1(y_t).$$

[9] Throughout, we use the term semiparametric to mean that the process for the regimes is taken parametrically (i.e., K-regime Markov), but the distribution of the observable, conditional on the regime is modeled nonparametrically.

Now consider an alternative model ($\widetilde{M}$) in which the within regime distributions are simply convex combinations of the previous within regime distributions,

$$\begin{aligned} y_t &= \varepsilon_{r,t} \\ \widetilde{\varepsilon}_{0,t} &\sim (1-\delta)f_0(\cdot) + \delta f_1(\cdot)\ (iid). \\ \widetilde{\varepsilon}_{1,t} &\sim (1-\gamma)f_0(\cdot) + \gamma f_1(\cdot)\ (iid). \\ P_{00} &= 1-\widetilde{P} \\ P_{11} &= \widetilde{P} \\ \widetilde{P} &= \frac{-\delta}{\gamma-\delta} + \frac{P}{\gamma-\delta} \end{aligned}$$

for some values of (δ, γ) inside the unit square. Algebraic manipulation readily yields that $\widetilde{f}(y_t|\Omega_{t-1}) = f(y_t|\Omega_{t-1})$. Accordingly, the true model (M) is distributionally equivalent to the alternative model ($\widetilde{M}$). This simple example highlights the fact that a necessary condition for model identification is that the regimes must exhibit some form of persistence, i.e. $P_{00} \neq 1 - P_{11}$. Persistent regimes imply that the conditional density ($f(y_t|\Omega_{t-1})$) exhibits dynamics necessary for identification. Below we state a proposition concerning the sufficiency of this restriction for identification of a certain class of two state regime switching models.

Proposition 4.1 *Suppose* $f_r(y_t|\Omega_{t-1}) = f_r(y_t) \equiv f_r(\varepsilon_{r,t})$ *for* $r = 0, 1$ *with* $f_0 \neq f_1$ *and that* $P_{00} \neq 1 - P_{11}$ *then the model* $M = \{f_0, f_1, P_{00}, P_{11}\}$ *is identified in the sense that there does not exist another model* $\widetilde{M} = \{\widetilde{f}_0, \widetilde{f}_1, \widetilde{P}_{00}, \widetilde{P}_{11}\}$ *such that* $f(y_t|\Omega_{t-1}) = \widetilde{f}(y_t|\Omega_{t-1})$ *except for the trivial re-classification of regimes* $\widetilde{M} = \{f_1, f_0, P_{11}, P_{00}\}$ *i.e., re-labeling regime 1 as regime 0 and vice versa.*

Proof. See Appendix. ■

While the details of the proof are not instructive and hence relegated to the appendix, its basic structure is helpful in understanding the primary source of identification. The proof proceeds first by demonstrating that any alternative model $\widetilde{M}$ must posses the following properties,

$$\begin{aligned} \widetilde{f}_0(y_t) &= (1-\delta)f_0(y_t) + \delta f_1(y_t) \\ \widetilde{f}_1(y_t) &= (1-\gamma)f_0(y_t) + \gamma f_1(y_t) \\ \widetilde{\pi}^1_{t+1|t} &= -\frac{\delta}{(\gamma-\delta)} + \frac{1}{(\gamma-\delta)}\pi^1_{t+1|t} \end{aligned} \tag{4.2}$$

where (δ, γ) lie in the unit square. The first and second property establish that all feasible alternative models are simply rotations of the true model and the last property imposes a linear structure between the one-step ahead forecast probabilities of the true and alternative model. The remainder of the proof exploits the dynamic structure of $\pi_{t+1|t}$ and shows that the only values of

(δ, γ) consistent with the law of motion for $\pi^1_{t+1|t}$ imposed by the regime switching structure are $(1,0)$ and $(0,1)$.

Having established that misspecification of $f_r(\cdot)$ yields an inconsistent QMLE and the conditions under which regime switching models with general structures for $f_r(\cdot)$ are identified (i.e., $P_{00} \neq 1 - P_{11}$); we now introduce a semiparametric estimator for the regime switching model (M) that is robust to within regime non-normality.

4.2 The Discrete Case

First we consider a simple class of regime switching models to motivate the intuition behind the more general estimation procedure. Consider a two regime model in which the support of y is discrete. Using the previous notation, the model can be characterized as $M = \{p_0 \equiv \{p_{k,0}\}_{k=1}^{K_0}, p_1 \equiv \{p_{k,1}\}_{k=1}^{K_1}, P_{00}, P_{11}\}$ where $p_{k,r} \equiv \Pr(y = y_k|r)$.[10] Conditional on the history of states $R \equiv \{r_0, r_1, ..., r_T\}$, the likelihood takes the form,

$$p(Y_T, R; \theta) = \rho_{r_0} p(r_1|r_0)...p(r_T|r_{T-1})(p_{1,0})^{n_{10}}...(p_{K_0,0})^{n_{K_0,0}}...(p_{1,1})^{n_{1,1}}...(p_{K_1,1})^{n_{K_1,1}} \tag{4.3}$$

where ρ_{r_0} refers to the initial probability of being in regime one or zero, $p(r_t|r_{t-1})$ refers to one of $P_{00}, 1 - P_{00}$, P_{11} or $1 - P_{11}$; $n_{k,r}$ is the number of times y_k was observed during regime $r = \{0, 1\}$ and $\sum_{k,r} n_{k,r} = T$. Since the underlying regime, r_t, is latent so too is the entire history R and as a result we must integrate R out of the likelihood leaving us with the following expression for the likelihood.

$$p(Y_T; \theta) = \sum_R p(Y_T, R; \theta) \tag{4.4}$$

where $\sum_R$ denotes summation over every possible history of regimes. In principle, θ, could be obtained by directly maximizing the above expression. In practice, however, this is intractable. A sample of only 20 observations would require the computation of $2^{20} \approx 10^6$ different summands. While this representation of the likelihood is not useful for computation, it is useful, as first noted by Hamilton (1990), in implementing the EM algorithm as a means of computing maximum likelihood estimates. Below we briefly outline the mechanics of the EM algorithm as it relates to this problem and we direct interested readers to Hamilton (1990) and Dempster, Laird and Rubin (1976) for further details.

[10] Note that this model contains no conditional mean or volatility dynamics apart from those generated by the regime switching channel. This can be relaxed by allowing $p_{k,r}$ to depend on the value of an indicator variable X. Accordingly, the model can be re-stated in terms of $p_{j,k,r} \equiv \Pr(y = y_k|r, X = x_j)$.

4.2.1 Model Estimation

The EM algorithm proceeds by solving a sequence of maximization programs of the form:

$$\max_{\theta_{l+1}\in\Theta} Q(\theta_{l+1}; Y_T, \theta_l) = \max_{\theta_{l+1}\in\Theta} \sum_R \ln(p(Y_T, R, \theta_{l+1})) \cdot p(Y_T, R, \theta_l) \tag{4.5}$$

where θ_l was obtained from a previous iteration or an initial value. The procedure continues until θ_l converges. Dempster, Laird and Rubin show, under general conditions, that the algorithm converges to a maximum of the likelihood. In particular they show that satisfying the first-order conditions of (4.5) is equivalent to satisfying the first-order conditions of (4.4) and that each successive iteration of (4.5) results in an increase in the value of the likelihood (4.4).

In the case under consideration here, $Q(\theta_{l+1}; Y_T, \theta_l)$ can be decomposed as:

$$\begin{aligned} &\sum_R \ln(\rho_{r_0}^{l+1}) \cdot p(Y_T, R, \theta_l) + \sum_R \sum_{t=1}^{T} \ln(p^{l+1}(r_t|r_{t-1})) \cdot p(Y_T, R, \theta_l) \\ &\sum_R \sum_{k=1}^{K} \sum_{t=1}^{T} 1\{y_t = y_k\} 1\{r_t = r\} \ln(p_{k,r}^{l+1}) \cdot p(Y_T, R, \theta_l) \end{aligned} \tag{4.6}$$

and $Q(\theta_{l+1}; Y_T, \theta_l)$ is maximized subject to the constraint that each regime conditional distribution sums to unity ($\sum_{k=1}^{K} p_{k,r} = 1;\ r = \{0, 1\}$). Note that $Q(\theta_{l+1}; Y_T, \theta_l)$ is separable in terms of the initial regime probability (ρ_{r_0}), the transition probabilities ($p(r_t|r_{t-1})$) and the regime conditional distributions ($p_{k,r}$). Furthermore, the constraints do not involve any parameters except for those related to the regime conditional distributions ($p_{k,r}$). As a result, the first order conditions for ρ_{r_0} and $p(r_t|r_{t-1})$ can be solved independently of those for $p_{k,r}$. Following Hamilton (1990) it is shown in the appendix that the following recursions characterize the EM algorithm for the model parameters

Algorithm 4.2

$$P_{00}^{l+1} = \sum_{t=2}^{T} \Pr(r_t = 0, r_{t-1} = 0|\Omega_T, \theta_l) \cdot [\sum_{t=2}^{T} \Pr(r_{t-1} = 0|\Omega_T, \theta_l)]^{-1} \tag{4.7}$$

$$P_{11}^{l+1} = \sum_{t=2}^{T} \Pr(r_t = 1, r_{t-1} = 1|\Omega_T, \theta_l) \cdot [\sum_{t=2}^{T} \Pr(r_{t-1} = 0|\Omega_T, \theta_l)]^{-1} \tag{4.8}$$

$$\rho_{r_0}^{l+1} = \Pr(r_0 = 0|\Omega_T, \theta_l) \tag{4.9}$$

$$\widehat{p}_{k,r}^{l+1} = \frac{1}{\sum_{t=1}^{T} \widehat{\pi}_{t|T}^{r}} \sum_{t=1}^{T} 1\{y_t = y_k\} \widehat{\pi}_{t|T}^{r} \tag{4.10}$$

where $\Pr(r_t = r, r_{t-1} = r|\Omega_T, \theta_l)$, $\Pr(r_{t-1} = 1|\Omega_T, \theta_l)$ and $\Pr(r_0 = 1|\Omega_T, \theta_l)$ are posterior probabilities that are conditioned on the full information set Ω_T. The expressions for these objects can be found in Hamilton (1990).

The recursion for all model parameters is initiated with an initial value $\theta_0 \equiv \left(p_0^0, p_1^0, P_{00}^0, P_{11}^0, \rho_{r_0}^0\right)$ and the equations constituting Algorithm 4.2 are iterated upon until convergence is achieved. Although the MLE for $p_{k,r}$ is a recursive set of nonlinear equations, the expression for $p_{k,r}$ is quite intuitive. In the case that the regimes are observable, the MLE for $p_{k,r}$ would simply be a histogram, $p_{k,r} = \frac{\sum 1\{y_t=y_k\}1\{r_t=r\}}{\sum 1\{r_t=r\}}$. In the case that regimes are unobserved, we replace $1\{r_t = r\}$ with the posterior probability that regime r was realized at time t. This form of the estimator also accords with the intuition behind Lemma 2.2. Once convergence is achieved the estimator takes the form, $\widehat{p}_{k,r} = \frac{1}{\sum \widehat{\pi}_{t|T}^r} \sum \left[1\{y_t = y_k\}\widehat{\pi}_{t|T}^r\right]$. Under the true parameters θ_0 we have shown in Lemma 2.2 that $E\left[1\{y_t = y_k\}\pi_{t|T}^r\right][\pi^r]^{-1} = E\left[1\{y_t = y_k\}|r_t = r\right] = p_{k,r}$. The MLE simply replaces population parameters θ_0 with their estimates $\widehat{\theta}$ and computes sample averages.

In light of the previous discussion concerning model identification, as long as $P_{00} \neq 1 - P_{11}$ the model $M = \{p_0 \equiv \{p_{k,0}\}_{k=1}^{K_0}, p_1 \equiv \{p_{k,1}\}_{k=1}^{K_1}, P_{00}, P_{11}\}$ is asymptotically identified. Given, asymptotic identification and other regularity conditions (see for example, Handbook of Econometrics, Ch. 38, Vol. 4) the MLE is both consistent and asymptotically normal. Accordingly, standard errors may be computed as usual. We should note that the claim of asymptotic normality stands in stark contrast to the Gaussian case, i.e. model (2.1) with $f_r(\cdot) = N(0, 1)$. In the Gaussian case one can show that the maximum of the likelihood does not exist, (i.e. no Type I MLE exists). The non-existence arises from allowing the variance parameters to exist in the half open interval $(0, c]$. In every sample, if the mean of regime 1, for example, is set to y_1 and σ_1 is allowed to converge towards zero then the likelihood becomes unbounded. As a result, no maximum exists. Kiefer (1978) shows that a consistent Type II MLE exists (i.e., asymptotically there exists a unique solution to the FOC in a closed neighborhood around the true parameter values.) in the case $P_{00} = 1 - P_{11}$, but that result has not yet been extended to the case considered here ($P_{00} \neq 1 - P_{11}$). The construction of the model in this context escapes this problem by considering a discrete support which guarantees that the likelihood is always bounded by unity.

