Panel Regression with Endogenous Regime Switching

Yoosoon Chang*, Alireza Marahel†, Joon Y. Park‡

This Draft: May 1, 2022

Abstract

This paper investigates the time variation in the Captial Asset Pricing Model (CAPM) betas by introducing a new approach that models panel regressions with endogenous regime-switching using a latent autoregressive factor. For our estimation, we model the CAPM using portfolio returns sorted on book-to-market ratio, where the factor loadings, the pricing errors, and the volatility of the error terms can vary across high and low volatility states of the market. We find that the behavior of this asset pricing model significantly differs across different volatility regimes and its performance improves significantly, especially when it is evaluated during the times where the market is in the low volatility regime.

1 Introduction

The Capital Asset Pricing Model (CAPM) introduced by Sharpe (1964) and Lintner (1965) is among the first and most important benchmark models in the asset pricing literature where it considers the market return as the sole factor to explain the variations observed in the stock excess returns. However, numerous papers, including Fama and French (1992), have evaluated the performance of the CAPM with constant factor loadings and found that the estimated betas do not explain the variation observed in the average returns across different portfolios. After this point, many researchers have tried to propose an asset-pricing model that employs multiple factors to explain the excess stock returns, given in the form of $R^e_{i,t} = \alpha_i + \sum_{n=1}^{n=N} \beta_{n,i} f_{n,i} + \varepsilon_{i,t}$, where $R^e_{i,t}$ is the excess return and $f_{n,i}$'s are the risk factors. The number of different factors proposed is quite overwhelming. Harvey, Liu, and Zhu (2015) documented 316 significant factors pricing the cross-section of stock returns

*Department of Economics, Indiana University.
†Department of Economics, Indiana University: amarahel@iu.edu.
‡Department of Economics, Indiana University.

identified by the literature, with the majority being in the last 15 years. The main objective of introducing new factors is to construct a model that could explain the observed cross-sectional variations in the stock returns more than what is already established. However, the abnormal cross-sectional returns were still found to be persistent since the pricing errors seem to remain significant in all the models proposed so far. A possible explanation for the failure of the original CAPM and other asset-pricing models is the dramatic intertemporal variation in the stock prices, which we believe is the main reason that these traditional models cannot perform well. The main issue in these models is one of their principal assumptions which states the volatility level in the market and betas are both constant over time. This inspires the idea that there is more than a single regime existing in the market. Given this hypothesis, there is literature that follows this logic and studies the models with time-varying betas. Broadly speaking, there are two types of approaches to implement the time variation of factor loadings to the model specification. One way to do so is to consider continuous changes in the betas. For instance, Jagannathan and Wang (1996) evaluate the performance of conditional CAPM where it assumes that the CAPM holds in a conditional sense and the betas and the market risk premium can vary over time. Among this type of papers, some use instrumental variables to proxy time-variation observed in the factor loading and market risk premium and to identify the covariance between them (e.g. Lettau and Ludvigson (2002); Petkova and Zhang (2005)). In another group of articles, there has been an effort to apply the mentioned hypothesis to expand the multi-factor models with multiple regimes, each of which is associated with a different distribution of asset returns (e.g. Tu (2010); Abdymomunov and Morley (2011); Chen and Kawaguchi (2018)). To put it another way, the regime-switching models proposed by these articles create a setting that assumes that the stock excess returns are drawn from different distributions, with a well-defined stochastic process determining the likelihood that each return is drawn from a given distribution. However, they simply apply an exogenous Markov-switching model, which was introduced by Hamilton (1989), where the process of determining the regimes is completely independent of all other features of the model.

In another approach, to test the hypothesis that the time variation in the betas is discrete (having multiple regimes in the model), this paper proposes a new approach to model panel regressions with endogenous regime-switching using an autoregressive latent factor that was first introduced by Chang, Choi, and Park (2017). Under our specification, the state of the market–high volatility or low volatility state–is determined by whether a latent regime factor, that is extracted from the observed time series, takes a value above or below a threshold level. The innovation of the latent factor is assumed to be correlated with the previous stock return shock and as a result, the shock to the stock returns will affect the stock market

regime-switching in the following period. There are a couple of advantages of using this regime switching model. First, the ability of our model to extract the latent factor enables us to efficiently get more information on the regime-switching from the observed stock return data and look for the key determinant of the state of the market. Second, our model implies that the future state of the market is determined by not only the current state but also the realized values of the stock return, which is what one may normally expect.

Our empirical findings, consistent with the discussion made in section 4, demonstrates that the time-varying betas can help explaining the portfolio returns much better than the original CAPM, especially when market volatility level is relatively low. The results reported by previous articles, obtain from applying the regime-switching specification to an asset pricing model, commonly provided contrary evidence to the theoretical positive relationship between risk and return. To justify the anomalies observed to the contrary of this theory, these articles have discussed that even though the investors may react to the information about the true volatility regimes, it is more reasonable to assume that there is a time delay in the process of digesting the news and information about the current volatility level of the market. However, before we could make such a judgment, we think there should be a distinction between the aggregate uncertainty level in the market and the relative uncertainty observed in each portfolio with respect to the market which will be discussed in section 4. It is demonstrated that the positive relationship between risk and return still holds with respect to the relative risk observed in the portfolios, but only when the market is in the low volatility regime.

The model introduced in this paper can be applied to any multi-factor asset pricing model given in the form of $R^e_{i,t} = \alpha_i + \sum_{n=1}^{n=N} \beta_{n,i} f_{n,i} + \varepsilon_{i,t}$. To show how our model works, we consider the CAPM that measures the systematic risk of a security relative to the overall market. The overall market excess return, which is among the most promising factors expressed in the literature, is the only risk factor used in this model. We expect the model to correctly identify the price of risk when the market is in the low volatility regime. Our model specification may simply be extended by adding an additional latent factor to consider the possibility that the pricing errors can follow a separate state process. Additionally, this endogenous regime-switching specification can be further extended to a version that considers more than two volatility regimes to evaluate the performance of any asset pricing model.

2 Model

In this section, we introduce a panel regression model with endogenous regime switching and describe how it can simplify to a model based on the conventional regime switching.

The model for a panel (y_{it}) is specified as

$$y_{it} = \alpha_i(s_t) + \beta_i(s_t)'x_t + \varepsilon_{it}(s_t) \tag{1}$$

for $t = 1, ..., T$ and $i = 1, ..., N$, where (x_t) is a vector of covariates and $\varepsilon_{it}(s_t)$ represents the regime dependent error term, which is further specified as

$$\varepsilon_{it}(s_t) = \pi_i(s_t)u_t + \sigma_i e_{it}, \tag{2}$$

where in turn (u_t) and (e_{it}) are normal random variables with zero mean and unit variance, independent from each other, and also both serially and cross-sectionally. The specification in (2) implies that the error term in our panel regression a factor structure with a single common factor u_t whose loadings $\pi_i(s_t)$ are regime dependent. Only the common factor component has regime dependence, and all the idiosyncratic components are set to be independent of regimes. The coefficients $\alpha_i(s_t)$ and $\beta_i(s_t)$ in (1) are also set to be dependent upon regimes determined by the common state variable (s_t). In our model, the state variable (s_t) is defined by

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

where τ is an unknown threshold, $1\{\cdot\}$ is the indicator function, and (w_t) is a latent autoregressive factor generated as

$$w_t = \lambda w_{t-1} + v_t \tag{4}$$

with $|\lambda| < 1$ and i.i.d. standard normal innovations (v_t). Therefore, we have two regimes, denoted as 1 and 0 respectively according to the value of (s_t), depending upon whether the latent factor (w_t) takes a value above or below the threshold τ. Chang, Choi, and Park (2017) show that the transition given by (3) and (4) have a one-to-one correspondence with those of the general two-state Markov transitions: we can find a pair of λ and τ so that (3) and (4) yield the same transition for any two-state Markov transition, as well as the choice of a pair of λ and τ in (3) and (4) uniquely determine a two-state Markov transition.

The reformulation of a two-sate Markov transition as in (3) and (4) has some clear advantages. First, by introducing a latent factor, we may extract information on the *strength* of regimes, as well as regimes themselves. Secondly, and more importantly, our formulation makes it possible to allow for endogeneity in the regime switching. In fact, we introduce correlation between the common factor of the error term (u_t) and (v_t) in (2) and (4), and in particular let

$$\rho = \mathbb{E}(u_t v_{t+1}) \tag{5}$$

and allow $\rho \neq 0$. For nonzero ρ, the regime determined by the value of the latent factor w_{t+1} at time $t+1$ is affected by the realization of $\varepsilon_{it}(s_t)$ at time t (which itself depend on the realization of the regime at time t), implying the presence of a feedback effect of the common factor of the error term (u_t) on the regime. As in the conventional regime switching model, we assume that (u_t) and (v_t) are all jointly normal, and that they have zero mean and, for identification, unit variance. With this specification, if $\rho < 0$, the lagged common factor of the innovation u_t of the time series y_t at time t becomes negatively correlated with the innovation v_{t+1} of the latent autoregressive factor w_{t+1} at time $t+1$. This implies that a negative shock to y_t in the current period will cause an increase to the volatility level in the next period. The opposite is true when $\rho > 0$.

Under our specification in (1), (2) and (5), we can decompose (u_t) as

$$u_t = \rho v_{t+1} + \sqrt{1-\rho^2}\eta_t \tag{6}$$

with $|\rho| \leq 1$, and (η_t) is an i.i.d. standard normal random variable being independent of (v_t) at all leads and lags. With (1) and (6) taken together, our model may be rewritten as

$$y_{it} = \alpha_i(s_t) + \beta_i(s_t)'x_t + \rho\pi_i(s_t)v_{t+1} + \sqrt{1-\rho^2}\pi_i(s_t)\eta_t + \sigma_i e_{it}, \tag{7}$$

which is a general panel regression with regime switching, where we allow for both fixed and random effects, as well as heterogeneity. clearly, our model specified by (1), (2), (3), (4), (5) and (6) may also be given by (7) with (3) and (4).