4.3 The Continuous Case

Now we turn our attention to the more complex task of constructing estimators for $\theta \equiv \{\theta^1, \theta^2\} \equiv \{P_{00}, P_{11}\}, \{f_0(y_t), f_1(y_t)\}$ without nesting $f_r(y_t)$ within a finite dimensional parametric family. Before describing the proposed estimation procedure we build some intuition for the estimator by considering a locally weighted likelihood approach to estimating a simple univariate density. In a nonparametric setting, it is difficult to think about maximizing a likelihood since the likelihood is only available if we have a parametric form for $f(y_t; \theta)$. Instead of thinking about the likelihood $f(Y_T; \theta)$, we will define the notion of a pseudo-likelihood $\widetilde{f}(Y_T; \theta)$. Consider the case in which y is an *iid* multinomial random variable. Abstracting from the regime switching set up, the likelihood

of T observations can be written as:

$$p(Y_T;\theta) = (p(y_1))^{n_1}(p(y_2))^{n_2}...(p(y_k))^{n_K}$$

where $p(y_k)$ is the probability of observing $y = y_k$ and n_k is the number of times y_k was observed over T periods. Note that we can also represent $p(Y_T, \theta)$ in the following manner:

$$p(Y_T;\theta) = \prod_{j=1}^{K}\prod_{t=1}^{T} p(y_j)^{w_t^j}$$

where the weighting function, w_t^k, takes the particular form $w_t^k = 1(y_t = y_k)$. Now consider the case where we wish to estimate $f(y)$ and the support is continuous. Consider a partition of y, $\{y_1, y_{2,...,}y_K\}$ and the associated partition of function values $\{f(y_1), f(y_2), ..., f(y_K)\}$, also let $\Delta y_k \equiv y_k - y_{k-1}$. Now we define the sample pseudo likelihood as follows:

$$\widetilde{f}(Y_T;\theta) = \prod_{j=1}^{K}\prod_{t=1}^{T} f(y_j)^{w_t^j}$$

where w_t^k is a weighting function. A convenient choice for the weighting function is $w_t^k = \frac{1}{h}K(\frac{|y_t - y_k|}{h})\Delta y_k$ where $K(\cdot)$ is a symmetric density. Note that if we were to choose the weighting function $w_t^k = 1(y_t = y_k)$ and y was in fact discrete then the pseudo likelihood, $\widetilde{f}(Y_T;\theta)$, and the actual likelihood, $p(Y_T;\theta)$, would coincide. Before extending this analysis to the regime switching case, it is interesting to note the relation between MLE of $p(Y_T;\theta)$ and maximization of $\widetilde{f}(Y_T;\theta)$. In the first case we wish to choose $\theta = \{p_1, p_2, ..., p_K\}$ to maximize $p(Y_T;\theta)$ subject to the constraint that $\sum_{k=1}^{K} p_k = 1$. The resulting MLE is simply $\widehat{p}_k = n_k/T$, i.e. a histogram estimator. Now in the case of the pseudo likelihood our objective is to choose $\theta = \{f(y_1), f(y_2), ..., f(y_K)\}$ in order to maximize $\widetilde{f}(Y_T, \theta)$ subject to the constraint $\sum_{k=1}^{K} f(y_k)\Delta y_k = 1$. Accordingly, we set up the lagrangian and take first derivatives:

$$L = \log(\widetilde{f}(Y_T;\theta)) - \mu(\sum_{k=2}^{K} f(y_k)\Delta y_k - 1)$$

$$\frac{\partial L}{\partial f(y_k)} = \frac{1}{f(y_k)}\sum_{t=1}^{T} w_t^k - \mu\Delta y_k.$$

Using the constraint we can see:

$$\mu = \sum_{k=1}^{K}\sum_{t=1}^{T} w_t^k,$$

and if we recall that $w_t^k = \frac{1}{h}K(\frac{|y_t - y_k|}{h})\Delta y_k$ where $K(\cdot)$ is a symmetric density and if we assume that the partition on y is taken to be very fine then:

$$\mu = \sum_{t=1}^{T}\sum_{k=1}^{K}\frac{1}{h}K(\frac{|y_t - y_k|}{h})\Delta y_k =$$

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{1}{h}K(\frac{z_{t,k}}{h})\Delta z_k \approx$$

$$\sum_{t=1}^{T}\int_{-\infty}^{\infty}\frac{1}{h}K(u_t)du_t \cdot h = \sum_{t=1}^{T}1$$

$$\mu = T.$$

Now setting the derivative of the lagrangian to zero:

$$\frac{1}{f(y_k)}\sum_{t=1}^{T}w_t^k = \mu\Delta y_k$$

$$\frac{1}{f(y_k)}\sum_{t=1}^{T}\frac{1}{h}K(\frac{|y_t - y_k|}{h})\Delta y_k = \mu\Delta y_k$$

$$\frac{1}{f(y_k)}\sum_{t=1}^{T}\frac{1}{h}K(\frac{|y_t - y_k|}{h}) = T$$

$$\widehat{f}(y_k) = \frac{1}{Th}\sum_{t=1}^{T}K(\frac{|y_t - y_k|}{h}).$$

The resulting estimator is the standard Nadarya-Watson kernel density estimator. Viewed in this light, it can be seen that maximizing a pseudo likelihood produces a smoothed histogram as a density estimator in contrast to the discrete support case, in which case, the MLE is the traditional histogram.

4.3.1 Model Estimation

Now our aim is to incorporate the pseudo likelihood into the regime switching context. We define the pseudo regime conditional likelihood as follows:

$$\widetilde{f}(Y_T, R; \theta) = \rho_{r_0}p(r_1|r_0)...p(r_T|r_{T-1})\prod_{k=1}^{k=K}\prod_{t=1}^{T}f(y_k|r=0)^{w_t^{0,k}}\prod_{k=1}^{k=K}\prod_{t=1}^{T}f(y_k|r=1)^{w_t^{1,k}} \tag{4.11}$$

where now the weight function takes the form, $w_t^{r,k} = \frac{1}{h}K(\frac{|y_t-y_k|}{h})1\{r_t = r\}\Delta y_k$. The unconditional pseudo likelihood is simply given by $\widetilde{f}(Y_T;\theta) = \sum_R \widetilde{f}(Y_T, R, \theta)$. Note that the pseudo-likelihood is composed of two parts. The first is completely parametric and represents the probability of a given regime path (R). The second component is nonparametric and represents the pseudo-likelihood of Y_T given R. We maximize $\widetilde{f}(Y_T;\theta)$ in the same manner as we maximized $p(Y_T;\theta)$ in the discrete case, namely we employ the EM algorithm.

As in the discrete case, note that $\ln(\widetilde{f}(Y_T, R; \theta))$ can be separated into three components: $\ln(\widetilde{f}(Y_T, R; \theta)) = \ln(\rho_{r_0}) + \sum_{t=2}^{T} \ln(p(r_t|r_{t-1})) + \sum_{t=1}^{T} \sum_{r=0}^{1} \sum_{k=1}^{K} w_t^{r,k} \log(f(y_k|r))$. The first component represents the contribution to the likelihood from the initial regime probability, the second reflects the contribution from the transition probabilities and the third the contribution from the regime conditional pseudo-likelihood for y. Since the log pseudo likelihood is additively separable and since only the portion representing the contribution from $f_r(y_k)$ differs from the parametric case, determining the initial regime vector and the transition probabilities does not differ from the discrete case (Algorithm 4.2).

All that remains is to obtain expressions for $\{f_0(y_1), ..., f_0(y_K); f_1(y_1), ..., f_1(y_K)\}$. We show in the appendix that the appropriate recursion for $f(y_k|r)$ is given by:

$$\widehat{f}^{l+1}(y_k|r) = \frac{1}{h} \sum_{t=1}^{T} K(\frac{|y_t - y_k|}{h_0}) \frac{\widehat{\pi}^r_{t|T}}{\sum_t \widehat{\pi}^r_{t|T}}. \tag{4.12}$$

Estimation proceeds by choosing, $P^0_{00}, P^0_{11}, f^0_0(\cdot), f^0_1(\cdot)$ and then iterating on the above equation as well as the recursions for $P^l_{r,r}$. The fully semiparametric estimator "smooths out" the discrete estimator through the use of a smooth weighting function $K(\frac{|y_t - y_k|}{h_0})$ rather than the discontinuous weighting function $1(y_t = y_k)$. Here we note that while it is quite intuitive, this estimator is considerably more complex than the estimator in the discrete case due to the fact that part of the model space is finite dimensional, $\theta_1 = (P_{00}, P_{11})$, and part is infinite dimensional, $\theta_2 = (f_0(\cdot), f_1(\cdot))$. Here we make no claim to establish consistency rates or asymptotic distribution theory for the fully semiparametric estimator. We only remark that due the significant nonlinear nature of the model that this is a challenging task that we leave to future research.

5 U.S. Short Term Interest Rate: Empirical Results

5.1 Motivation and Model Specification

Interest rates have received considerable attention in the regime switching literature, Ang and Bekaert (2001), Bansal and Zhou (2002), Dahlquist and Gray (2000), Garcia and Perron (1996), Gray (1996), Cai (1994). The widespread application of regime switching models to interest rates stems from the natural association between the notion of regimes that underlie the econometric model and the large economy-wide shocks that have strong and persistent influences on the behavior of interest rates. In this way, the regime structure of the model is more than a mere device used to filter the data. For example, Ang and Bekaert (2001) argue that the regime classification in a two regime model of U.S. nominal short term rates corresponds reasonably well with business cycles. Bansal and Zhou (2002) draw a similar conclusion in their regime switching analysis of the U.S.

term structure. Dahlquist and Gray (2000) argue that the regimes identified in their study of a sample of EMS countries reflects changes in monetary policy regimes as central banks attempt to maintain pre-specified currency target zones. In this way, the regime structure of the model makes a close connection with the economics of interest rate determination.

In what follows we will focus on the following model of the weekly U.S. short term rate between 1970 and 1994:

$$\begin{aligned} \Delta i_t &= \alpha_{0r} + \alpha_{1r} i_{t-1} + \sigma_r \varepsilon_t;\ r = \{0,1\} \\ \varepsilon_{r,t} &\sim f_r(\cdot)\ (iid) \end{aligned} \tag{5.13}$$

where P_{00} and P_{11} are defined as in (2.1) and $f_r(\cdot)$ refers to a well-behaved, smooth density function. The specification described above with $f_0(\cdot) = f_1(\cdot) = N(0,1)$ corresponds to Gray's (1996) study of the one-month U.S. Treasury bill rate rate. The specification allows for mean reversion in interest rates that varies across regimes. Table 1 reproduces Gray's model estimates. For comparative purposes we use the exact same data used by Gray throughout the paper. Figure 1 presents plots of the U.S. short rate (i_t) between 1970 and 1994, the weekly difference in the short rate (Δi_t), the estimated smoothed probability of regime 0 ($\widehat{\pi}^0_{t|T}$) and a histogram of Δi_t. Readers interested in a more detailed description of the data are referred to Gray (1996).

Figure 1 About Here

Before discussing the model results we highlight some important features of the data. Looking at the time series plot of Δi_t, it is apparent that the series displays considerable and persistent heteroscedasticity. The periods surrounding the OPEC oil crises, the Volcker monetary policy regime and the period surrounding the October stock market crash of 1987 all display increased variability. Looking at the histogram of Δi_t, it appears that the most striking feature of the unconditional distribution is extreme leptokurtosis (b_2=28.4). As the time series switches between periods of high and low variability the unconditional distribution inherits a tall peak near the origin and thick tails.

Table 1 About Here

Examining Table 1 shows that the two regimes are characterized by high levels of persistence ($P_{00}, P_{11} > 0.95$) and widely differing levels of volatility ($\sigma_0 = 0.6716$, $\sigma_1 = 0.1496$). Examining the plot of Δi_t and $\widehat{\pi}^0_{t|T}$ in Figure 1 shows that the model identifies regime 0 almost exclusively with the aforementioned instances of increased interest rate variability. The model fails to recognize any meaningful mean dynamics given the small estimates of the autoregressive parameters. Additionally, the intercept parameters are indistinguishable from zero across both regimes, implying that the short term interest rate may be loosely characterized as a driftless, heteroscedastic random

walk. In what follows, we will make use of a simplified version of model (5.13) which assumes that $\alpha_{0,0} = \alpha_{0,1} = \alpha_{1,0} = \alpha_{1,1} = 0$.

A conclusion that mean reversion is unimportant based on the estimates of a parametric model needs to be reconciled with the fact that model estimates are inconsistent if (as we will shortly argue) $f_r(\cdot)$ is misspecified. The rationale for abstracting from any conditional mean dynamics is as follows. First, though the QMLE may not be consistent it is not completely uninformative. Even contemplating a range of mean reversion parameters within two QMLE standard errors of the sample estimates would not imply very strong mean reversion. Second, the question of mean reversion in interest rates is simply beyond the scope of this paper. Rather, our main goal is to explore whether or not there are any interesting regime dynamics in higher order moments (i.e., skewness and kurtosis) in the distribution of short term interest rate shocks.

The assumption that volatility is constant once the regime has been controlled for is, perhaps, more objectionable. Any attempt to characterize the regime conditional distribution of interest rate shocks should explicitly recognize any within regime volatility dynamics. Moreover, excess kurtosis or skewness in $f_r(\cdot)$ may easily be confused with unmodeled volatility dynamics. Many researchers who have previously explored interest rates in a regime switching context have allowed for within regime volatility dynamics. Popular specifications of the regime conditional volatility function include low order ARCH specifications, $\sigma_{r(t)} = \omega_{r(t)} + \alpha_{r(t)}\varepsilon_{t-1}^2$, as in Cai (1994) and regime dependent CIR processes, $\sigma_{r(t)} = \omega_{r(t)} + \beta_{r(t)}\sqrt{i_{t-1}}$, as in Dahlquist and Gray (2000) as well as more complex GARCH specifications as in Gray (1996) and Ang and Bekaert (2000). It should be noted, however, that while within regime volatility dynamics are often included it is not clear that they are necessary once regime dependent level effects in volatility have been recognized as in the above specification (5.13). Gray (1996) reports, "[t]he squared standardized residuals [exhibit] no evidence of serial correlation. The simple regime switching model can capture much of the stochastic volatility of short term interest rates." In all of the studies cited above, the evidence for complex volatility dynamics is sharply reduced after controlling for regime effects in the level of volatility. In light of these findings, we take the simple regime switching specification (5.13) to be an adequate point of departure for our exploration into the regime conditional distribution of interest rate shocks. Before considering semiparametric alternatives to the parametric regime switching model we investigate the validity of the within regime normality assumption using the extended Jarque-Bera test.

5.2 Empirical Results of the Normality Test

Before considering the extended JB test (JB_{RS}), we examine the estimated within regime skewness and kurtosis as well as the informal Jarque-Bera statistics as an informal way of checking the adequacy of the normality assumption.

Table 2 About Here

Table 2 contains point estimates of within regime skewness and kurtosis as given by equations (2.5), (2.6) as well as the informal JB statistics. Examining these estimates shows that the amount of skewness within each regime is small relative to the estimated kurtosis. The kurtosis in the low volatility regime ($\widehat{S}_r = 3.70$) is roughly consistent with normality. The estimated kurtosis in the high volatility regime ($\widehat{S}_r = 8.5$), however, seems too extreme to reconcile with the normality assumption.

Evidence of excess kurtosis in the more volatile regime may help to explain one of the simple regime switching models largest shortcomings. The lower panel of Table 1 reports the estimated value of unconditional kurtosis generated by the parametric regime switching model (K=9.62). Estimation uncertainty aside, this value seems too small relative to the sample estimate of unconditional kurtosis from the data ($\widehat{K}$=28.38). Note that in a Gaussian regime switching model with a constant mean and switching variances (i.e, the restricted form of model (5.13)) unconditional kurtosis is given by,

$$\begin{aligned} b_2 &= \frac{a}{b}, \\ a &= 3\pi^1(1-\pi^1)(\sigma_0^2-\sigma_1^2)^2, \\ b &= \left[\sigma_0^2+\pi^1(\sigma_1^2-\sigma_0^2)\right]^2, \end{aligned} \tag{5.14}$$

where π^1 denotes the ergodic (unconditional) probability of regime 1.[11] Notice that the only mechanism for generating excess kurtosis is by varying π^1 or $\sigma_0^2 - \sigma_1^2$. Allowing for the possibility of within regime excess kurtosis in the distribution of $\varepsilon_{r,t}$ can improve the model's ability to match the amount of kurtosis in the data. It is straightforward to show that allowing for excess kurtosis within each regime leads to the following expression for unconditional kurtosis,

$$b_2' = b_2 + \pi^0\left(\frac{\sigma_0^2}{\sigma^2}\right)^2 b_2^0 + \pi^1\left(\frac{\sigma_1^2}{\sigma^2}\right)^2 b_2^1, \tag{5.15}$$

where $\sigma^2 = \sigma_0^2 + \pi^1(\sigma_1^2 - \sigma_0^2)$ and $b_2^r = \frac{E[\varepsilon_r^4]}{E[\varepsilon_r^2]^2} - 3$ where ε_r represents the residual from regime r. Observe that the unconditional kurtosis within each regime, b_2^r, is amplified by the square of the ratio of the regime conditional to the unconditional variance $\left(\frac{\sigma_r^2}{\sigma^2}\right)$. In a two regime setting, this implies that the regime with the larger within regime variance will have a substantially larger effect on unconditional kurtosis. Accordingly, Gaussian regime switching estimates of unconditional kurtosis that fall short of the kurtosis in the data may signal that the more volatile regime is also more leptokurtic.

Turning attention to the extended JB tests in Table 2, the JB_{RS}^0 and JB_{RS}^1 statistics test the null of normal residuals within regime 0 and 1, respectively, and are asymptotically distributed $\chi^2(2)$.

[11] Timmerman (2000) provides expressions for the moments of regime switching models.

The JB_{RS} statistic tests the null of normal residuals across both regimes and is asymptotically distributed $\chi^2(4)$. Examining the JB^0_{RS} statistic (338.12) casts further doubt on the normality assumption. While the corresponding normality test for regime 1 (JB^1_{RS}) also rejects the null at any reasonable significance level, the size of the discrepancy is much larger for regime 0 (338.12 vs. 92.42) than for regime 1. Qualitatively, based on these tests, the normality assumption appears to be a more reasonable approximation to the distribution of the residuals for regime 1 than for regime 0. It is important to note, however, that while the data rejects a model of within regime normality ($JB_{RS} = 428.68$), comparing the size of the modified JB test to the standard (single regime) JB test (see Figure 1) computed from the raw data ($JB = 34,465$) suggests that introducing multiple regimes goes a long way towards improving the model's specification.

5.3 Estimation Results When $f_r(\cdot)$ is Discrete

We investigate the regime dependent distribution of interest rate shocks (Δi_t) using the discretized regime switching model.[12] The model is defined as $M = \{p_0, p_1, P_{00}, P_{11}\}$ and is estimated using Algorithm 4.2. [13] The initial parameter θ^0 was taken to be the one implied by the results of the parametric model (See Table 1).

Table 3 About Here

Figure 2 About Here

Table 3 shows model estimates and the probabilities that would obtain under normality using Gray's (1996) estimates from Table 1.[14] Figure 2 plots an overlay of the estimated histograms and the normal density implied by Gray's (1996) estimates. The estimated histograms confirm the qualitative results of the extended Jarque-Bera tests. The table shows that the less volatile regime accords closely with normality while the more volatile regime appears considerably more peaked in the middle and thinner in the tails than the corresponding normal distribution. It also appears that the difference between the distribution of Δi_t in the volatile regime is driven by more than a few outliers. If this were the case, one would not expect such persistent deviation from normality in the center of the distribution.

These results suggest that higher order moments (e.g., skewness and kurtosis) of interest rate shocks (Δi_t) vary across regimes. This calls into question simple location scale (e.g, GARCH(p,q)) models of interest rate shocks that only assume time variation in conditional volatility. Models such as Hansen's (1994) Autoregressive Conditional Density (ACD), provide for richer dynamics

[12] We discretize the domain of Δi_t into the following set $D = \{(-\infty, -0.80), [-0.80, -0.60), [-0.60, -0.40), ..., [0.80, \infty)\}$

[13] In all empirical applications we take r_0 to be an independent draw from the ergodic distribution of regimes, i.e. $\Pr(r_0 = 1) = \frac{(1-p_{00})}{2-P_{11}-P_{00}}$, hence ρ_{r_0} is not treated as a free parameter.

[14] As noted previously, $\alpha_{0,r}, \alpha_{1,r}$ are set to zero.

that account for time variation in volatility, skewness and kurtosis but are hampered by the need to specify a separate time series model for each moment. This modeling strategy typically results in a large number of parameters to be estimated. The current model builds on the ACD approach by allowing for variation in higher order moments that are driven by a single shock (r_t).

While these results are suggestive of a regime that is reasonably approximated by normality and one that is not, it is difficult to compare previous results that employ continuous distributions with the current discrete model. We now turn to the estimation results for the case of continuous regime conditional distributions in order to make a more complete comparison with earlier research.

5.4 Estimation Results When $f_r(\cdot)$ is Continuous

Figure 3 shows plots of the two regime dependent distributions for the weekly change in the U.S. short rate using the method outlined above overlayed against the normal distribution implied by Gray's (1996) estimates. The estimation algorithm was begun by using the final estimates from the discrete model in the previous section as the initial parameter vector $(P_{00}, P_{11}, f_0(\Delta i_t), f_1(\Delta i_t))$. Then, the modified EM algorithm was carried out until convergence was achieved. In effect, Figure 3 simply smooths out the histogram estimates from Figure 2. The less volatile regime's estimated distribution fits rather closely with the associated Gaussian distribution except near the peak. The more volatile regime shows more significant signs of misspecification. The continuous density estimates confirm what the earlier analysis has shown. The less volatile regime is well characterized by a Normal distribution but the more volatile regime exhibits negative skewness and thicker tails than a normal density. Again, these results suggests that interest risk is only adequately characterized by the variance of interest shocks during periods of low volatility. During periods of excessive interest rate variability higher order moments are also important in characterizing interest rate risk. In particular, these empirical findings would be of importance to the pricing of interest rate sensitive securities. These results suggest that pricing would critically depend on whether or not (or the relative likelihood) the economy was in the midst of a calm or volatilie interest rate regime.

Figure 3 About Here

6 Conclusion

This paper has developed a set of diagnostic tools and tests that can be used to shed light on the plausibility of the normality assumption in a regime switching model. These diagnostics, when applied to U.S. short term interest rate shocks (Δi_t), cast doubt on the normality assumption in a two regime model. An extension of the Jarque-Bera test to the two regime setting rejects the

null of normality at all reasonable significance levels. Additionally, we show that QML estimation of regime switching models is inconsistent. In light of the need for more general alternatives to the Gaussian regime switching model, we show that semiparametric alternatives are identified and propose two different estimators. When these estimators are applied to the U.S. short rate series the estimator of the more volatile regime reveals a distribution that is negatively skewed and leptokurtic relative to the normal distribution. Other researchers who have examined and rejected the Gaussian framework for the short rate (e.g., Thompson (2000)) have argued that a process with unconditionally fat tails such as a Levy process would provide a better approximation to the short rate process. These findings suggest that fat tailed innovations are only relevant during the more volatile regime. Accordingly, we suggest a model that allows for considerable excess kurtosis only during periods of high volatility.

References

[1] Ang, A., Bekaert, G. (2002), "Regime Switches in Interest Rates," *Journal of Business and Economic Statistics*,20,2,163-182.

[2] Bansal, R., Zhou, H. (2002), "Term Structure of Interest Rates Under Regime Shifts," *Journal of Finance*, October.

[3] Bera, A.K., Jarque, C.M. (1982), "Model Specification Tests - A Simultaneous Approach," *Journal of Econometrics*, 20, 59-82.