To simplify our notation, let us label the states as 1 (low volatility) or 0 (high volatility) when state variable takes a value of 1 or 0, respectively. If we denote ξ_i as a generic notation for the state dependent parameters of the model, e.g. $\alpha_i(s_t)$, we may write

$$\xi_i(s_t) = \xi_{i,0}(1-s_t) + \xi_{i,1}s_t,$$

where $\xi_{i,0}$ and $\xi_{i,1}$ are the values the state dependent parameter can get, depending on whether we have $w_t < \tau$ or $w_t \geq \tau$. For the identification of the parameters of our model, we characterize the state of the market by its uncertainty, similar to what investors tend to do. We assume that $\pi_{i,0} > \pi_{i,1}$ for all i, which simply means that the level of uncertainty is relatively higher in the high volatility state than the low volatility state (note that $Var(\varepsilon_{it}(s_t)) = \pi_i^2(s_t) + \sigma_i^2$). If $\lambda = 1$, the latent factor (w_t) becomes a random walk and we further have to face the issue of joint identification for the initial value w_0 of (w_t) and the threshold level τ. In this case, the latent autoregressive process becomes $w_t = w_{t-1} + \sum_{t=1}^{T} v_t$ for all t. We set $w_0 = 0$ since any transformation of the form w_0 to $w_0 + c$ for any constant

c will result in the transformation of w_t to $w_t + c$ and τ to $\tau + c$, and will not affect state process (s_t) defined in (3). However, in the case of $|\lambda| < 1$, the identification problem of the initial value w_0 of (w_t) does not arise and if we let

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

the latent factor (w_t) becomes a strictly stationary process. Therefore, one may easily see that the autoregressive parameter λ determines the level of persistency observed in the regime changes. In particular, if the regime changes in the market is highly persistent in a specific time period, the autoregressive parameter will be close to 1 for that period.

If we let $\rho = 0$, the state process defined in (3) reduces to conventional Markov switching process where the innovation $\varepsilon_t(s_t)$ of the time series y_t becomes independent of the innovation v_{t+1} of the latent autoregressive factor w_{t+1}. To see how this works, we assume $\rho = 0$ for the rest of this section. It follows that the transition probabilities will depend on the latent factor autoregressive coefficient λ and the threshold level τ. We may easily see that

$$\mathbb{P}\{s_t = 0|w_{t-1}\} = \mathbb{P}\{w_t < \tau|w_{t-1}\} = \Phi\left(\tau - \lambda w_{t-1}\right) \tag{8}$$

$$\mathbb{P}\{s_t = 1|w_{t-1}\} = \mathbb{P}\{w_t \geq \tau|w_{t-1}\} = 1 - \Phi\left(\tau - \lambda w_{t-1}\right). \tag{9}$$

If we let $|\lambda| < 1$, it follows that the transition probabilities of the state process (s_t) from the low volatility state to the low volatility state and from the high volatility state to the high volatility state is given by

$$\mathbb{P}\{s_t = 0|s_{t-1} = 0\} = \frac{\displaystyle\int_{-\infty}^{\tau\sqrt{1-\lambda^2}} \Phi\left(\tau - \frac{\lambda x}{\sqrt{1-\lambda^2}}\right)\varphi(x)dx}{\Phi(\tau\sqrt{1-\lambda^2})} \tag{10}$$

$$\mathbb{P}\{s_t = 1|s_{t-1} = 1\} = 1 - \frac{\displaystyle\int_{\tau\sqrt{1-\lambda^2}}^{\infty} \Phi\left(\tau - \frac{\lambda x}{\sqrt{1-\lambda^2}}\right)\varphi(x)dx}{1 - \Phi(\tau\sqrt{1-\lambda^2})}. \tag{11}$$

Let us define the conditional transition density $p(s_t|s_{t-1})$ as

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

where $\omega = \omega(s_{t-1})$ is the transition probability of (s_t) to the low volatility state conditional

on the previous state and the past values of the observed times series and is given by

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

On the other hand, if we let $\lambda = 1$, the state process (s_t) defined in (3) becomes nonstationary and its transition evolves with time t. For $t = 1$, the transitions are given by $\mathbb{P}\{s_1 = 0|s_0 = 0\} = \Phi(\tau)$ where $\mathbb{P}\{s_{t-1} = 0\} = 1$ if $\tau > 0$, and $\mathbb{P}\{s_1 = 1|s_0 = 1\} = 1 - \Phi(\tau)$ where $\mathbb{P}\{s_{t-1} = 1\} = 1$ if $\tau \leq 0$. For $t \geq 2$, we define the transition probabilities explicitly as functions of time as

$$\mathbb{P}\{s_t = 0|s_{t-1} = 0\} = \frac{\int_{-\infty}^{\tau/\sqrt{t-1}} \Phi\left(\tau - x\sqrt{t-1}\right)\varphi(x)dx}{\Phi\left(\tau/\sqrt{t-1}\right)} \tag{13}$$

$$\mathbb{P}\{s_t = 1|s_{t-1} = 1\} = 1 - \frac{\int_{\tau/\sqrt{t-1}}^{\infty} \Phi\left(\tau - x\sqrt{t-1}\right)\varphi(x)dx}{1-\Phi\left(\tau/\sqrt{t-1}\right)}. \tag{14}$$

3 Estimation

Our model can be estimated by the maximum likelihood method. For the maximum likelihood estimation of our model based on the sample (y_{it}) for $i = 1, ..., N$ and $t = 1, ..., T$, we let $y_t = (y_{1t}, ..., y_{Nt})'$ and

$$\mathcal{F}_t = \sigma\Big((y_s)_{s=1}^t\Big)$$

which is the information given by $y_1, ..., y_t$ for $t = 1, ..., T$. The log-likelihood function is then given by

$$\ell(y_1, ..., y_T) = \log p(y_1) + \sum_{t=2}^{T} \log p(y_t|\mathcal{F}_{t-1})$$

where $p(\cdot)$ and $p(\cdot|\cdot)$ denote the density and conditional density functions, respectively. The objective is to maximize the log-likelihood function over a matrix of unknown parameters $\theta \in \Theta$, which includes, for example, the coefficients of the model such as α_i, the latent autoregressive factor coefficient λ, etc. Then, the maximum likelihood estimator $\hat{\theta}$ of θ is

given by

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(y_1, ..., y_T)$$

where θ consists of the set of state dependent coefficients $(\alpha_{i0}, \alpha_{i1})$ and (β_{i1}, β_{i1}), the volatility parameters (π_{i0}, π_{i1}) and σ_i, as well as the correlation coefficient ρ, the autoregressive coefficient of the latent factor λ, and the threshold level τ.

As in the conventional regime switching model, the log-likelihood function can be obtained in two stages: prediction and updating steps. In what follows, we let $\varepsilon_t(s_t)$ denote $(\varepsilon_{1t}(s_t), ..., \varepsilon_{Nt}(s_t))'$, analogous to the definition of (y_t).

Prediction The prediction step is defined as

$$p(y_t|\mathcal{F}_{t-1}) = \sum_{s_t} p(y_t|s_t, \mathcal{F}_{t-1}) p(s_t|\mathcal{F}_{t-1}).$$

We may easily deduce that $p(y_t|s_t, \mathcal{F}_{t-1}) = p(y_t|s_t)$, and that

$$p(y_t|s_t) = \left(\frac{1}{2\pi}\right)^{N/2} \left(\sqrt{\det \Omega(s_t)}\right)^{-1} \exp\left(-\frac{1}{2}\varepsilon_t'(s_t)\Omega^{-1}(s_t)\varepsilon_t(s_t)\right) \tag{15}$$

with $\varepsilon_{it}(s_t)$ specified by

$$\varepsilon_{it}(s_t) = y_{it} - \alpha_i(s_t) - \beta_i'(s_t)x_t$$

for $i = 1, .., N$, where

$$\Omega(s_t) = \begin{pmatrix} \pi_1^2(s_t) + \sigma_1^2 & \pi_1(s_t)\pi_2(s_t) & \cdots & \pi_1(s_t)\pi_N(s_t) \\ \pi_2(s_t)\pi_1(s_t) & \pi_2^2(s_t) + \sigma_2^2 & \cdots & \pi_2(s_t)\pi_N(s_t) \\ \vdots & \vdots & \ddots & \vdots \\ \pi_N(s_t)\pi_1(s_t) & \pi_N(s_t)\pi_2(s_t) & \cdots & \pi_N^2(s_t) + \sigma_N^2 \end{pmatrix} \tag{16}$$

is the covariance matrix of $\varepsilon_t(s_t)$. Note that $p(\varepsilon_t(s_t)) =_d \mathbb{N}(0, \Omega(s_t))$.

Moreover, we have

$$p(s_t|\mathcal{F}_{t-1}) = \sum_{s_{t-1}} p(s_t|s_{t-1}, \mathcal{F}_{t-1}) p(s_{t-1}|\mathcal{F}_{t-1}) \tag{17}$$

As in the conventional Markov switching filter, $p(s_{t-1}|\mathcal{F}_{t-1})$ is obtained from the previous updating step, which will be given below. Therefore, it suffices to get $p(s_t|s_{t-1}, \mathcal{F}_{t-1})$. Note

that

$$p(w_t|w_{t-1},\mathcal{F}_{t-1}) = p(w_t|w_{t-1},y_{t-1}) = p(w_t|w_{t-1},\varepsilon_{t-1}(s_{t-1}))$$

where w_{t-1} is independent of $\varepsilon_{t-1}(s_{t-1})$.

Let $|\rho|<1$ and $|\lambda|<1$. It follows that

$$p(v_{t+1}|s_t,\varepsilon_t(s_t)) =_d \mathbb{N}\left(\rho\pi'(s_t)\Omega^{-1}(s_t)\varepsilon_t(s_t), 1-\rho^2\pi'(s_t)\Omega^{-1}(s_t)\pi(s_t)\right). \tag{18}$$

where $\pi=(\pi_1,...,\pi_N)'$. Inverse of Covariance Matrix provides some useful insights to find the determinant and inverse of the covariance matrix $\Omega(s_t)$, analytically. For the sake of simplicity of our notation, we drop (s_t) from the state dependent parameters and variables.

It follows that

$$\mathbb{P}\Big\{s_t=0|s_{t-1}=0,\mathcal{F}_{t-1}\Big\} = \frac{\displaystyle\int_{-\infty}^{\tau\sqrt{1-\lambda^2}} \Phi\left(\frac{\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}} - \frac{\lambda x}{\sqrt{(1-\lambda^2)\left(1-\rho^2\pi'\Omega^{-1}\pi\right)}}\right)\varphi(x)dx}{\Phi\left(\tau\sqrt{1-\lambda^2}\right)}$$

and

$$\mathbb{P}\Big\{s_t=1|s_{t-1}=1,\mathcal{F}_{t-1}\Big\} = 1-\frac{\displaystyle\int_{\tau\sqrt{1-\lambda^2}}^{\infty} \Phi\left(\frac{\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}} - \frac{\lambda x}{\sqrt{(1-\lambda^2)\left(1-\rho^2\pi'\Omega^{-1}\pi\right)}}\right)\varphi(x)dx}{1-\Phi\left(\tau\sqrt{1-\lambda^2}\right)}$$

with $\varepsilon_{i,t-1}$ given by

$$\varepsilon_{i,t-1} = y_{i,t-1} - \alpha_i - \beta_i' x_{t-1}$$

for $i=1,..,N$. Similar to Chang, Choi, and Park (2017), let us define the conditional transition density $p(s_t|s_{t-1},\mathcal{F}_{t-1})$ as

$$p(s_t|s_{t-1},\mathcal{F}_{t-1}) = (1-s_t)\omega_\rho + s_t(1-\omega_\rho) \tag{19}$$

where $\omega_\rho = \omega_\rho(s_{t-1},\mathcal{F}_{t-1})$ is the transition probability of (s_t) to the low volatility state conditional on the previous state and the past values of the observed times series. We can

easily see that

$$\omega_\rho = \frac{\left[(1-s_{t-1})\int_{-\infty}^{\tau\sqrt{1-\lambda^2}} + s_{t-1}\int_{\tau\sqrt{1-\lambda^2}}^{\infty}\right]\Phi\left(\frac{\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}} - \frac{\lambda x}{\sqrt{(1-\lambda^2)(1-\rho^2\pi'\Omega^{-1}\pi)}}\right)\varphi(x)dx}{(1-s_{t-1})\Phi\left(\tau\sqrt{1-\lambda^2}\right) + s_{t-1}\left[1-\Phi\left(\tau\sqrt{1-\lambda^2}\right)\right]} \tag{20}$$

Now, if we let $\lambda = 1$, the progression of latent autoregressive factor process defined in (4) becomes a random walk, which makes the state process defined in (3) nonstationary and its transition evolves with time t. For $t = 1$, $\omega_\rho(s_0) = \Phi(\tau)$ with $\mathbb{P}\{s_0 = 0\} = 1$ and $\mathbb{P}\{s_0 = 1\} = 1$ when $\tau > 0$ and $\tau \leq 0$, respectively. For $t \geq 2$

$$\omega_\rho = \frac{\left[(1-s_{t-1})\int_{-\infty}^{\tau/\sqrt{t-1}} + s_{t-1}\int_{\tau/\sqrt{t-1}}^{\infty}\right]\Phi\left(\frac{\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1}-x\sqrt{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)\varphi(x)dx}{(1-s_{t-1})\Phi\left(\tau/\sqrt{t-1}\right) + s_{t-1}\left[1-\Phi\left(\tau/\sqrt{t-1}\right)\right]}. \tag{21}$$

The conditional transition of y_t is fully specified by (20) or (21) in case of $|\rho| < 1$. Mathematical Proofs provides the steps required to obtain the above expressions.

If $|\rho| = 1$, we will have perfect endogeneity and the conditional transition of the state process (s_t) in (20) and (21) is not valid anymore. Under this condition, the current shock ε_t of the model will fully describes the realization of latent factor w_{t+1} in the next period. We modify the transition probability ω_ρ of (s_t) to the low volatility state, conditional on the previous state and past values of the observed time series for different values of the autoregressive coefficient λ, as follows:

(i) $\lambda = 0$

$$\omega_\rho = 1\{\rho\pi'\Omega^{-1}\varepsilon_{t-1} < \tau\}$$

(ii) $0 < \lambda < 1$

$$\omega_\rho = (1-s_{t-1})\min\left(1, \frac{\Phi\left(\frac{\sqrt{1-\lambda^2}}{\lambda}(\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)}{\Phi\left(\tau\sqrt{1-\lambda^2}\right)}\right)$$
$$+ s_{t-1}\max\left(0, \frac{\Phi\left(\frac{\sqrt{1-\lambda^2}}{\lambda}(\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1})\right) - \Phi\left(\tau\sqrt{1-\lambda^2}\right)}{1-\Phi\left(\tau\sqrt{1-\lambda^2}\right)}\right)$$

(iii) $-1 < \lambda < 0$

$$\omega_\rho = (1-s_{t-1}) \max \left(0, \frac{\Phi\left(\tau\sqrt{1-\lambda^2}\right) - \Phi\left(\frac{\sqrt{1-\lambda^2}}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)}{\Phi\left(\tau\sqrt{1-\lambda^2}\right)} \right)$$
$$+s_{t-1} \min \left(1, \frac{1 - \Phi\left(\frac{\sqrt{1-\lambda^2}}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)}{1 - \Phi\left(\tau\sqrt{1-\lambda^2}\right)} \right)$$

(iv) $\lambda = 1$

$$\omega_\rho = (1-s_{t-1}) \min \left(1, \frac{\Phi\left(\frac{1}{\sqrt{t-1}}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)}{\Phi\left(\tau/\sqrt{t-1}\right)} \right)$$
$$+s_{t-1} \max \left(0, \frac{\Phi\left(\frac{1}{\sqrt{t-1}}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right) - \Phi\left(\tau/\sqrt{t-1}\right)}{1 - \Phi\left(\tau/\sqrt{t-1}\right)} \right)$$

Clearly, the transition density of the state process (s_t) depends on the lagged values of the observed times series and consequently, is not a Markov process. However, if we let $\rho = 0$, the model simplifies to the standard 1st order Markov process, independent of (y_t) similar to the conventional Markov switching model.

Updating The updating step is exactly the same as that of the conventional Markov switching model, and given by

$$p(s_t|\mathcal{F}_t) = p(s_t|y_t, \mathcal{F}_{t-1}) = \frac{p(y_t|s_t, \mathcal{F}_{t-1})p(s_t|\mathcal{F}_{t-1})}{p(y_t|\mathcal{F}_{t-1})} \tag{22}$$

where $p(y_t|s_t, \mathcal{F}_{t-1})$ is given by (15). Therefore, we can easily obtain $p(s_t|\mathcal{F}_t)$ from $p(s_t|\mathcal{F}_{t-1})$ and $p(y_t|\mathcal{F}_{t-1})$ computed from the prediction step.

Latent Factor As mentioned before, the way we defined our regime switching filter for the state process (s_t) enables us to easily extract the latent autoregressive factor (w_t) through the prediction and updating steps defined in (17) and (22). In the prediction step for the latent factor, we may write

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

where $p(s_{t-1}|\mathcal{F}_{t-1})$ is obtained from the previous updating step for the state process (s_t).

In order to compute $p(w_t, s_{t-1}|\mathcal{F}_{t-1})$, we need to find the transition density of the latent factor conditional on the previous state and the information based the lagged values of the observed time series. The expression for this transition density is derived for different values the autoregressive coefficient λ of the latent factor process and the endogeneity parameter ρ, as follows:

(i) $|\lambda| < 1$ and $|\rho| < 1$

$$p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) = \frac{1-\Phi\left(\sqrt{\frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\rho^2\pi'\Omega^{-1}\pi}}\left(\tau-\frac{\lambda(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}\right)\right)}{1-\Phi\left(\tau\sqrt{1-\lambda^2}\right)} \times \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, \frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\lambda^2}\right)$$

$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\Phi\left(\sqrt{\frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\rho^2\pi'\Omega^{-1}\pi}}\left(\tau-\frac{\lambda(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}\right)\right)}{\Phi\left(\tau\sqrt{1-\lambda^2}\right)} \times \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, \frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\lambda^2}\right)$$

(ii) $|\lambda| < 1$ and $|\rho| = 1$

- $0 < \lambda < 1$

$$p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) = \frac{\frac{\sqrt{1-\lambda^2}}{\lambda}\varphi\left(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\lambda}\sqrt{1-\lambda^2}\right)}{1-\Phi\left(\tau\sqrt{1-\lambda^2}\right)} 1\{w_t \geq \lambda\tau+\rho\pi'\Omega^{-1}\varepsilon_{t-1}\}$$

$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\frac{\sqrt{1-\lambda^2}}{\lambda}\varphi\left(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\lambda}\sqrt{1-\lambda^2}\right)}{\Phi\left(\tau\sqrt{1-\lambda^2}\right)} 1\{w_t < \lambda\tau+\rho\pi'\Omega^{-1}\varepsilon_{t-1}\}$$

- $-1 < \lambda < 0$

$$p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) = \frac{\frac{\sqrt{1-\lambda^2}}{\lambda}\varphi\left(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\lambda}\sqrt{1-\lambda^2}\right)}{1-\Phi\left(\tau\sqrt{1-\lambda^2}\right)} 1\{w_t \leq \lambda\tau+\rho\pi'\Omega^{-1}\varepsilon_{t-1}\}$$

$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\frac{\sqrt{1-\lambda^2}}{\lambda}\varphi\left(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\lambda}\sqrt{1-\lambda^2}\right)}{\Phi\left(\tau\sqrt{1-\lambda^2}\right)} 1\{w_t > \lambda\tau+\rho\pi'\Omega^{-1}\varepsilon_{t-1}\}$$

(iii) $\lambda = 1$ and $|\rho| < 1$

$$p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) = \frac{1-\Phi\left(\sqrt{\frac{t-\rho^2\pi'\Omega^{-1}\pi}{(t-1)(1-\rho^2\pi'\Omega^{-1}\pi)}}\left(\tau - \frac{(t-1)(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{t-\rho^2\pi'\Omega^{-1}\pi}\right)\right)}{1-\Phi\left(\tau/\sqrt{t-1}\right)} \times \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, t-\rho^2\pi'\Omega^{-1}\pi\right)$$

$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\Phi\left(\sqrt{\frac{t-\rho^2\pi'\Omega^{-1}\pi}{(t-1)(1-\rho^2\pi'\Omega^{-1}\pi)}}\left(\tau - \frac{(t-1)(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{t-\rho^2\pi'\Omega^{-1}\pi}\right)\right)}{\Phi\left(\tau/\sqrt{t-1}\right)} \times \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, t-\rho^2\pi'\Omega^{-1}\pi\right)$$

(iv) $\lambda = 1$ and $|\rho| = 1$

$$p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) = \frac{\frac{1}{\sqrt{t-1}}\varphi\left(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{t-1}}\right)}{1-\Phi\left(\tau/\sqrt{t-1}\right)} 1\{w_t \geq \tau + \rho\pi'\Omega^{-1}\varepsilon_{t-1}\}$$

$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\frac{1}{\sqrt{t-1}}\varphi\left(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{t-1}}\right)}{\Phi\left(\tau/\sqrt{t-1}\right)} 1\{w_t < \tau + \rho\pi'\Omega^{-1}\varepsilon_{t-1}\}.$$

Similar to the state process (s_t), the updating step for the latent autoregressive factor is given by

$$p(w_t, s_{t-1}|\mathcal{F}_t) = \frac{p(y_t|w_t, s_{t-1}, \mathcal{F}_{t-1})p(w_t, s_{t-1}|\mathcal{F}_{t-1})}{p(y_t|\mathcal{F}_{t-1})}. \tag{24}$$

It follows that

$$p(w_t|\mathcal{F}_t) = \sum_{s_{t-1}} p(w_t, s_{t-1}|\mathcal{F}_t),$$

which enables us to extract the inferred factor,

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

for $t = 1, ..., T$, when the parameters that maximize the likelihood function are found.