[4] Bollen, N. B.P., Gray S. F. and Whaley R. E. (2000), "Regime Switching in Foreign Exchange Rates: Evidence from Currency Option Prices," *Journal of Econometrics*, 94, 239-276.

[5] Cai, J. (1994), "A Markov Model of Switching regime ARCH," *Journal of Business and Economic Statistics*, 12, 309-316.

[6] Cechetti, S., Lam, P. and Mark, N. (1990), "Mean Reversion in Equilibrium Asset Prices," *American Economic Review*, 80, 398-418.

[7] Cechetti, S., Lam, P. and Mark, N. (2000), "Asset Pricing with Distorted Beliefs: Are Equity Returns Too Good to be True," *American Economic Review*, 90, 787-805.

[8] Campbell, S., and Li, C. (2001), "Option Pricing with Unobserved and Regime-Switching Volatility," Manuscript, University of Pennsylvania.

[9] Dahlquist, M. and Gray, S. F. (2000), "Regime-Switching and Interest Rates in the European Monetary System,"*Journal of International Economics*, 50, 399-419.

[10] Dempster, A.P., Laird N.M. and Rubin D.B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society* B, 39, 1-38.

[11] Diebold F.X, Lee J.H., Weinbach G. (1994), "Regime Switching with Time-Varying Transition Probabilities," in C. Hargreaves (ed.) *Nonstationary Time-Series analysis and Cointegration*, 283-302, New York: Oxford University Press (Reprinted in F.X. Diebold and G.D. Rudebusch Business Cycles: Durations, Dynamics, and Forecasting (1999), Princeton: Princeton University Press).

[12] Engel, C., Hamilton, J. D. (1990), "Long Swings in the Dollar: Are They in the Data and Do Markets Know It?," *American Economic Review*, 80, 689-713.

[13] Engle, R., Gonzalez-Rivera, J. D. (1991), "Semiparametric ARCH Models," *Journal of Business and Economic Statistics*, 9, 345-359.

[14] Engle, R., McFadden D. (ed.), *Handbook of Econometrics*(1994), Vol. 4, Amsterdam: Elsevier Science.

[15] Fama, E., French K. (1988), "Permanent and Temporary components of Stock Prices," *Journal of Political Economy*, 96, 246-273.

[16] Filardo, A. J. (1994), "Business-Cycle Phases and Their Transitional Dynamics," *Journal of Business and Economic Statistics*, 12, 299-308.

[17] Garcia, R. (1998), "Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models," *International Economic Review*, 39, 763-788.

[18] Garcia, R., Perron, P. (1996), "An Analysis of the Real Interest Rate under Regime Shifts," *Review of Economics and Statistics*, 78(1), 111-125.

[19] Glosten, L., Jagannathan R. and Runkle D. (1993), "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks " *Journal of Finance*, 48, 1779-1801.

[20] Gray, S. F. (1996), "Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process" *Journal of Financial Economics*, 42, 27-62.

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

[22] Hamilton J.D. (1990), "Analyis of Time Series Subject to Changes in Regime," *Journal of Econometrics*, 45, 39-70.

[23] Hamilton J.D. (1996), "Specification Testing in Markov -Switching Models," *Journal of Econometrics*, 70, 127-157.

[24] Hamilton J.D., Susmel R. (1994), "Autoregressive Conditional Heteroskedasticity and changes in Regime," *Journal of Econometrics*, 64, 307-333.

[25] Hammersley J.M., Handscomb D.C. (1964), Monte Carlo Methods. New York: John Wiley and Sons Inc.

[26] Hansen, B. (1994), "Autoregressive Conditional Density Estimation," *International Economic Review*, 35, 705-730.

[27] Hansen, B. (1992), "The Likelihood Ratio Test Under Nonstandard Conditions: Testing the Markov Switching Model of GNP," *Journal of Econometrics*, 7, 53-74.

[28] Kendall, M., Stuart A. (1973), The Advanced Theory of Statistics. New York: Hafner.

[29] Kim C.J., Nelson C. R. (1999), State-Space Models with Regime Switching. Cambridge: MIT Press.

[30] Kiefer, N., (1978), "Discrete Parameter Variation: Effieicent Estimation of a Switching Regression Model," *Econometrica*, 2, 427-434.

[31] McCurdy, T. H., Durland, M. J. (1994), "Duration Dependent Transitions in a Markov Model of U.S. GNP Growth," *Journal of Business and Economic Statistics*, 12, 279-288.

[32] Perez-Quiros, G., Timmerman, A. (2000), "Business Cycle Asymmetries in Stock Returns: Evidence from Higher Order Moments and Conditional Densities," *Journal of Econometrics*, forthcoming.

[33] Schaller, H., Van Norden S. (1997), "Regime Switching in Stock Market Returns," *Applied Financial Economics*, 7, 177-91.

[34] Silverman B.W. (1986), Density Estimation for Statistics and Data Analysis. New York: Chapman and Hall.

[35] Thompson, S. (2000), "Specification Tests for Continuous Time Models", Harvard University, Mimeo.

[36] Timmerman, A. (2000), "Moments of Markov Switching Models," *Journal of Econometrics*, 96, 75-111.

[37] Turner, C. M., Startz, R. and Nelson, C. R. (1989), "A Markov Model of Heteroskedasticity, Risk, and Learning in the Stock Market," *Journal of Financial Economics*, 13, 521-547.

[38] Whitelaw, R (2000), "Stock Market Risk and Return: An Equilibrium Approach," *Journal of Financial Economics*, 25, 3-22.

A Appendix

A.1 Proof of Lemma 2.2

Lemma 3.1 *Let $z_t = [y_t, x_t]'$. Given a regime-switching model of the form in (2.1) $E\left(\pi^r_{t|t+\tau} m(z_t)\right) = \pi^r E\left(m(z_t)|r\right)$ for any $\tau \geq 0$ and $r = 0, 1$.*

Proof.
The proof is shown for the case $\tau = T - t$. Generalizing the proof for any $\tau \geq 0$ is straightforward. Let $Z_j = (z_1, z_2, ..., z_j)$.

$$\begin{aligned}
\pi^r_{t|T} &= \Pr(r_t = r|Z_T) \\
\pi^r_{t|T} &= f(Z_T|r_t = r)\left[\frac{\pi^r}{f(Z_T)}\right] \\
\pi^r_{t|T} &= \pi^r \left[\frac{f(Z_{t-1}|r_t = r)}{f(Z_{t-1})}\right]\left[\frac{f_r(z_t|Z_{t-1})}{f(z_t|Z_{t-1})}\right]\left[\prod_{i=t+1}^{i=T} \frac{f(z_i|r_t = r, Z_{i-1})}{f(z_i|Z_{i-1})}\right]
\end{aligned}$$

Consider each term of the form $\frac{f(z_i|r_t=r,Z_{i-1})}{f(z_i|Z_{i-1})}$. It will be convenient to demonstrate:

$$E\left[\frac{f(z_i|r_t = r, Z_{i-1})}{f(z_i|Z_{i-1})}|Z_{i-1}\right] = 1$$

The result follows from observing:

$$E\left[\frac{f(z_i|r_t = r, Z_{i-1})}{f(z_i|Z_{i-1})}|Z_{i-1}\right] = \int_{-\infty}^{\infty}\left(\frac{f(z_i|r_t = r, Z_{i-1})}{f(z_i|Z_{i-1})}\right) f(z_i|Z_{i-1})dy_i = \int_{-\infty}^{\infty} f(z_i|r_t = r, Z_{i-1})dy_i = 1$$

setting $\frac{f(z_i|r_t=r,Z_{i-1})}{f(z_i|Z_{i-1})} = \theta_i$ we have

$$E\left[\pi^r_{t|T} m(z_t)\right] = \pi^r E\left[m(z_t)\left(\left[\frac{f(Z_{t-1}|r_t = r)}{f(Z_{t-1})}\right]\left[\frac{f_r(z_t|Z_{t-1})}{f(z_t|Z_{t-1})}\right]\left[\prod_{i=t+1}^{i=T}\theta_i\right]\right)\right]$$

Repeated application of the law of iterated expectations yields:

$$E\left[\pi^r_{t|T} m(z_t)\right] = \pi^r E\left[\left[\frac{f(Z_{t-1}|r_t = r)}{f(Z_{t-1})}\right] E\left[m(z_t)\left(\left[\frac{f_r(z_t|Z_{t-1})}{f(z_t|Z_{t-1})}\right]\right)|Z_{t-1}\right]\right] =$$

$$\pi^r E\left[\left[\frac{f(Z_{t-1}|r_t = r)}{f(Z_{t-1})}\right]\int_{-\infty}^{\infty} m(z_t)\left(\left[\frac{f_r(z_t|Z_{t-1})}{f(z_t|Z_{t-1})}\right]\right) f(z_t|Z_{t-1})dz_t\right] =$$

$$\pi^r E\left[\left[\frac{f(Z_{t-1}|r_t = r)}{f(Z_{t-1})}\right] E\left[m(z_t)|r_t = r, Z_{t-1}\right]\right] =$$

$$\pi^r \int_{Z_{t-1}} [f(Z_{t-1}|r_t = r)]\, E\left[m(z_t)|r_t = r, Z_{t-1}\right] \quad =$$

$$\pi^r E\left[m(z_t)|r_t = r\right]$$

■

A.2 Proof or Proposition 4.1

Proposition 4.1 *QMLE estimation of regime-switching models as specified in (2.1) leads to inconsistent estimates of the model parameters* $(\beta_0, \beta_1, \sigma_r, \sigma_1, P_{00}, P_{11})$.

Proof. We construct an example of a regime-switching process for which the Gaussian QMLE is inconsistent. The proof consists of showing that the true model parameters $\theta_0 = (\beta_0, \beta_1, \sigma_r, \sigma_1, P_{00}, P_{11})$ do not constitute a fixed point of the EM algorithm. Consider the following two regime-switching model, $r = \{0, 1\}$

$$\begin{aligned} y_t &= \varepsilon_{r,t} \\ \varepsilon_{0,t} &\sim 2\Phi_{z\geq 0}(\cdot) \\ \varepsilon_{1,t} &\sim 2\Phi_{z\leq 0}(\cdot) \\ P_{00} = P_{11} &= \frac{1}{2} \end{aligned}$$

where $2\Phi_{z0}$ refers to the (left or right) truncated standard normal distribution. As a result, note that $\sigma_0^2 = \sigma_1^2 = 1 - 4\phi(0)^2$ and $\mu_0 = 2\phi(0), \mu_1 = -2\phi(0)$. Moreover, we endow the econometrician with the knowledge that $\sigma_0 = \sigma_1 = 1 - 4\phi(0)^2$ and $P_{00} = P_{11} = \frac{1}{2}$. Accordingly, the only goal is to estimate μ_0, μ_1.