4 Results

In this section, we delve into the evaluation of the Capital Asset Pricing Model (CAPM), a widely accepted asset pricing model that has been the subject of extensive studies over the years. Developed by Sharpe (1964) and Lintner (1965), CAPM is a benchmark model in the academic literature that provides insights into how investors make investment decisions based on market risk and expected return. Our goal is to improve the performance of the CAPM by relaxing one of its main underlying assumptions that states the model coefficients are constant. More specifically, we allow the pricing errors α_i's and the market excess return loading β_i's to vary across market conditions under the setup described in section 2. By allowing the β to vary across market conditions, we recognize that the risk associated with a particular stock may not be constant over time, but instead may vary depending on the prevailing market conditions. Under the traditional CAPM, β is assumed to be a constant that does not change over time, but in reality, the β of a stock can change depending on factors such as economic conditions, changes in industry dynamics, shifts in investor sentiment, etc. By allowing the β to have discrete shifts across market conditions, we can better capture these changes in risk and adjust the expected return accordingly. For instance, one may argue that during times of economic expansion, the β of a cyclical stock might increase as the company becomes more exposed to the ups and downs of the economy. By adjusting the β to reflect these changing market conditions, one can more accurately estimate the expected return of the stock and make better investment decisions.

Our study focuses on evaluating the performance of CAPM under the model specification outlined in section 2. In doing so, the market is assumed to be in one of two regimes, each with a different level of volatility. When the market is in a high volatility regime, the relationship between the asset returns and the market portfolio is expected to be different than when the market is in a low volatility regime. Under our specification, the timing of changes in β, which corresponds to changes in the market volatility levels, is determined directly by the return data through our endogenous regime switching specification. This is in contrast to the traditional approach of imposing changes in market volatility levels through an exogenous Markov-switching process. To identify different volatility states of the market, we define a latent variable ω_t as defined in section 2. When the latent factor exceeds a certain threshold τ, the market is assumed to be in a high volatility regime ($s_t = 1$), and when it is below that value, the market is assumed to be in a low volatility regime ($s_t = 0$). The use of the endogenous regime switching specification allows us to examine the behavior of CAPM in a more realistic setting, where changes in market volatility levels are not predetermined but rather emerge from the data itself. By doing so, we can better understand how the model

performs in real-world scenarios and its ability to capture the dynamic nature of financial markets. Our analysis aims to contribute to the ongoing debate on the effectiveness of CAPM as an asset pricing model and provide insights into its limitations and potential areas for improvement.

To estimate the regime-dependent CAPM, we consider the monthly data for excess stock returns on value-weighted tertile and decile portfolios of all stocks listed on the New York Stock Exchange (NYSE), AMEX and NASDAQ, sorted separately by BE/ME ratios (BE/ME portfolios). The dataset covers the period from July 1963 to June 2022, which corresponds to a total of 708 months. At the end of each June, stocks are allocated to three and ten BE/ME groups (Low to High) using NYSE breakpoints. In the sort for June of year t, B is book equity at the end of the fiscal year ending in year $t-1$ and M is the market cap at the end of December of year $t-1$, adjusted for changes in shares outstanding between the measurement of B and the end of December. This data is readily available and are downloaded from Kenneth R. French's official website.

Table I reports the estimation results for the model using ten portfolios sorted on BE/ME, respectively. The parameters estimated for the low volatility level of the market are significantly different from the parameters estimated for the high volatility model. To be more specific, the estimated β's for the portfolios with highest B/M (i.e. value portfolios) are significantly different across different states of the market. Contrary to the theoretical models that suggest a positive relation between risk and return, we find that β of the value portfolio in the low volatility regime is higher than the β in the high volatility regime. However, what we have found is consistent with some of the previous studies. For instance, Lakonishok et al. (1994) report that the β's for portfolios with higher book-to-market ratios are higher than β's for portfolios with lower book-to-market ratios in good times (low volatility state). For the portfolios with the lowest B/M, however, we observe that our results are consistent with the positive relationship between risk and return. To further discuss the risk-return relationship, we need to consider the distinction between the aggregate uncertainty level in the market and the relative uncertainty observed in each portfolio (or each stock specifically) with respect to the market uncertainly level.

Given the specification of the error term $\varepsilon_{it}(s_t)$ in (2), the overall risk of each portfolio is measured by the estimated values of $\pi_i(s_t)$ and σ_i. It is clear that the overall risk of each portfolio during the high volatility regime is higher than its corresponding value in the low volatility regime. However, by comparing the overall risk of different portfolios during each state separately, we observe that returns behave differently across different state. In relative terms and within the same regime, the portfolios that have higher book-to-market ratios carry a higher level of risk relative to the market and consequently, they should have betas

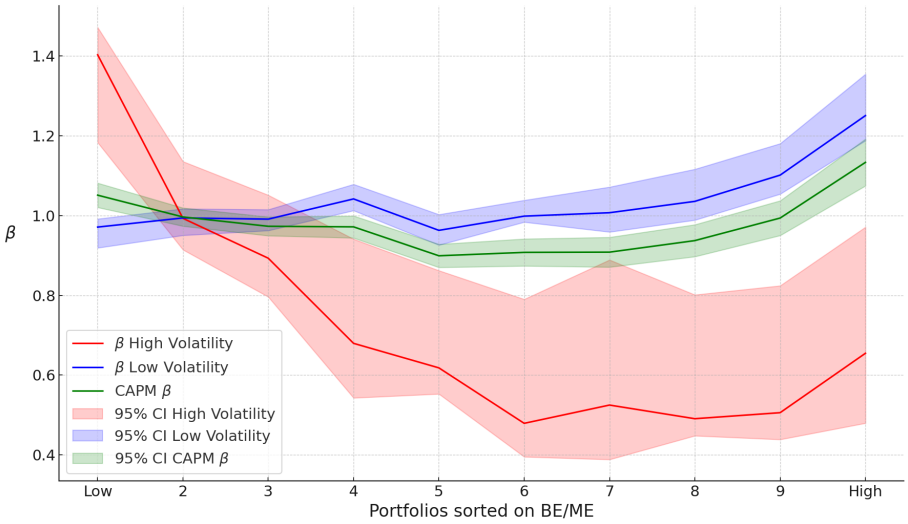
Table I: Regime-Dependent CAPM (10 Portfolios Sorted on BE/ME, 1963 - 2022)

		Growth	2	3	4	5	6	7	8	9	Value
Low Volatility	α	0.256***	0.3131***	0.349***	0.3604***	0.4262***	0.5402***	0.461***	0.5043***	0.6707***	0.5451***
		(0.8739)	(0.8102)	(0.7647)	(0.7159)	(0.6438)	(0.5196)	(0.6028)	(0.5672)	(0.4139)	(0.5646)
	β	0.9718***	0.9922***	0.9922***	1.0455***	0.9684***	1.0063***	0.9946***	1.0336***	1.0968***	1.2474***
		(0.1586)	(0.1347)	(0.1296)	(0.0877)	(0.1462)	(0.11)	(0.1176)	(0.0943)	(0.0824)	(0.1974)
	π	0***	0***	0***	0.013***	0.2956***	0.7313***	1.1836***	1.6909***	1.9429***	2.5532***
		(1.0998)	(1.0998)	(1.0998)	(1.0709)	(0.8636)	(0.5078)	(0.1732)	(0.5491)	(0.8241)	(1.4943)
High Volatility	α	0.3714	1.1813**	0.8452*	0.6872	0.59	0.7681	0.4425	1.19*	0.9583*	1.2918*
		(1.0013)	(0.4111)	(0.4585)	(0.6922)	(0.81)	(0.793)	(0.9638)	(0.5483)	(0.5971)	(0.7209)
	β	1.3648***	1.0022***	0.8987***	0.6843***	0.6324***	0.5269***	0.5815***	0.5486***	0.5871***	0.6761***
		(0.2758)	(0.1276)	(0.1948)	(0.3724)	(0.4258)	(0.5214)	(0.4822)	(0.5321)	(0.4752)	(0.4134)
	π	0***	0.7565***	2.0395***	2.3022***	2.3338***	3.2171***	3.3854***	1.8696***	2.0142***	2.5532***
		(1.0998)	(0.7676)	(1.1837)	(1.8611)	(2.044)	(3.2091)	(3.2328)	(1.7085)	(1.3986)	(1.5384)
	σ	1.7021***	1.3273***	1.2584***	1.3577***	1.4489***	1.3971***	1.4566***	1.4758***	1.5132***	2.2371***
		(0.6455)	(0.254)	(0.1756)	(0.2524)	(0.3148)	(0.2415)	(0.3177)	(0.3827)	(0.3879)	(1.0845)
	ρ					0.2495					
						(1.0597)					
	λ					0.6137***					
						(0.5566)					
	τ					1.3666***					
						(0.5553)					

Notes: The standard errors are calculated using wild residual bootstrap method and reported in parenthesis. Significance is computed using wild residual bootstrap confidence intervals. *p<0.1; **p<0.05; ***p<0.01

bigger than 1.

Figure I: β Estimates for High, Low Volatility Regimes, and CAPM with 95% Confidence Bands



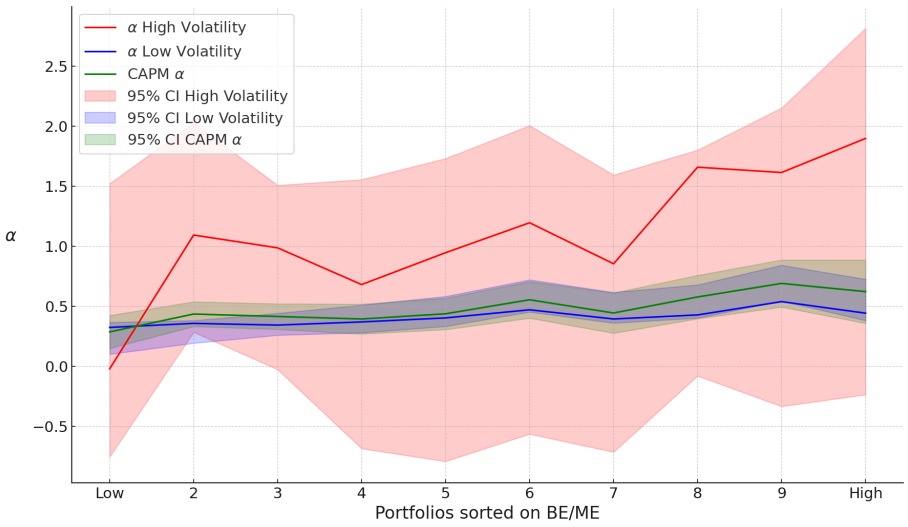
On the other hand, the portfolios with lower book-to-market ratios are expected to have betas less than 1 as they bear lower levels of risk compared to the market. As a general rule, for a model to price the relative risk of portfolios correctly, it should provide betas that are increasing in book-to-market ratios. Inline with this logic, the average excess returns of the first and last portfolio from the 10 portfolios sorted on the book-to-market ratio during the period 1963 - 2022 are given by 0.8676 and 1.2528, which corresponds to the portfolio with the lowest and the highest book-to-market ratio, respectively. This shows that portfolios with higher book-to-market ratio had higher average returns because of the higher level of risk they bear and therefore, the model should consider higher betas for them. Our estimation results show that when the market is in the low volatility regime, the realized risk of portfolios with higher B/M are higher than the realized risk of the portfolios with low B/M ratio. Furthermore, consistent with the theoretical frameworks for risk-return relationship, portfolios with higher B/M have higher β's than the portfolios with lower B/M. When the market is in the high volatility regime, however, we do not observe the same behavior anymore.