Recall that under the assumed normality of $\varepsilon_{r,t}$ the MLE for the regime-dependent means,μ_0, μ_1, is a fixed point of the following set of equations:

$$\begin{aligned} \widehat{\mu}_0 &= \frac{1}{\frac{1}{T}\sum \widehat{\pi}^0_{t|T}} \frac{1}{T} \sum \widehat{\pi}^0_{t|T} y_t \\ \widehat{\mu}_1 &= \frac{1}{\frac{1}{T}\sum \widehat{\pi}^1_{t|T}} \frac{1}{T} \sum \widehat{\pi}^1_{t|T} y_t \end{aligned}$$

the proposition is proved by showing that, in large samples, $\{\widehat{\mu}_0, \widehat{\mu}_1\} = \{\mu_0, \mu_1\}$ is not a fixed point of the above system of equations (recall that $\sigma_0, \sigma_1, P_{00}, P_{11}$) are known to the econometrician. Consider the first equation.

$$\widehat{\mu}_0 = \frac{1}{\frac{1}{T}\sum \widehat{\pi}^0_{t|T}} \frac{1}{T} \sum \widehat{\pi}^0_{t|T} y_t$$

Since $P_{00} = P_{11} = \frac{1}{2}$ the model reduces to a static mixture model. Accordingly, past and future values of y_t provide no information about the value of r_t beyond that contained in y_t. As a result,

we have $\pi^0_{t|T} = \pi^0_{t|t}$ (see Hamilton, 1994). Using this fact, we can re-express the above equation as:

$$\widehat{\mu}_0 = \frac{\frac{1}{T}\sum \widehat{\pi}^0_{t|t} y_t}{\frac{1}{T}\sum \widehat{\pi}^0_{t|t}} = \frac{A_T}{B_T}$$

We wish to show that given $\{\mu_0, \mu_1\}$, $Plim(\widehat{\mu}_0) \neq \mu_0$. We evaluate $Plim(\widehat{\mu}_0)$ by appealing to Slutsky's Rule (i.e. $Plim(\frac{A_T}{B_T}) = \frac{Plim(A_T)}{Plim(B_T)}$). First consider $Plim(A_T)$.

$$Plim(A_T) = Plim(\frac{1}{T}\sum \widehat{\pi}_{t|t} y_t) = E\left(\widehat{\pi}_{t|t} y_t\right)$$

by Kolmogorov's strong law of large numbers. We now evaluate $E\left(\widehat{\pi}_{t|t} y_t\right)$.

$$E\left(\pi_{t|t} y_t\right) = \int \widehat{\pi}^0_{t|t} y_t \frac{(\phi_{z\geq 0}(y_t) + \phi_{z\leq 0}(y_t))}{2} dy_t =$$

$$\int y_t \frac{\phi(y_t; \mu_0, \sigma)}{\phi(y_t; \mu_0, \sigma) + \phi(y_t; \mu_1, \sigma)} (\phi_{z\geq 0}(y_t) + \phi_{z\leq 0}(y_t)) dy_t =$$

$$\int y_t \phi(y_t; \mu_0, \sigma) R(y_t) dy_t = E_0\left(y_t R(y_t)\right) = c\mu_0; \text{ for some finite } c$$

where $\phi(y_t; \mu, \sigma)$ represents the normal pdf with mean μ and variance σ^2, $R(y_t) = \frac{\phi_{z>0}(y_t) + \phi_{z<0}(y_t)}{\phi(y_t; \mu_0, \sigma) + \phi(y_t; \mu_1, \sigma)}$ and E_0 refers to the expectation taken with respect to $\phi(\cdot; \mu_0, \sigma)$. Before proceeding with the proof we note some useful properties of R that are straightforward to verify. First, $R(y_t)$ is symmetric about 0 and $R'(y_t)0$ as $y_t 0$.

Turning our attention to B_T we wish to calculate $Plim(B_T) = Plim(\frac{1}{T}\widehat{\pi}^0_{t|t}) = E\left(\widehat{\pi}^0_{t|t}\right)$.

$$E\left(\widehat{\pi}^0_{t|t}\right) = \int \phi(y_t; \mu_0, \sigma) R(y_t) dy_t = E_0\left(R(y_t)\right)$$

The proof is completed by showing that $E_0\left(R(y_t)\right) \neq c$. We proceed by contradiction. Suppose that $E_0\left(R(y_t)\right) = c$. If this is the case then it must also be the case that

$$E_0\left(y_t R(y_t)\right) = E_0\left(y_t\right) E_0\left(R(y_t)\right)$$

which implies that y_t and $R(y_t)$ are uncorrelated. Since the expectation is taken with respect to $\phi(\cdot; \mu_0, \sigma)$ we can invoke Stein's Lemma (see Cochrane, (2001)) to compute $Cov_0(y_t, R(y_t))$. Stein's Lemma dictates that if y_t is normally distributed then $Cov_0(y_t, R(y_t)) = E_0\left(R'(y_t)\right)\sigma^2_{y_t}$. Since $R'(y_t) = -R'(-y_t)$, $R'(y_t) > 0$ whenever $y_t > 0$ and since $\mu_0 > 0$ it follows that $E_0\left(R'(y_t)\right) > 0$ which contradicts the maintained assumption that y_t and $R(y_t)$ are uncorrelated, thus completing the proof . ■

At this point we note two characteristics of the DGP which may be questionable to some readers. First, since $P_{00} = P_{11} = \frac{1}{2}$ there is no difference between the conditional and unconditional distribution of y_t. As a result, one could argue that recovering μ_r is not interesting. Instead the only parameter of interest is $\frac{\mu_0+\mu_1}{2}$. The transition probabilities are chosen so that we can make use of $\pi^r_{t|t}$ instead of $\pi^r_{t|T}$ in the proof which greatly simplifies the analysis. It would be tedious but trivial to extend the proof to the case where $P_{00} = P_{11} = \frac{1}{2} + \varepsilon$ in which case μ_0 is of direct interest. Secondly, the DGP consists of two distributions with only partial support. This assumption is maintained to simplify the proof and the proof can easily be extended to the case where each distribution has full support.

A.3 Proof of Proposition 5.1

First we note that the proof below only strictly applies to the case where the only element of x_{t-1}. Extending this proof to the case of a linear conditional mean only changes the details and not the basic structure of the proof.

Proposition 5.1 *Suppose* $f_r(y_t|Y_{t-1}) = f_r(y_t - X'_{t-1}\beta_r) \equiv f_r(\varepsilon_{r,t})$ *for* $r = 0, 1$ *with* $f_0 \neq f_1$ *and that* $P_{00} \neq 1 - P_{11}$ *then the model* $M = \{f_0, f_1, P_{00}, P_{11}\}$ *is identified in the sense that there does not exist another model* $\widetilde{M} = \{\widetilde{f}_0, \widetilde{f}_1, \widetilde{P}_{00}, \widetilde{P}_{11}\}$ *such that* $f(y_t|Y_{t-1}) = \widetilde{f}(y_t|Y_{t-1})$ *except for the trivial re-classification of regimes* $\widetilde{M} = \{f_1, f_0, P_{11}, P_{00}\}$ *i.e., re-labeling regime 1 as regime 0 and vice versa.*

The proof depends on three useful lemmas. They are given below and proven at the end of the proof.

Proof.

Lemma A.1 *If an observationally equivalent model,* $\widetilde{M} = \{\widetilde{f}_0, \widetilde{f}_1, \widetilde{P}_{00}, \widetilde{P}_{11}\}$*, exists then the following conditions must hold.*

$$\widetilde{f}_0(y) = (1-\delta)f_0(y) + \delta f_1(y) \tag{A.6}$$

$$\widetilde{f}_1(y) = (1-\gamma)f_0(y) + \gamma f_1(y) \tag{A.7}$$

$$\widetilde{\pi}^1_{t+1|t} = -\frac{\delta}{(\gamma-\delta)} + \frac{1}{(\gamma-\delta)}\pi^1_{t+1|t} \tag{A.8}$$

for some constants δ, γ*.*

At this point we note that if $\delta = \gamma$ then $\widetilde{f}_0 = \widetilde{f}_1$ in which case the model is reduced to a single-regime model which is ruled out from the beginning. Accordingly, in what follows, we always assume $\delta \neq \gamma$.

Lemma A.2 *If an observationally equivalent model,* $\widetilde{M} = \{\widetilde{f}_0, \widetilde{f}_1, \widetilde{P}_{00}, \widetilde{P}_{11}\}$*, exists then* $\widetilde{P}_{01}$ *has the following two representations:*

$$\widetilde{P}_{01} = \frac{\gamma - P_{11}}{\delta - \gamma} \; if \; \delta \neq 0 \tag{A.9}$$

and

$$\widetilde{P}_{01} = \frac{P_{01} - \delta}{\gamma - \delta} \; if \; \delta \neq 1 \tag{A.10}$$

for the same δ, γ in the previous lemma.

Lemma A.3 *If an observationally equivalent model,* $\widetilde{M} = \{\widetilde{f}_0, \widetilde{f}_1, \widetilde{P}_{00}, \widetilde{P}_{11}\}$*, exists then* $\widetilde{P}_{11}$ *has the following two representations:*

$$\widetilde{P}_{11} = \frac{\widetilde{P}_{01}\gamma(P_{11} - P_{01} - 1) + \delta(P_{11} - P_{01} - 1) + P_{01}(1 + \widetilde{P}_{01})}{\delta(P_{11} - P_{01} - 1) + P_{01}} \tag{A.11}$$

and

$$\widetilde{P}_{11} = \frac{\delta(P_{01} - 1) + P_{11}(1 - \delta)}{\gamma(1 - \delta) + (\gamma - 1)\delta} \tag{A.12}$$

for the same δ, γ in the previous lemma.

The remainder of the proof proceeds as follows. First we show that the only possible values of δ are 0 and 1. Then we show that given δ, γ is either 1 or 0. These facts along with the formulae provided in lemmas A.2 and A.3 finish the proof.

Suppose that δ is neither 0 or 1. Lemma 2 then requires that the following equality be satisfied.

$$\frac{\gamma - P_{11}}{\delta - \gamma} = \frac{P_{01} - \delta}{\gamma - \delta}$$

Algebraic manipulation shows that this condition is tantamount:

$$P_{11}(\delta - \gamma) = P_{01}(\delta - \gamma)$$

Suppose that $\delta \neq \gamma$. In this case the above condition implies $P_{11} = P_{01}$ which is contrary to the maintained hypothesis. The only other way to satisfy the above condition is if $\delta = \gamma$. When $\delta = \gamma$ then lemma 1 shows that $\widetilde{f}_0 = \widetilde{f}_1$ in which case $\widetilde{f}(y_{t+1}|Y_t)$ is not affected by the history $\{y_{1,}y_2, ..., y_t\}$ but under the assumption that $P_{11} \neq P_{01}$ and $f_0 \neq f_1$, $f(y_{t+1}|Y_t)$ is affected by the history $\{y_1, y_2, ..., y_t\}$. Consequently it can not be the case that $f(y_{t+1}|Y_t) = \widetilde{f}(y_{t+1}|Y_t)$. Accordingly, the only permissible values of δ are 0 and 1.

Next we show that whenever $\delta = 1$ then $\gamma = 0$ and vice versa. Suppose that $\delta = 1$. In this case lemma 3 requires that:

$$\frac{\widetilde{P}_{01}\gamma(P_{11} - P_{01} - 1) + (P_{11} - P_{01} - 1) + P_{01}(1 + \widetilde{P}_{01})}{P_{11} - 1} = \frac{(P_{01} - 1)}{(\gamma - 1)}$$

Also recall that lemma 2 requires that:

$$\widetilde{P}_{01} = \frac{(1 - P_{11})}{(1 - \gamma)}$$

These two conditions imply that:

$$\gamma(P_{11} - P_{01}) = 0$$

This can only be satisfied if $\gamma = 0$ since $P_{11} \neq P_{01}$ by assumption.

Now suppose that $\delta = 0$. In this case lemma A.3 requires that

$$\frac{\widetilde{P}_{01}\delta(P_{11} - P_{01} - 1) + \delta(P_{11} - P_{01} - 1) + P_{01}(1 + \widetilde{P}_{01})}{\delta(P_{11} - P_{01} - 1) + P_{01}} = \frac{P_{11}}{\gamma}$$

and lemma 2 requires that:

$$\widetilde{P}_{01} = \frac{P_{01}}{\gamma}$$

These two conditions imply that:

$$\frac{(\gamma - 1)}{\gamma}(P_{11} - P_{01}) = 0$$

Since $P_{11} \neq P_{01}$ by assumption it must be the case that $\gamma = 1$.

Now we have shown that the only admissible (δ, γ) pairs are $(1, 0)$ and $(0, 1)$. Appealing to the equations in lemmas 2 and 3 readily shows that when $(\delta, \gamma) = (1, 0)$ then $\widetilde{P}_{11} = 1 - P_{01}$ and $\widetilde{P}_{01} = 1 - P_{11}$. Likewise, when $(\delta, \gamma) = (0, 1)$ then $\widetilde{P}_{11} = P_{11}$ and $\widetilde{P}_{01} = P_{01}$. This completes the proof. ■

Proofs of Lemma A.1, A.2, and A.3 are provided below.