Even though the overall realized risk of the portfolio tend to increase when the value of the B/M becomes higher in the portfolio, the risk-return relationship no longer is captured by the CAPM β's and they seem to move in the opposite direction. One may argue that when the market is in the high volatility regime, where there is a substantial increase in the overall risk of the portfolios, the risk associated with the increase of the B/M becomes less relevant and the overall risk of the market only matters. Given that stocks are grouped

into portfolios based on their B/M characteristics, we would expect changes in the B/M to be dominant determinant of the overall changes in risk of the portfolios. However, as we will show later in our adaptive lasso estimation of the unobserved factor, there area other factors that contribute to the changes in the overall risk of the portfolios. Our estimation results show that when the market is in the high volatility state, the pricing errors (α's) becomes insignificant at %5 significance level (with the exception of the second portfolio), which is consistent with the argument made above. Even though the market β's are no longer increasing with the value of B/M, yet the market factor completely explains the cross-sectional variations in returns. However, when the market is in the low volatility regime, all the pricing errors are highly significant even at %1 significance level.

This result leads us to conclude while the capital asset pricing model doesn't perform well, when it is considered unconditionally, its performance improves substantially when it is evaluated separately in different state of the market. More specifically, while the CAPM fails in the low volatility regime, it holds almost perfectly when the market is in the high volatility regime. It is still worth noting that the point estimates of the pricing errors are noticeable in the order of the total excess returns. That means the performance of the model could further improve if we include other risk factors to more accurately explain the stock excess return. We further discuss the possible choices for additional factors improving the performance of the model in the next section.

Figure II: α Estimates for High, Low Volatility Regimes, and CAPM with 95% Confidence Bands



To provide insights into how volatile the behavior of the market is, we need to evaluate the level of persisitency of the state process and form expectations based on the historical data. Given that our analysis covers a long time span of almost six decades, starting from

1963 to 2022, we can expect that the persistency of the regimes would not be close to 1. The United States economy has experienced several periods of significant economic fluctuations, financial crises, and uncertainty, such as the oil shocks in the 1970s, the global financial crisis in 2008, and the COVID-19 pandemic in 2020. Such periods of high volatility in the economy may lead to more abrupt shifts between the different market conditions and decrease the persistency of the regimes. Moreover, it is well known that the behavior of the economy is subject to several exogenous shocks that can significantly affect its trajectory, such as changes in government policies, geopolitical tensions, and technological innovations. These factors can further increase the uncertainty and volatility of the economy, and hence the persistency of the regimes may not be very high. As discussed in section 2, the autoregressive coefficient λ of the process defined for the latent factor (ω_t) represents the persistency of the state process (s_t). A value of λ close to 1 implies that the states are highly persistent, while a value close to 0 indicates that the states are less persistent and can change more rapidly over time. On the other hand, by looking at any indicators of economics uncertainty, such as the Economic Policy Uncertainty (EPU) Index introduced by Baker et al. (2016), one can easily see that economics uncertainty levels have often been realtively low in the US economy. This can also be an indication that level of persistency of the regimes should be decently far from 0. The estimated value of the autoregressive coefficient in our baseline model is 0.6137, which is reasonably consistent with what one may expect from the US economy. This implies that the states are relatively persistent but still subject to change over time, and the market conditions are not entirely fixed. Hence, investors need to be aware of the possibility of sudden shifts in the market conditions and adjust their investment strategies accordingly. Additionally, the estimated value of the autoregressive coefficient can be useful in predicting the future behavior of the market and assessing the risks associated with different investment strategies.

Figure III presents the extracted latent factors for ten portfolios sorted on BE/ME ratio. The shaded areas represent the periods of high volatility, which consists of the recession periods announced by the National Bureau of Economic Research (NBER), the Dot-com bubble where most of the internet companies lost almost half of their value of the from March to December 2000 and the bear market that started in march 2022. As stated before, when the value of the latent factor is above the threshold τ (red dashed line), the market is in the high volatility regime and it is expected to be synchronous with the periods of financial crises or macroeconomic recession periods. When the latent factor is below the threshold τ, the market is in the low volatility regime which is generally characterized by the behavior of the market in normal times. The recession periods stated by the NBER are all recognized by the extracted latent factor as periods of high stock market volatility, except the early 1990s,

the first period of early 1980s and the relatively mild 1970 recessions. Overall, it seems that there is an alignment between periods of high macroeconomic uncertainty (recessions) and financial instability. In contrast, the low volatility periods identified by the extracted latent factor correspond to periods of relative stability in the market, where investors may be less concerned about macroeconomic risks and more focused on company-specific factors. It is worth noting that this does not necessarily mean that the market is free from risks or that it is immune to sudden shocks. In fact, even during low volatility periods, unexpected events such as geopolitical tensions or natural disasters can still trigger sudden changes in the market.

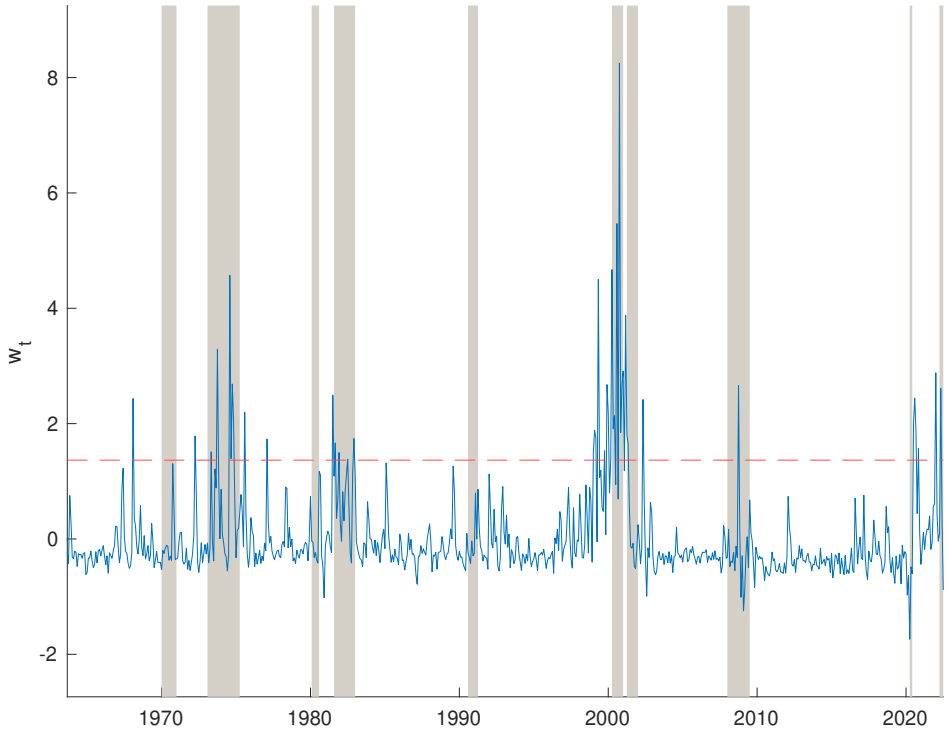
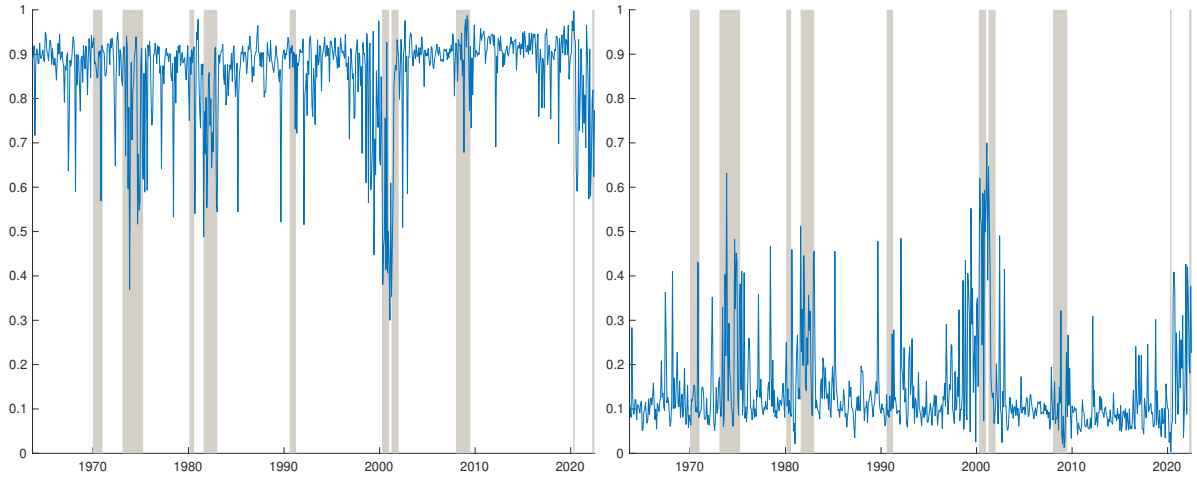


Figure III: Extracted latent factor. *Notes*: This figure presents the sample path of the latent factor extracted from the endogenous volatility switching model (solid blue line) and the threshold τ (dashed red line) along with the NBER recession periods (grey shaded area) for 3 and 5 portfolios sorted on the book-to-market ratio for the period 1964–2021, respectively, on the left and right vertical axis.

The smoothed state probabilities of being in the low and high volatility regime for the ten portfolios estimated from our endogenous regime switching model are displayed in the left- and right-hand side graphs of Figure V, respectively. As described above, the shaded areas represent the periods of high volatility. The information that can be extracted from these graphs is inline with what was discussed before. At the times that our model predicts the market should be in the high volatility state, the state probability of being the high volatility

is observed to be high. The opposite is observed for periods of low volatility. Furthermore, the the periods of high volatility, which were indicated independent of the state process, are all experiencing a high probability of being in the high volatility regime (and low probability of being in the low volatility regime) which is consistent with the results from the extracted latent factor.