Proof of Lemma A.1

Suppose an observationally equivalent model ($\widetilde{M}$) exists then it must be the case that

$$f(y_{t+1}|Y_t) = \widetilde{f}(y_{t+1}|Y_t), \forall y_{t+1}, Y_t$$

This condition is simply the definition of observational equivalence. If the conditional distributions differ over any part of the support or for any history Y_t then the two models are indeed discernible from each other. The conditional distribution in a regime switching model is given by:

$$f(y_t|Y_{t-1}) = \pi^1_{t|t-1} f_1(y_t) + (1 - \pi^1_{t|t-1}) f_0(y_t)$$

where $\pi^1_{t+1|t}$ is given by the following recursion:

$$\begin{aligned} \pi^1_{t|t} &= \frac{\pi^1_{t|t-1} f_1(y_t)}{\pi^1_{t|t-1} f_1(y_t) + \pi^0_{t|t-1} f_0(y_t)} \\ \pi^1_{t+1|t} &= P_{11}\pi^1_{t|t} + P_{01}\pi^0_{t|t} \end{aligned}$$

Accordingly, observational equivalence implies:

$$f_0(y_{t+1}) + \pi^1_{t+1|t}(f_1(y_{t+1}) - f_0(y_{t+1})) = \widetilde{f}_0(y_{t+1}) + \widetilde{\pi}^1_{t+1|t}(\widetilde{f}_1(y_{t+1}) - \widetilde{f}_0(y_{t+1}))$$

or

$$\pi^1_{t+1|t} = \frac{(\widetilde{f}_0(y_{t+1}) - f_0(y_{t+1}))}{(f_1(y_{t+1}) - f_0(y_{t+1}))} + \frac{(\widetilde{f}_1(y_{t+1}) - \widetilde{f}_1(y_{t+1}))}{(f_1(y_{t+1}) - f_0(y_{t+1}))}\widetilde{\pi}^1_{t+1|t}$$

Moreover, for any two distinct histories the pair $\left(\pi^1_{t+1|t}, \widetilde{\pi}^1_{t+1|t}\right), \left(\pi^{1,\prime}_{t+1|t}, \widetilde{\pi}^{1,\prime}_{t+1|t}\right)$ lie on a straight line for any y_{t+1}. Accordingly, it must be the case that

$$\frac{(\widetilde{f}_0(y_{t+1}) - f_0(y_{t+1}))}{(f_1(y_{t+1}) - f_0(y_{t+1}))} = \delta$$

and

$$\frac{(\widetilde{f}_1(y_{t+1}) - \widetilde{f}_1(y_{t+1}))}{(f_1(y_{t+1}) - f_0(y_{t+1}))} = \gamma - \delta$$

for some constants δ and γ. This proves equation (3) of Lemma 1. Manipulation of the two equations above yields:

$$\begin{aligned} \widetilde{f}_0(y_{t+1}) &= \delta f_1(y_{t+1}) + (1-\delta)_1 f_0(y_{t+1}) \\ \widetilde{f}_0(y_{t+1}) &= \gamma f_1(y_{t+1}) + (1-\gamma) f_0(y_{t+1}) \end{aligned}$$

This completes the proof of Lemma A.1

Proof of Lemma A.2

To economize on notation in what follows we write f instead of $f(\cdot)$.

Begin by assuming that the state probabilities are linearly related. Using the fact that $\pi_{t+1|t} = P\pi_{t|t} + Q(1 - \pi_{t|t})$ it must necessarily be the case that:

$$\widetilde{\pi}^1_{t|t} = \frac{(P_{01} - \delta + (\delta - \gamma)\widetilde{P}_{01})}{(\gamma - \delta)(\widetilde{P}_{11} - \widetilde{P}_{01})} + \frac{(P_{11} - P_{01})}{(\gamma - \delta)(\widetilde{P}_{11} - \widetilde{P}_{01})}\pi_{t|t}$$

Note that this expression is always valid since $P_{11} \neq P_{01}$ it must necessarily be the case that $\widetilde{P} \neq \widetilde{P}_{01}$ also $\delta = \gamma$ is also inconsistent with the true model since it is assumed that $f_0 \neq f_1$ and if $\delta = \gamma$ then $\widetilde{f}_0 = \widetilde{f}_1$. Now we use the fact that $\pi^1_{t|t} = \frac{\pi^1_{t|t-1} f_1(y_t)}{\pi^1_{t|t-1} f_1(y_t) + \pi^0_{t|t-1} f_0(y_t)}$ and rewrite the above relationship in terms of $\widetilde{\pi}^1_{t|t-1}$ and $\pi^1_{t|t-1}$. The resulting expression can be written as follows:

$$A\widetilde{\pi}^1_{t|t-1} + B\pi^1_{t|t-1} + C\widetilde{\pi}^1_{t|t-1}\pi^1_{t|t-1} + D = 0$$

where the coefficients are given by:

$$A = \frac{\widetilde{f}_1 f_0(\gamma - \delta)(\widetilde{P}_{11} - \widetilde{P}_{01}) + \kappa(\widetilde{f}_1 - \widetilde{f}_0 f_0}{(\gamma - \delta)(\widetilde{P}_{11} - \widetilde{P}_{01})}$$

$$B = \frac{\widetilde{f}_0\{(f_0 - f_1)\kappa - f_1(P_{11} - P_{01})\}}{(\gamma - \delta)(\widetilde{P}_{11} - \widetilde{P}_{01})}$$

$$C = \frac{\widetilde{f}_1(f_1 - f_0)(\gamma - \delta)(\widetilde{P}_{11} - \widetilde{P}_{01}) + (\widetilde{f}_0 - \widetilde{f}_1)\{\kappa(f_1 - f_0) + f_1(P_{11} - P_{01})\}}{(\gamma - \delta)(\widetilde{P}_{11} - \widetilde{P}_{01})}$$

$$D = \frac{-\kappa \widetilde{f}_0 f_0}{(\gamma - \delta)(\widetilde{P}_{11} - \widetilde{P}_{01})}$$

$$\kappa = P_{01} - \delta + (\delta - \gamma)\widetilde{P}_{01}$$

In order to maintain the relationship $\widetilde{\pi}^1_{t+1|t} - \frac{1}{(\gamma-\delta)}\pi^1_{t+1|t} + \frac{\delta}{(\gamma-\delta)} = 0$ it must be the case that:

$$\frac{B}{A} = -\frac{1}{(\gamma - \delta)}$$

$$C = 0$$

and

$$\frac{D}{A} = \frac{\delta}{(\gamma - \delta)}$$

an implication of these restrictions is that $\frac{D}{B} = -\delta$. We examine this implication in more detail below. Using the fact that $\widetilde{f}_0 = (1 - \delta)f_0 + \delta f_1$ and $\widetilde{f}_1 = (1 - \gamma)f_0 + \gamma f_1$ along with the constraint that $\frac{D}{B} = -\delta$ we find that:

$$\widetilde{P}_{01} = \frac{af_1 + bf_0}{cf_1 + df_0}$$

where:

$$a = \delta(\delta - P_{11})$$

$$b = P_{01}(\delta - 1) + \delta(1 - \delta)$$

$$c = \delta(\delta - \gamma)$$

$$d = \delta(1 - \delta) + \gamma(\delta - 1)$$

Now, to ensure that $\widetilde{P}_{01}$ is a constant its dependence on f_0 and f_1 must vanish. To eliminate this dependence it must be the case that

$$a = c\widetilde{P}_{01}$$

$$b = d\widetilde{P}_{01}$$

These relationships can be expressed as:

$$\delta(\delta - P_{11}) = \widetilde{P}_{01}\delta(\delta - \gamma)$$

and

$$P_{01}(\delta - 1) + \delta(1 - \delta) = \widetilde{P}_{01}\{\delta(1 - \delta) + \gamma(\delta - 1)\}$$

We rewrite the above expressions in terms of $\widetilde{P}_{01}$.

$$\widetilde{P}_{01} = \frac{(\delta - P_{11})}{(\delta - \gamma)} \; if \; \delta \neq 0$$

and

$$\widetilde{P}_{01} = \frac{P_{01}(\delta - 1) + \delta(1 - \delta)}{\delta(1 - \delta) + \gamma(\delta - 1)} \; if \; \delta \neq 1$$

This completes the proof of Lemma A.2.

Proof of Lemma A.3 Using equation (3) from Lemma A.1 we have:

$$E(\widetilde{\pi}^1_{t+1|t}) = -\frac{\delta}{(\gamma - \delta)} + \frac{1}{(\gamma - \delta)}E(\pi^1_{t+1|t})$$

It is easy to show that $E(\widetilde{\pi}^1_{t+1|t}) = \pi^{1*} = \frac{P_{01}}{1 - P_{11} + P_{01}}$. Accordingly, we have:

$$\frac{\widetilde{P}_{01}}{(1 - \widetilde{P}_{11} + \widetilde{P}_{01})} = -\frac{\delta}{(\gamma - \delta)} + \frac{1}{(\gamma - \delta)}\frac{P_{01}}{1 - P_{11} + P_{01}}$$

Algebraic manipulation yields

$$\widetilde{P}_{11} = \frac{\widetilde{P}_{01}\gamma(P_{11} - P_{01} - 1) + \delta(P_{11} - P_{01} - 1) + P_{01}(1 + \widetilde{P}_{01})}{\delta(P_{11} - P_{01} - 1) + P_{01}}$$

Next we provide another expression for $\widetilde{P}_{11}$. We claim that the following expression is valid.

$$\widetilde{P}_{11} = \frac{\delta(P_{01} - 1) + P_{11}(1 - \delta)}{\gamma(1 - \delta) + (\gamma - 1)\delta}$$

We prove the assertion in the following steps. First recall that earlier we showed that $A\widetilde{\pi}^1_{t|t-1} + B\pi^1_{t|t-1} + C\widetilde{\pi}^1_{t|t-1}\pi^1_{t|t-1} + D = 0$ and that in order to preserve linearity between $\widetilde{\pi}^1_{t|t-1}$ and $\pi^1_{t|t-1}$ we require that $C = 0$. Setting $C = 0$ and re-arranging terms yields the following:

$$\widetilde{P}_{11} = \frac{af_1 + bf_0}{cf_1 + df_0}$$

where

$$\begin{aligned} a &= P_{11} - \delta + \widetilde{P}_{01}\delta \\ b &= \widetilde{P}_{01}(1 - \delta) + \delta - P_{01} \\ c &= \gamma \\ d &= 1 - \gamma \end{aligned}$$

Now the argument proceeds in the same way as Lemma 2. In order to ensure that $\widetilde{P}_{11}$ is constant we need to restrict the coefficients as follows:

$$\begin{aligned} a &= \widetilde{P}_{11}c \\ b &= \widetilde{P}_{11}d \end{aligned}$$

equating coefficients and collecting terms results in two expressions for $\widetilde{P}_{01}$.

$$\begin{aligned} \widetilde{P}_{01} &= \frac{\widetilde{P}_{11}\gamma - P_{11} + \delta}{\delta} \\ \widetilde{P}_{01} &= \frac{\widetilde{P}_{11}(1-\gamma) - \delta + P_{01}}{(1-\delta)} \end{aligned}$$

Equating the two expressions results in the following expression for $\widetilde{P}_{11}$.

$$\widetilde{P}_{11} = \frac{\delta(P_{01} - 1) + P_{11}(1-\delta)}{\gamma(1-\delta) + (\gamma - 1)\delta}$$

This completes the proof of Lemma A.3

A.4 Proof of Algorithm 5.2

In this section we provide some details on the mechanics of the EM algorithm used in estimation. A note about notation. In this section of the appendix we replace the notation for the time-t information set Ω_t with Y_t since we abstract from any role for covariates.

A.4.1 The Parametric Case

First we consider the parametric case of section 4.2. Suppose that the support of y is discrete. Further label the two state dependent probability distributions $p_1 \equiv \{p_{k,0}\}_{k=1}^K$ and $p_1 \equiv \{p_{k,1}\}_{k=1}^K$. Conditional on a regime path $R = \{r_1, r_2, ..., r_T\}$, the likelihood takes the following form:

$$p(Y_T, R; \theta) = \rho_{r_0} p(r_1|r_0)...p(r_T|r_{T-1})(p_{1,0})^{n_{10}}...(p_{K,0})^{n_{K,0}}...(p_{1,1})^{n_{1,1}}...(p_{K,1})^{n_{K,1}}$$

where θ represents all model parameters, P_{00}, P_{11}, p_0, p_0, $n_{k,r}$ is the number of times y_k was observed during state r and $\sum_{k,r} n_{k,r} = T$. Instead of summing over all possible regime paths, R, to construct the likelihood, the EM algorithm is employed. Recall from section ?? that the EM algorithm proceeds by iteratively maximizing the following function:

$$\begin{aligned} Q(\theta^{l+1}; Y_T, \theta^l) = &\sum_R \ln(\rho_{r_0}) \cdot p(Y_T, R, \theta_l) \\ &+ \sum_R \sum_{t=1}^T \ln(p(r_t|r_{t-1})) \cdot p(Y_T, R, \theta^l) \\ + \sum_R \sum_{k=1}^K \sum_{t=1}^T &1\{y_t = y_k\} 1\{r_t = r\} \ln(p_{kr}) \cdot p(Y_T, R, \theta^l) \end{aligned}$$

$Q(\theta^{l+1}; Y_T, \theta^l)$ is maximized subject to the constraint that each state conditional distribution sums to unity ($\sum_{k=1}^{K} p_{kr} = 1;\ r = 0, 1$). Note that Q is separable in terms of the initial state probability (ρ_{r_0}), the transition probabilities ($p(r_t|r_{t-1})$) and the state conditional distributions (p_{kr}). Furthermore, the constraints do not involve any parameters except for those related to the state conditional distributions (p_{kr}). As a result, the first order conditions for ρ_{r_0} and $p(r_t|r_{t-1})$ can be solved independently of those for p_{kr}. Moreover, the first order conditions relating to the transition probabilities and the initial state probability are identical to those in Hamilton (1990). Accordingly, we do not reproduce the algebraic manipulations used to solve for these parameters. The interested reader may consult Hamilton (1990) for the full details. Setting the FONC relating to $P_{r,r}$ and ρ_{r_0} yields the following recursion:

$$P_{r,r}^{l+1} = \sum_{t=2}^{T} \Pr(r_{t-1} = 1, r_{t-1} = 1|Y_T, \theta^l) \cdot [\sum_{t=2}^{T} \Pr(r_{t-1} = 1|Y_T, \theta^l)]^{-1}$$

$$\rho_{r_0}^{l+1} = \Pr(r_0 = 1|Y_T, \theta^l)$$

Turning our attention to the parameters governing the state conditional distributions (p_{kr}) we find that the first order conditions take the following form:

$$\frac{\partial Q(\theta^{l+1}; Y_T, \theta)}{\partial p_{kr}} : \sum_{R} \sum_{t=1}^{T} \left\{ \frac{1}{p_{kr}} 1\{y_t = y_k\} 1\{r_t = r\} \right\} \cdot p(Y_T, R, \theta^l) = \mu_r$$

$$\sum_{r_t=0}^{1} \sum_{t=1}^{T} \left\{ \frac{1}{p_{kr}} 1\{y_t = y_k\} 1\{r_t = r\} \Pr(r_t = r|Y, \theta^l) \right\} = \frac{\mu_r}{p(Y_T, \theta^l)}$$

$$\sum_{t=1}^{T} \left\{ 1\{y_t = y_k\} \Pr(r_t = r|Y_T, \theta^l) \right\} = p_{kr} \frac{\mu_R}{p(Y_T, \theta^l)}$$

The constant, $\frac{\mu_r}{p(Y_T, \theta^l)}$, is pinned down by summing over the support of the distribution.

$$\sum_{k=1}^{K} \sum_{t=1}^{T} \left\{ 1\{y_t = y_k\} \Pr(r_t = r|Y_T, \theta^l) \right\} = \frac{\mu_r}{p(Y_T, \theta^l)} \sum_{k=1}^{K} p_{kr} = \frac{\mu_r}{p(Y_T, \theta^l)}$$

$$\frac{\mu_r}{p(Y_T, \theta^l)} = \sum_{t=1}^{T} \Pr(r_t = r|Y_T, \theta^l)$$

Now we find the solution for p_{kr},

$$p_{kr}^{l+1} = \frac{\sum_{t=1}^{T} \left\{ 1\{y_t = y_k\} \Pr(r_t = r|Y_T, \theta^l) \right\}}{\sum_{t=1}^{T} \Pr(r_t = r|Y_T, \theta^l)}$$

The algorithm is completed by providing expressions for $\Pr(r_t = r, r_{t-1} = r; Y_T, \theta^l)$ and $\Pr(r_t = r; Y_T, \theta^l) \equiv \widehat{\pi}_{t|T}^{r}$. The formulae describing these objects can be obtained from Appendix B of Hamilton (1990), or by making use of the approximation of Kim (1999).

A.4.2 The semiparametric case

We briefly discuss how the likelihood (and EM algorithm) are modified in the semiparametric case. Conditional on a regime path, R, we define the pseudo regime conditional likelihood as follows:

$$\widetilde{f}(Y_T, R; \theta) = \rho_{r_0} p(r_1|r_0)...p(r_T|r_{T-1}) \prod_{t=1}^{T} p_0(y_1)^{w_t^{0,1}} ... \prod_{t=1}^{T} p_0(y_K)^{w_t^{0,K}} \prod_{t=1}^{T} p_1(y_1)^{w_t^{1,1}} ... \prod_{t=1}^{T} p_1(y_K)^{w_t^{1,K}}$$

where now the weight function $w_t^{r,k} = \frac{1}{h} K(\frac{|y_t - y_k|}{h}) 1\{r_t = r\} \Delta y_k$. Our approach will be to take the exact same approach to maximizing $\widetilde{f}(Y_T, \theta)$ as we did to maximizing $p(Y_T, \theta)$, namely we will employ the EM algorithm.

Now we characterize the solution to the problem of maximizing $\widetilde{Q}(\theta^{l+1}; Y_t, \theta^l)$.

First note that $\ln(\widetilde{f}(Y_T, R; \theta))$ can be separated into three parts: $\ln(\widetilde{f}(Y_T, R; \theta)) = \ln(\rho_{r_0}) + \sum_{t=2}^{T} \ln(p(r_t|r_{t-1})) + \sum_{t=1}^{T} \sum_{r=0}^{1} \sum_{k=1}^{K} w_t^{r,k} \ln(f_r(y_k))$. The first part represents the contribution to the likelihood from the initial state probability, the second piece reflects the contribution from the transition probabilities and the third the contribution from the state conditional pseudo-likelihood for y. Since the last piece is the only section of the log pseudo likelihood which differs from the parametric case, determining the initial state vector and the transition probabilities do not differ from the parametric case. Namely,

$$P_{r,r}^{l+1} = \sum_{t=2}^{T} \Pr(r_t = 1, r_{t-1} = 1|Y_T, \theta^l) \cdot [\sum_{t=2}^{T} \Pr(R_{t-1} = 1|Y_T, \theta^l)]^{-1}$$

$$\rho_{r_0}^{l+1} = \Pr(r_0 = r|Y_T, \theta^l)$$

The only part of the pseudo likelihood that differs from the parametric case is the last term which contains the parameters $\{f_0(y_1), ..., f_0(y_K); f_1(y_1), ..., f_1(y_K)\}$. Now we take the derivative of $\widetilde{Q}(\theta^{l+1}; Y_T, \theta^l)$ with respect to $f_r(y_k)$ and recognize the constraints $\sum_{k=1}^{K} f_0(y_k) \Delta y_k = 1$ and $\sum_{k=1}^{K} f_1(y_k) \Delta y_k = 1$.

Recall,

$$Q(\theta^{l+1}; Y_T, \theta^l) = \sum_{R} \ln(\widetilde{f}(Y_T, R, \theta^{l+1})) \cdot \widetilde{f}(Y_T, R, \theta^l)$$

$$\frac{\partial Q}{\partial f_r(y_k)} = \sum_{R} \sum_{t=1}^{T} \left\{ \frac{1}{f_s(y_k)} \frac{1}{h} K(\frac{|y_t - y_k|}{h}) 1\{r_t = r\} \Delta y_k \right\} \cdot \widetilde{f}(Y_T, R, \theta^l) =$$

$$\sum_{r_t=0}^{1} \sum_{t=1}^{T} \left\{ \frac{1}{f_r(y_k)} \frac{1}{h} K(\frac{|y_t - y_k|}{h}) 1\{r_t = r\} \Delta y_k \Pr(r_t = r|Y_T, \theta) \right\} \cdot \widetilde{f}(Y_T, \theta^l) =$$

$$\sum_{t=1}^{T} \left\{ \frac{1}{f_s(y_k)} \frac{1}{h} K(\frac{|y_t - y_k|}{h}) \Delta y_k \Pr(r_t = r|Y_T, \theta) \right\} \cdot \widetilde{f}(Y_T, \theta^l)$$

Now recognize the constraint:

$$\sum_{t=1}^{T} \left\{ \frac{1}{f_s(y_k)} \frac{1}{h} K(\frac{|y_t - y_k|}{h}) \Delta y_k \Pr(r_t = r|Y, \theta) \right\} \cdot \widetilde{f}(Y, \theta^l) - \mu_r \Delta y_k = 0$$

Using the constraint $\sum_{k=2}^{K} f_r(y_k)\Delta y_k = 1$ we can pin down the constant:

$$\frac{\mu_r}{\widetilde{f}(Y_T, \theta^l)} = \sum_t \sum_k \frac{1}{h} K(\frac{|y_t - y_k|}{h}) \Delta y_k \Pr(r_t = r | Y_T, \theta)$$

and using the fact that $K(\cdot)$ is a density we can reduce $\frac{\mu_r}{\widetilde{f}(Y_T, \theta^l)}$ to

$$\sum_t \Pr(r_t = r | Y_T, \theta)$$

Now setting the first derivative of the lagrangian to zero :

$$\sum_{t=1}^{T} \left\{ \frac{1}{f_r(y_k)} \frac{1}{h} K(\frac{|y_t - y_k|}{h}) \Pr(r_t = r | Y_T, \theta) \right\} = \sum_t \Pr(r_t = r | Y_T, \theta)$$

$$\widehat{f_r(y_k)}^{l+1} = \frac{1}{h} \sum_{t=1}^{T} K(\frac{|y_t - y_k|}{h}) \frac{\Pr(r_t = r | Y_T, \theta^l)}{\sum_t \Pr(r_t = r | Y_T, \theta^l)}$$

Table 1: Gaussian Regime Switching Model

Estimates and Model Statistics

	Regime 0	**Regime 1**
Constant ($\alpha_{0,r}$)	0.1687	-0.0057
	(0.1398)	(0.0180)
Auto-regressive Parameter ($\alpha_{1,r}$)	-0.0190	0.0015
	(0.0149)	(0.0032)
Standard Deviation (σ_r)	0.6716	0.1496
	(0.0778)	(0.0344)
Transition Probability (P_{00}, P_{11})	0.9680	0.9905
	(0.0088)	(0.0024)
Log - Likelihood	111.1109	

Model Statistics

Long-Run Probability (π)	23%	77%
Regime Duration	31.25 weeks	105.26 weeks
Unconditional Mean (μ)	0.00	0.00
Unconditional Std. Dev. (σ)	0.36	0.36
Unconditional Skewness ($\sqrt{b_1}$)	0.00	0.00
Unconditional Kurtosis (b_2)	9.62	9.62

Table 1: The top panel presents parameter estimates from Gray (1996): $\Delta i_t = \alpha_{0,r} + \alpha_{1,r} i_{t-1} + \sigma_r \varepsilon_t, \Pr(r_t = j | r_{t-1} = j) = P_{jj}$ and the bottom panel presents implied model statistics. Standard errors appear in parentheses.

Table 2: Regime Skewness, Kurtosis and Normality Tests

	Regime 0	**Regime 1**
Skewness	-0.80	0.09
Kurtosis	8.50	3.70
JB_0	396.16	–
JB_1	–	21.31
	Normality Tests	
JB^0_{RS}	338.12	–
JB^1_{RS}	–	92.42
JB_{RS}	428.68	428.68

Table 2: The top panel displays estimates of within-regime skewness and kurtosis along with the informal Jarque-Bera tests. The bottom panel displays the formal normality tests. The null model is $\Delta i_t = \mu_r + \sigma_r \varepsilon_{r,t}$ $JB^{0,1}_{RS}$ are asymptotically distributed $\chi^2(2)$. The asymptotic 5% and 1% critical values of the test are 5.99 and 9.21 respectively. JB_{RS} is asymptotically distributed $\chi^2(4)$. The asymptotic 5% and 1% critical values for the test are 9.49 and 13.28 respectively.