(a) Low Volatility Regime (b) High Volatility Regime

Figure IV: Smoothed high and low State Probabilities. *Notes*: This figure presents the time series of the probabilities of being in the high and low volatility regimes (solid blue line) along with the NBER recession periods (grey shaded area). The left panel plots the low volatility probability series and the right panel plots the high volatility probability series obtained from the endogenous volatility switching model

To better understand what are the main drivers of the state process (s_t), we perform an adaptive lasso regression over the extracted latent factor ω_t. Adaptive lasso is a statistical technique that uses a penalty term to encourage certain coefficients to be shrunk to zero. The idea is to automatically select the most important variables by adding a penalty term to the loss function (the objective function that is set to be minimized) in the regression model. The adaptive Lasso technique is especially useful when the number of predictors is much larger than the number of observations, which is often the case in financial time series analysis. The resulting coefficients are then used to rank the importance of each variable. In our case, we are interested in identifying the main factors that affect financial uncertainty, as measured by the regimes of our endogenous switching model. To perform the adaptive lasso regression, we use the monthly Federal Reserve Economics Data (FRED). The FRED-MD is a macroeconomic database of 134 monthly U.S. indicators that was developed by the FRED data desk at the Federal Reserve Bank of St. Louis to facilitate macroeconometrics analysis. The database is a data-rich environment, meaning that it enables analysis of a

large number of variables without sacrificing information in the time series dimension. To obtain a more comprehensive understanding of the temporal dynamics, meaning whether the relationships between the predictors and the response variable changed over different time periods, we divided the data into 33 rolling windows of 25 years each. We then applied the adaptive lasso method separately to each window to obtain the regression coefficients and the corresponding variable importance scores. By doing so, we were able to examine how the importance of different predictors changed over time. To visualize the results, we plotted the variable importance scores over the rolling windows in a heat map. The variable on the horizontal axis represents the last year of the rolling window. The color of each cell in the heat map indicates the relative importance of the corresponding predictor for the given time period. By examining the heat map, we were able to identify the predictors that consistently had high importance scores over time, as well as those that exhibited temporal variations in importance. To improve the clarity of the visualization, we only plotted the variables that appeared in at least 5 of the 33 rolling windows of length 25 years. This criterion ensured that only the most robust variables were displayed on the heat map.

Figure V presents the heat map representing our results from the adaptive lasso over the latent factor. The heat map shows that certain factors had a more significant impact on the state of the market during certain periods, while their importance diminished during other periods. At the same time, some other factors only start affecting the market dynamics in the later periods. In a nutshell, the heat map suggests that the main factors affecting the regimes are: 1. VIX, which is a measure of market volatility and often used as a gauge of investor sentiment and risk appetite. 2. The spread between the 10-Year Treasury C Minus FEDFUNDS, which is an indicator of the yield curve slope and can signal the market's expectation of future economic growth and inflation. 3. The spread between the 3-Month Treasury C Minus FEDFUNDS, which is an indicator of short-term interest rate expectations and can also signal the market's expectation of future economic conditions. 4. Effective Federal Funds Rate, which is the interest rate at which depository institutions lend balances at the Federal Reserve to other depository institutions overnight. This is a key policy tool of the Federal Reserve and can signal the stance of monetary policy. 5. All Employees: Wholesale Trade, which is a measure of the employment in the wholesale trade sector and can indicate the level of economic activity and demand for goods. 6. New Private Housing Permits, Midwest (SAAR), which is a measure of the number of new building permits issued for private housing in the Midwest region and can signal the level of construction activity and the health of the housing market. These variables are all economically meaningful and have been studied extensively in the literature as indicators of financial uncertainty and market volatility. Our findings suggest that a combination of market volatility, monetary policy,

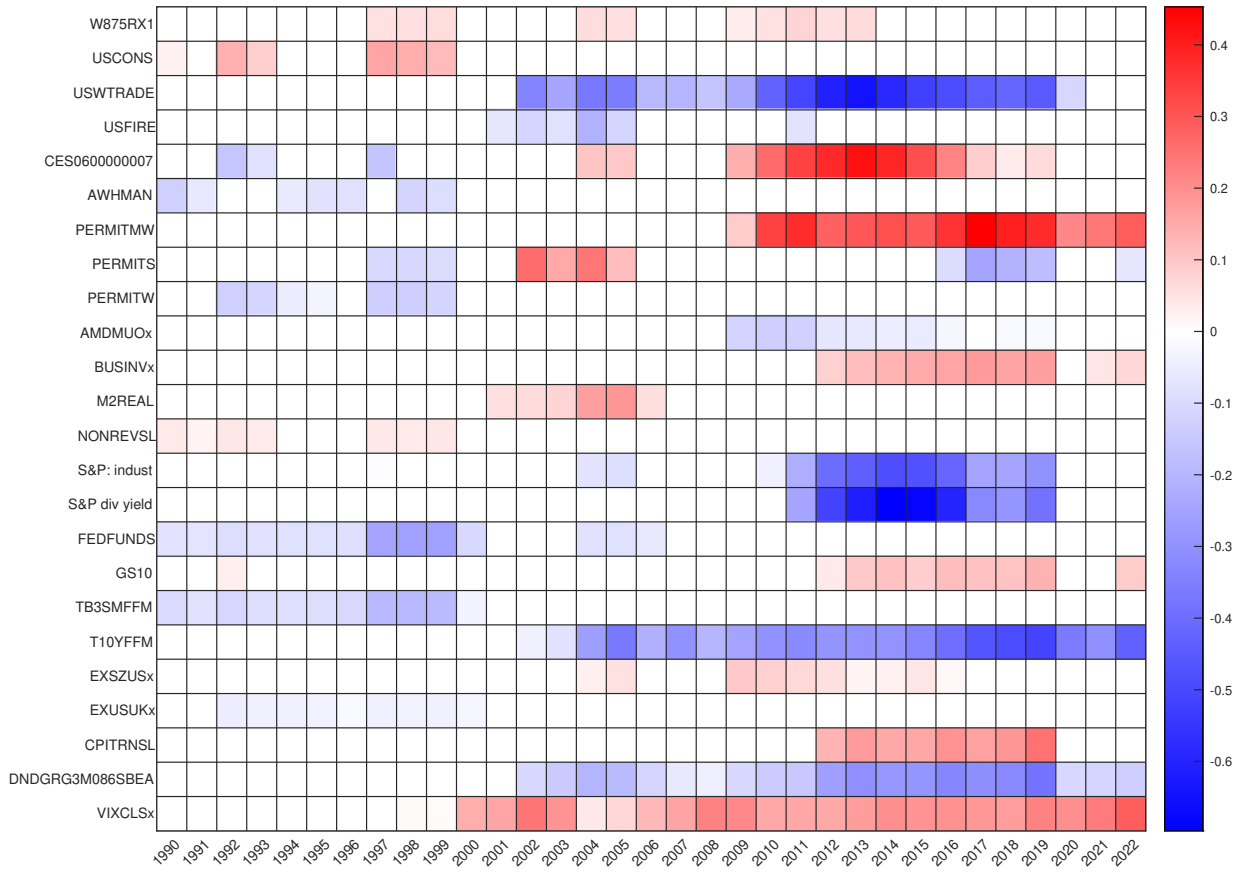


Figure V: Smoothed high and low State Probabilities. *Notes*: This figure presents the time series of the probabilities of being in the high and low volatility regimes (solid blue line) along with the NBER recession periods (grey shaded area). The left panel plots the low volatility probability series and the right panel plots the high volatility probability series obtained from the endogenous volatility switching model

economic activity, and housing market conditions are the main factors affecting financial uncertainty and provide further support for the validity of our regime-switching model.

To incorporate the argument that CAPM neglects the importance of other explanatory variables constructed based on the characteristics of the portfolios (stocks), we perform an adaptive lasso regression over the common component of the error term (u_t). As discussed before, the point estimates of the pricing errors are relatively large even at the high volatility regime and this approach allows us to identify variables that are correlated with the error term and thus can potentially improve the model's predictive power. To identify the u_t, we performed principal component analysis (PCA) on the error term $\varepsilon_{it}(s_t)$. PCA is a statistical technique that can be used to reduce the dimensionality of a data set by identifying linear combinations of variables that capture the most variation in the data. In this case, PCA is used to identify the dominant sources of variation in the error term across all the portfolios.

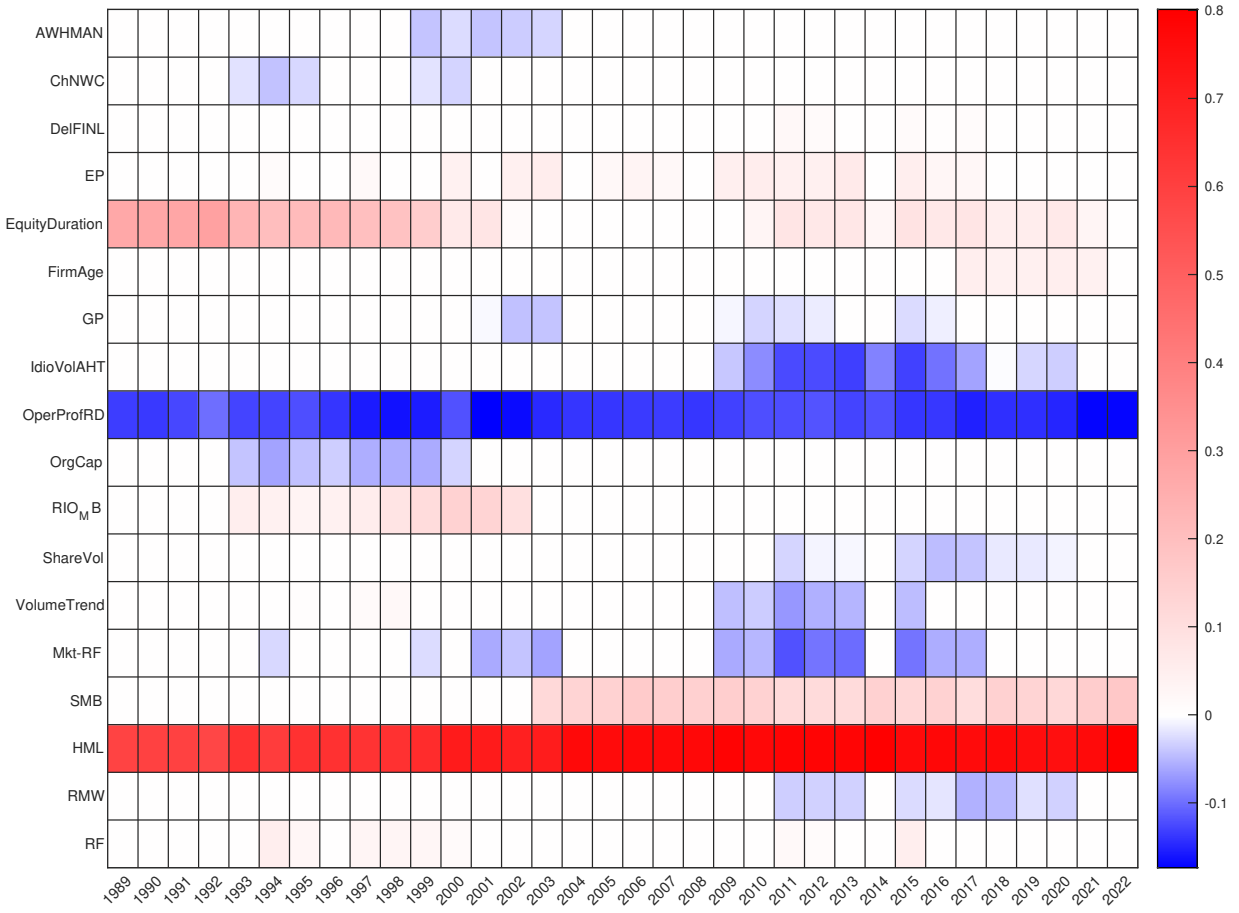


Figure VI: Smoothed high and low State Probabilities. *Notes*: This figure presents the time series of the probabilities of being in the high and low volatility regimes (solid blue line) along with the NBER recession periods (grey shaded area). The left panel plots the low volatility probability series and the right panel plots the high volatility probability series obtained from the endogenous volatility switching model

Following the extraction of u_t, which we refer to as the unobserved factor from here, we ran the adaptive lasso over the 34 rolling windows of length 25 years, similar to the previous analysis. To perform this adaptive lasso analysis, we used the 5 factors defined by Fama and French (2012) and 207 predictors reconstructed using stock's characteristics from Chen and Zimmermann (2021), in addition to the monthly Federal Reserve Economics Data (FRED).

The adaptive lasso over the unobserved factor yielded a heat map that highlights five key predictors: HML, Operating profitability R&D adjusted, Equity Duration, SMB, and Idiosyncratic risk. It's worth noting that HML and Operating profitability R&D adjusted appeared in all the rolling windows and were consistently identified as the most important factors. It's important to keep in mind that the unobserved factor is estimated from error terms that comes from a model where the dependent variables are the excess return of port-

folios sorted on book-to-market ratio. Therefore, it's not surprising that HML is identified as a principal factor with relatively large coefficients in all the windows. Equally important is the role of Operating profitability R&D adjusted, which is a measure of a company's profitability after accounting for research and development expenses. Equity Duration, SMB, and Idiosyncratic risk were also identified as key factors. Equity Duration measures the sensitivity of a stock's returns to changes in interest rates, while SMB is a measure of the size premium (i.e., the tendency for small companies to outperform larger ones). Finally, Idiosyncratic risk captures the risk that is specific to individual stocks and cannot be diversified away.

There are possible explanations for why other factors such as Operating profitability R&D adjusted were selected and not many others. One possible explanation is that the factor has a strong relationship with the dependent variable (the portfolio excess returns sorted on book-to-market ratio). For example, companies with higher operating profitability are likely to have higher earnings, which can lead to higher returns for investors. Additionally, adjusting for R&D expenses can account for the effects of investment in research and development on the profitability of a company. Another reason could be that the factor has low correlation with other factors in the model, making it a unique contributor to the variance of the dependent variable. This would make it more likely to be selected by the adaptive lasso as a significant factor. It's also possible that the other factors that were not selected by the adaptive lasso may have weak or non-existent relationships with the dependent variable, or may be redundant with having the market excess return as the sole factor in the model, making them less important in explaining the variation in the portfolio excess returns.

The analysis presented in this study highlights the importance of considering the dynamics of the market when analyzing the performance of an asset pricing model. The results demonstrate the significant impact of regime-switching behavior on the performance of the CAPM. Our findings suggest that a static CAPM may not adequately capture the complexities of the market, and that allowing for changes in the model parameters under different market regimes can improve its overall performance. The methodology presented in this study can be extended to more complex asset pricing models, providing a framework for analyzing portfolio returns under different states of the market. Furthermore, our endogenous regime-switching model can be expanded to include more than two volatility states, allowing for an even more nuanced evaluation of the performance of the CAPM or other asset pricing models. As demonstrated, the overall performance of an appropriate asset pricing model can be significantly improved by allowing the model parameters to vary between different volatility regimes of the market.

5 Alternative Portfolio Sorts and Sensitivity To Additional Risk Factors

To assess the robustness of our endogenous regime switching model, we apply it to the CAPM using various portfolio sorts and also to a two-factor model based on the B/M sort, as well as to Fama and French's three-factor model using portfolios sorted on B/M. Here, we present the results only for portfolios sorted on momentum, as similar results were obtained for portfolios sorted on investment, β, and E/P. We also conduct additional tests by limiting the time period to before 2020 (i.e., before the Covid-19 shock) or setting the start date to more recent years (1990, 1995, 2000) under the same model specification. Our findings show that the model's performance remains consistent and robust across all these alternative specifications.

In this section, we present the results of our endogenous regime model applied to portfolios sorted on momentum, as shown in Table II. The portfolios sorted on momentum are formed based on cumulative log returns from months $t-12$ through $t-2$ (11 months). The data for the portfolio returns is readily available on and downloaded from Kenneth French's website. We find that the estimated parameters are significantly different across different states of the market, similar to the case of portfolios sorted on B/M. Specifically, we observe that β tends to be higher for portfolios with higher momentum, and lower for portfolios with poor past performance. However, we find that almost all of the α coefficients become insignificant at the 5% significance level when the market is in the high volatility regime, whereas the opposite is true for the low volatility regime.

Figure VII, left- and right-hand side of Figure VII show the extracted latent factor, smoothed probability of being in the low and high volatility regime respectively.

As indicated above, to further investigate the behavior of the factor loadings under different states of the market, we added another factor, HML (High minus Low), which represents the effect of increasing the book-to-market ratio, keeping everything else constant. The HML is the difference between the returns of the portfolio with the highest BE/ME and the return of the portfolio with the lowest BE/ME. The two components of HML are returns on high and low BE/ME portfolios with about the same weighted average size and other features. Therefore, HML should be independent of other factors in returns, focusing on the different return behaviors of high and low BE/ME firms.

Table IV reports the estimation results for this new specification for 3 portfolios sorted on BE/ME. The results are consistent with the discussion made previously about the expectations one may have with respect to the behavior of the model in different volatility regimes. The market and HML factor coefficients, β and h, are both higher for the portfolios

Table II: Regime-Dependent CAPM (10 Portfolios Sorted on Momentum, 1963 - 2022)

		Low	2	3	4	5	6	7	8	9	High
Low Volatility	α	-0.7572***	-0.0681	0.2171*	0.3243***	0.277***	0.3674***	0.4095***	0.5811***	0.6802***	0.9633***
		(1.9343)	(1.2152)	(0.9341)	(0.8049)	(0.829)	(0.7246)	(0.68)	(0.5002)	(0.3955)	(0.161)
	β	1.1641***	1.0028***	0.8991***	0.9412***	0.9168***	0.9722***	0.9843***	1.0187***	1.0654***	1.2551***
		(0.0778)	(0.15)	(0.2336)	(0.182)	(0.2002)	(0.1363)	(0.1179)	(0.0929)	(0.0738)	(0.2195)
	π	2.5044***	2.2019***	1.278***	0.9325***	0.3351***	0***	0***	0***	0***	0***
		(1.7214)	(1.0398)	(0.2011)	(0.3348)	(0.8287)	(1.0825)	(1.0938)	(1.0938)	(1.0938)	(1.0938)
High Volatility	α	0.2097	0.5288	0.6405	0.7294*	0.6628**	0.5742*	0.5739**	0.3771**	0.1655	0.1939
		(1.1957)	(0.7255)	(0.5882)	(0.452)	(0.5371)	(0.6449)	(0.594)	(0.7496)	(0.9923)	(1.2336)
	β	2.2768***	1.7272***	1.3417***	1.0666***	0.9269***	0.8515***	0.6795***	0.6487***	0.7108***	0.9097***
		(1.2085)	(0.6835)	(0.3011)	(0.0997)	(0.1659)	(0.2485)	(0.4236)	(0.4638)	(0.4287)	(0.2778)
	π	3.9689***	3.9988***	4.3488***	3.2848***	2.4438***	2.1802***	1.2368***	0.3456***	0***	0***
		(3.2376)	(3.3876)	(3.7273)	(2.4675)	(1.5826)	(1.27)	(0.6189)	(0.8853)	(1.0576)	(1.0938)
	σ	3.5032***	1.5073***	1.4015***	1.3534***	1.3639***	1.3762***	1.5128***	1.5761***	1.8917***	3.1672***
		(2.3593)	(0.4205)	(0.2966)	(0.2746)	(0.2538)	(0.2745)	(0.377)	(0.416)	(0.7335)	(2.0134)
	ρ					-0.1755					
						(1.1539)					
	λ					0.8235***					
						(0.3197)					
	τ					1.561***					
						(0.5556)					

Notes: The standard errors are calculated using wild residual bootstrap method and reported in parenthesis. Significance is computed using wild residual bootstrap confidence intervals. *p<0.1; **p<0.05; ***p<0.01

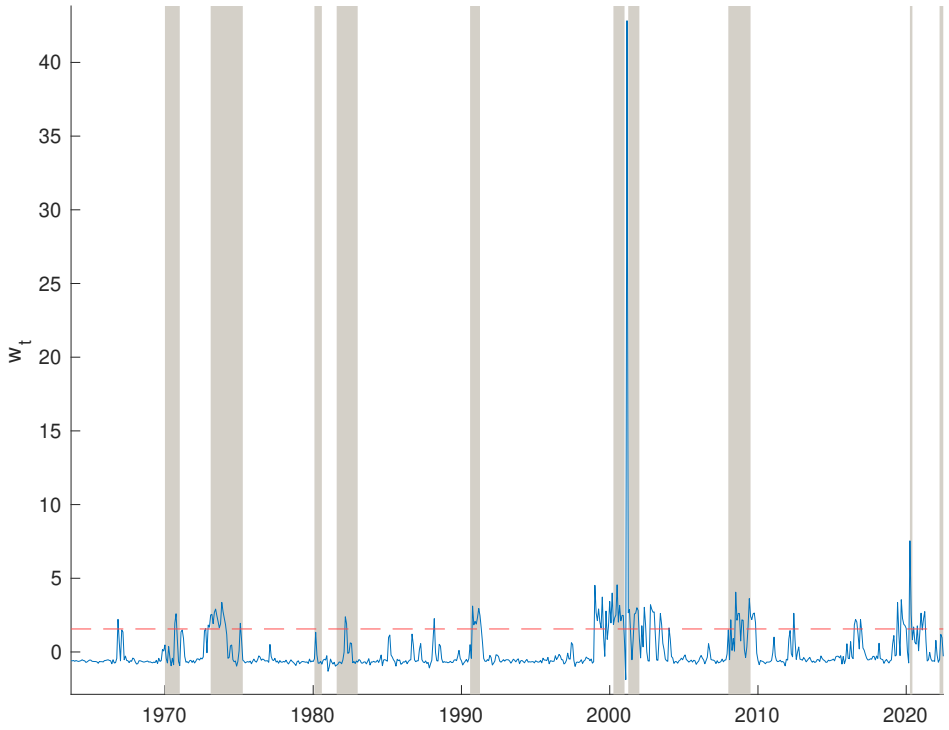


Figure VII: Extracted latent factor. *Notes*: This figure presents the sample path of the latent factor extracted from the endogenous volatility switching model (solid blue line) and the threshold τ (dashed red line) along with the NBER recession periods (grey shaded area) for 3 and 5 portfolios sorted on the book-to-market ratio for the period 1964–2021, respectively, on the left and right vertical axis.

with higher book-to-market ratios when the market is in the low volatility regime. However, this behavior no longer can be observed when we look at the estimated parameters in the high volatility regime.