Table 3: Histogram Regime Switching Estimates

	Regime 0			Regime 1		
Δi_t	**1**	**2**	**2-1**	**1**	**2**	**2-1**
$(-\infty, -0.8)$	0.074	0.126	-0.053	0.000	0.000	0.000
	(0.015)			(–)		
$[-0.8, -0.6)$	0.035	0.072	-0.037	0.000	0.000	0.000
	(0.011)			(–)		
$[-0.6, -0.4)$	0.108	0.092	0.016	0.001	0.003	0.002
	(0.018)			(0.002)		
$[-0.4, -0.2)$	0.110	0.1085	0.0011	0.069	0.0835	-0.015
	(0.017)			(0.008)		
$[-0.2, 0.0)$	0.143	0.117	0.026	0.4302	0.404	-0.0262
	(0.022)			(0.022)		
$[0.0, 0.2)$	0.190	0.116	0.074	0.412	0.415	-0.003
	(0.027)			(0.022)		
$[0.2, 0.4)$	0.156	0.105	0.052	0.079	0.090	-0.012
	(0.023)			(0.001)		
$[0.4, 0.6)$	0.090	0.087	0.003	0.005	0.004	0.001
	(0.017)			(0.003)		
$[0.6, 0.8)$	0.027	0.066	-0.039	0.000	0.000	0.000
	(0.009)			(–)		
$[0.8, \infty)$	0.068	0.111	-0.043	0.000	0.000	0.000
	(0.014)			(–)		
P_{00}, P_{11}	0.9789			0.9916		
	(0.009)			(0.003)		
Log-Likelihood	-1863.50					

Table 3: Column 1 reports parameter estimates. Standard errors are reported in parentheses. In the case that the parameter estimate is 0, no standard error is reported. Column 2 reports the probability that would be expected under normality. Column 2-1 shows the difference between columns 1 and 2.

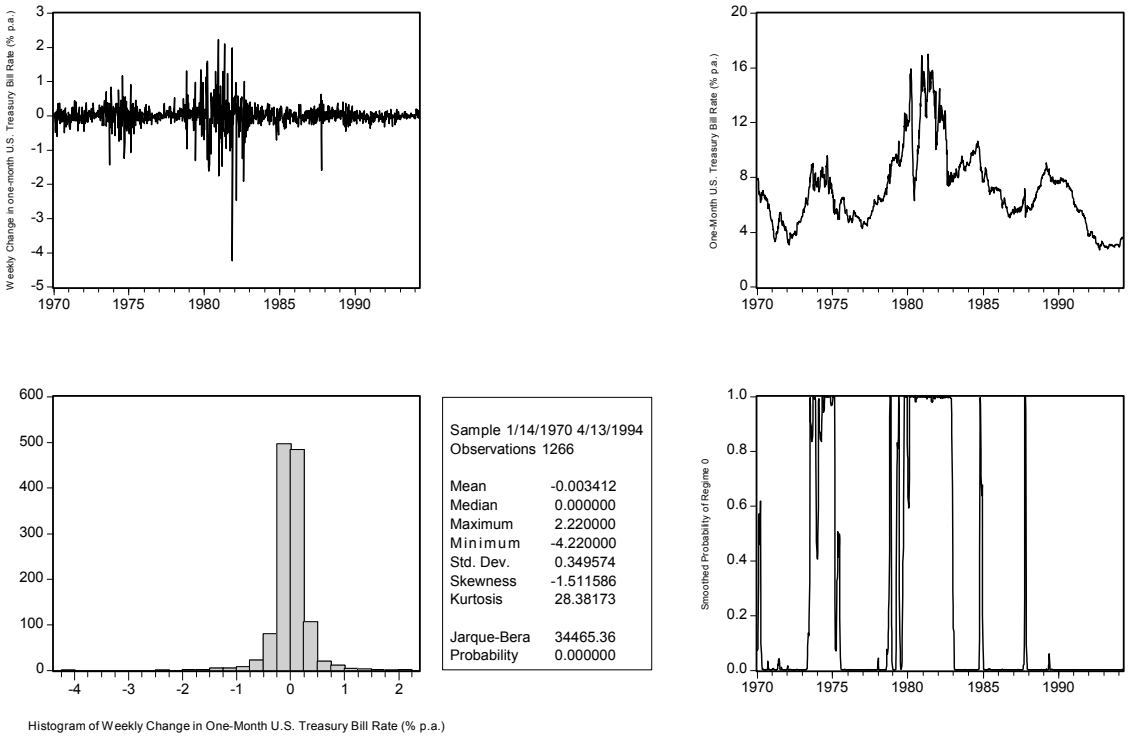

Figure 1: This figure summarizes the one-month U.S. Treasury bill rate between 1970-1994. The figure displays (left to right) a time-series plot of the weekly change in the Treasury bill rate, a time-series plot of the weekly level of the Treasury bill rate, a histogram of the weekly change, and lastly the smoothed probability of regime 0.

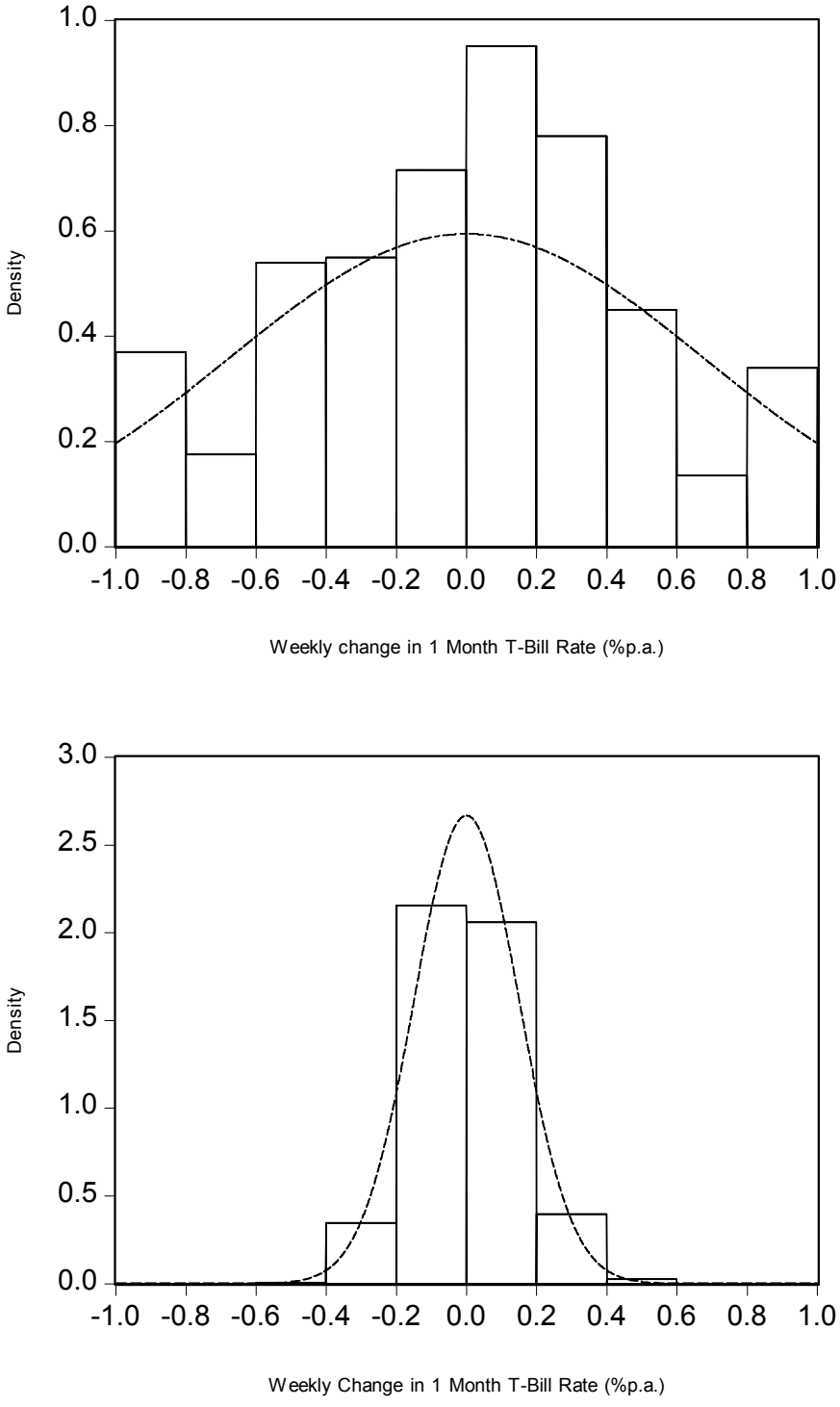

Figure 2: The graph above displays the estimated regime histograms (solid) for Δi_t along with the associated normal distribution (dashed) implied by Gray's (1996) estimates. The top panel displays Regime 0 and the bottom panel displays Regime 1. Note that the horizontal axes are identical but the vertical axes are not.

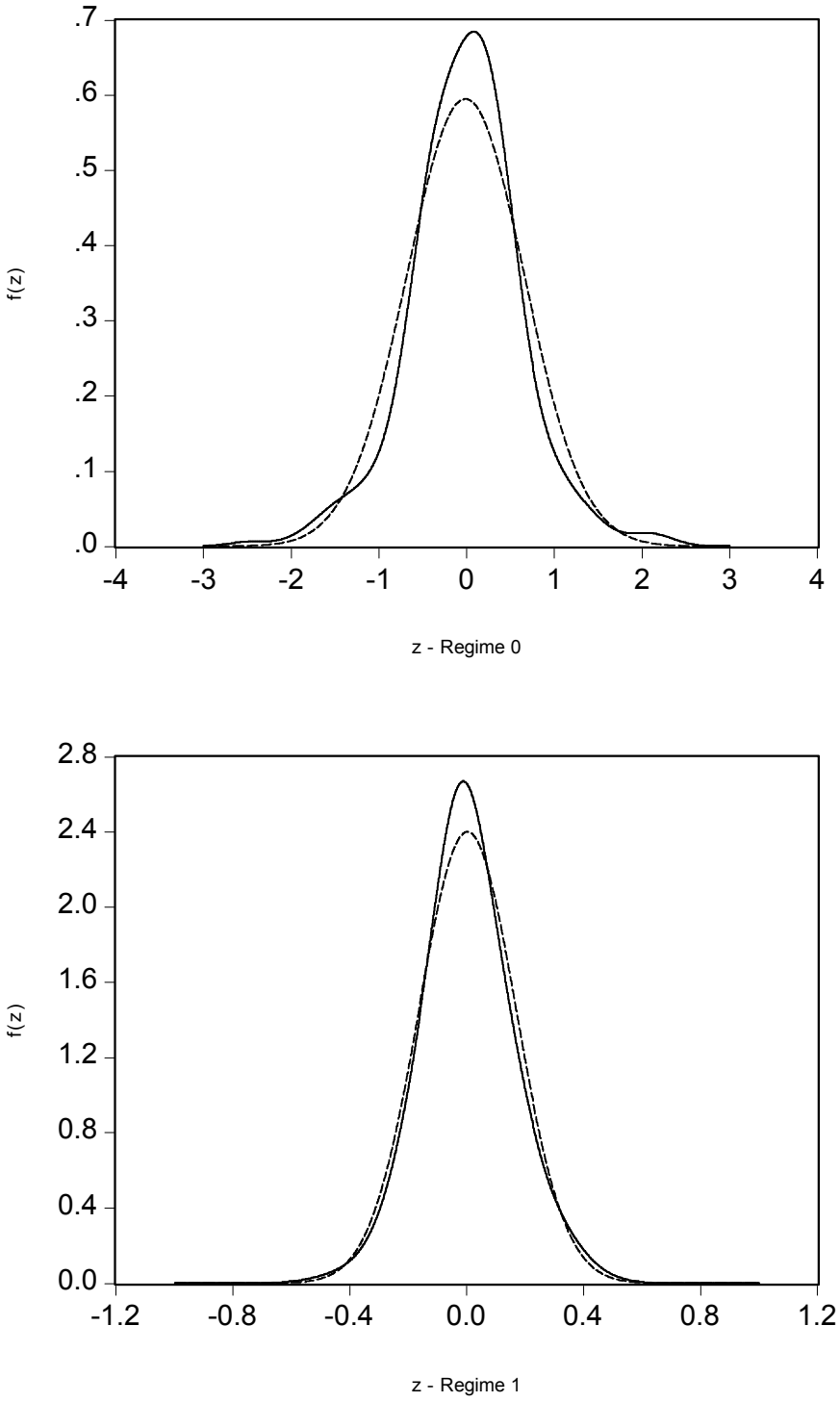

Figure 3: The graph above displays the estimated regime 0 (top panel) and regime 1 (bottom panel) distribution of weekly changes in the one-month U.S. Treasury bill rate (solid line) versus the Gaussian distribution implied by Gray's (1996) estimates (dotted line). Note that the axes are not identical.