The pricing errors are behaving similarly to the state-dependent CAPM where the α's in the low volatility regime are smaller than their corresponding value in the high volatility regime (except for the first portfolio). In addition, by comparing the pricing errors in the low volatility regime in this specification with the ones in the regime-dependent CAPM, we can easily see that the magnitude of the pricing errors is getting smaller (except for the first portfolio).

6 Conclusion

In this article, we proposed a new approach to model a panel regression with regime-switching using a latent autoregressive factor. In this setup, we test the performance of the Capital Asset Pricing Model (CAPM) where we allow for discrete time-variations in the CAPM betas

for portfolios that are sorted by the book-to-market ratios based on a two-state endogenous regime-switching process determined by the uncertainty observed in the stock market return behaviors. Our method has a couple of advantages by using an endogenous regime-switching setup rather than a Markov-switching process. We found that the behavior of this asset pricing model significantly differs across different volatility regimes and returns behave more closely to the rules of the proposed models. Even though the regime-dependent version of the CAPM can still be rejected, it provides strong evidence on how important it is to consider the occasional shifts observed in the market return when we want to evaluate the performance of an asset pricing model. In addition, based on the information that we can extract from the latent factor, there seems to be a correlation between the periods of high volatility and economic recessions such that the latter is a subset of the former, according to our empirical findings.

References

Abdymomunov, A. and J. Morley (2011). Time variation of capm betas across market volatility regimes. *Applied Financial Economics 21*, 1463–1478.

Chang, Y., Y. Choi, and J. Y. Park (2017). A new approach to model regime switching. *Journal of Econometrics 196*, 127–143.

Chen, J. and Y. Kawaguchi (2018). Multi-factor asset-pricing models under markov regime switches: Evidence from the chinese stock market. *International Journal of Financial Studies 6*, 1–19.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *The Journal of Finance 47*, 427–465.

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

Harvey, C. R., Y. Liu, and H. Zhu (2015). ... and the cross-section of expected returns. *The Review of Economics and Statistics 29*, 5–68.

Jagannathan, R. and Z. Wang (1996). The conditional capm and the cross-section of expected returns. *The Journal of Finance 51*, 3–53.

Lakonishok, J., A. Shleifer, and R. Vishny (1994). Contrarian investment, extrapolation, and risk. *The Journal of Finance 49*, 1541–1578.

Lettau, M. and S. Ludvigson (2002). Consumption, aggregate wealth, and expected stock returns. *The Journal of Finance 56*, 815–849.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics 47*, 13–37.

Petkova, R. and L. Zhang (2005). Is value riskier than growth? *Journal of Financial Economics 78*, 187–202.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance 19*, 425–442.

Tu, J. (2010). Is regime switching in stock returns important in portfolio decisions? *Management Science 56*, 1198–1215.

A Inverse of Covariance Matrix

We may find the determinant and inverse of covariance matrix $\Omega(s_t)$ of $\varepsilon_t(s_t)$ analytically, which would be very useful in computing the likelihood function. Define

$$\Sigma = \text{diag}\ (\sigma_1^2, ..., \sigma_N^2)$$

and write

$$\Omega(s_t) = \Sigma + \pi(s_t)\pi(s_t)' = \Sigma^{1/2}(I + \Sigma^{-1/2}\pi(s_t)\pi(s_t)'\Sigma^{-1/2})\Sigma^{1/2} \tag{25}$$

where $\Omega(s_t)$ is defined in (16).

Let $\tau(s_t) = \Sigma^{-1/2}\pi(s_t)$, and note that

$$\begin{aligned} I + \Sigma^{-1/2}\pi(s_t)\pi(s_t)'\Sigma^{-1/2} &= I + \tau(s_t)\tau(s_t)' \\ &= I + ||\tau(s_t)||^2 P_{\tau(s_t)} \\ &= (1 + ||\tau(s_t)||^2)P_{\tau(s_t)} + (I - P_{\tau(s_t)}), \end{aligned}$$

where $P_{\tau(s_t)} = \tau(s_t)\tau(s_t)'/||\tau(s_t)||^2$ is the orthogonal projection on the span of $\tau(s_t)$, from which it follows immediately that

$$\left(I + \Sigma^{-1/2}\pi(s_t)\pi(s_t)'\Sigma^{-1/2}\right)^{-1} = \frac{1}{1 + ||\tau(s_t)||^2}P_{\tau(s_t)} + (I - P_{\tau(s_t)}) = I - \frac{||\tau(s_t)||^2}{1 + ||\tau(s_t)||^2}P_{\tau(s_t)}. \tag{26}$$

Therefore, we may deduce from (25) and (26) that

$$\Omega^{-1}(s_t) = \Sigma^{-1/2}\left(I - \frac{||\tau(s_t)||^2}{1 + ||\tau(s_t)||^2}P_{\tau(s_t)}\right)\Sigma^{-1/2} = \Sigma^{-1} - \frac{1}{1 + \pi(s_t)'\Sigma^{-1}\pi(s_t)}\Sigma^{-1}\pi(s_t)\pi(s_t)'\Sigma^{-1} \tag{27}$$

Moreover, we have

$$\det \Omega(s_t) = (\det \Sigma)(1 + \pi(s_t)'\Sigma^{-1}\pi(s_t)) \tag{28}$$

due to (25).

Finally, we may also easily derive that

$$1 - \rho^2\pi(s_t)'\Omega^{-1}(s_t)\pi(s_t) = \frac{1 + (1 - \rho^2)\pi(s_t)'\Sigma^{-1}\pi(s_t)}{1 + \pi(s_t)'\Sigma^{-1}\pi(s_t)}$$

and

$$\Omega^{-1}(s_t)\pi(s_t) = \frac{\Sigma^{-1}\pi(s_t)}{1 + \pi(s_t)'\Sigma^{-1}\pi(s_t)}$$

from (27).

B Mathematical Proofs

We provide a brief proof for some of the expressions presented in the main part of the paper.

- **Equation** (8)**:**
  According to how we defined the latent factor in (4) and the assumption of normality for the error term, it follows that

$$\begin{aligned}\mathbb{P}\{w_t < \tau | w_{t-1}\} &= \mathbb{P}\{\lambda w_{t-1} + v_t < \tau | w_{t-1}\} \\ &= \mathbb{P}\{v_t < \tau - \lambda w_{t-1} | w_{t-1}\} \\ &= \Phi\left(\tau - \lambda w_{t-1}\right) \quad \square\end{aligned}$$

- **Equation** (10)**&**(11)**:**
  From (8), we may easily write

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

  It follows that

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

  since $w_{t-1}\sqrt{1-\lambda^2} =_d \mathbb{N}(0,1)$. Similarly, we have

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

from which it follows that

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

The proof for the case of $\lambda = 1$ in equations (13) and (14) is very similar to what we have done here, except that we have $w_{t-1}/\sqrt{t-1} =_d \mathbb{N}(0,1)$ for $t \geq 2$ in this case instead of $w_{t-1}\sqrt{1-\lambda^2} =_d \mathbb{N}(0,1)$ when $|\lambda| < 1$. □

- **Equation** (15)**:**
  For any normal random vector $X = (X_1, ..., X_k)'$ with mean μ and covariance matrix Σ, the probability density function can be written as

$$f_X(x) = \frac{exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^k|\Sigma|}}.$$

  In our panel regression model, we have

$$y_t = \alpha(s_t) + \beta(s_t)x_t + \sigma(s_t)u_t$$

  where

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_1' \\ \vdots \\ \beta_N' \end{pmatrix}, \ \sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \end{pmatrix}$$

  We may easily see

$$\begin{aligned}
\mathbb{E}(y_t|s_t) &= \alpha(s_t) + \beta'(s_t)x_t \\
Var(y_t|s_t) &= \mathbb{E}\Big(\varepsilon_t(s_t)\varepsilon_t(s_t)'|s_t\Big) = \Omega(s_t)
\end{aligned}$$

  since $\mathbb{E}\big(\varepsilon_t(s_t)|s_t\big) = 0_N$. If we apply the above density function to this regression

model, we easily derive (15). □

- **Equation** (18)**:**
  We may rewrite (6) as

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

  It follows that

$$\left. \begin{pmatrix} v_{t+1} \\ \varepsilon_t(s_t) \end{pmatrix} \right|_{s_t} =_d \mathbb{N}\left(0_{N+1}, \begin{pmatrix} 1 & \rho\pi'(s_t) \\ \rho\pi(s_t) & \Omega(s_t) \end{pmatrix} \right).$$

  Generally, if we partition a normal random vector $X =_d (\mu, \Sigma)$ as

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

  where X_1 and X_2 are n_1- and n_2-dimensional, respectively, we may write

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

  Then, the conditional distribution of X_1 given X_2 is given by

$$p(X_1|X_2) =_d \mathbb{N}(\mu_{1.2}, \Sigma_{11.2})$$

  where $\mu_{1.2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ and $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. If we apply the above formula to the times series and latent factor innovations, we may easily get

$$p(v_t|s_{t-1}, \varepsilon_{t-1}) =_d \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, 1 - \rho^2\pi'\Omega^{-1}\pi\right). \quad \square$$

- **Equation** (20)**:**
  Note that for the identification of our model in the case of $|\lambda| < 1$, we assumed that $w_{t-1} =_d N(0, 1/(1-\lambda^2))$. Let us define

$$z_t = \frac{w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1 - \rho^2\pi'\Omega^{-1}\pi}} - \frac{\lambda w_{t-1}}{\sqrt{1 - \rho^2\pi'\Omega^{-1}\pi}}$$

Based on what we had in (18), we can derive that

$$p(z_t|w_{t-1},\mathcal{F}_{t-1}) =_d \mathbb{N}(0,1).$$

It follows that

$$\begin{aligned}\mathbb{P}\left\{w_t<\tau|w_{t-1},\mathcal{F}_{t-1}\right\} &= \mathbb{P}\left\{z_t<\frac{\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}-\frac{\lambda w_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\middle| w_{t-1},\mathcal{F}_{t-1}\right\}\\ &= \Phi\left(\frac{\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}-\frac{\lambda w_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)\end{aligned}$$

Note that the latent factor at time t only depends on its lagged value of w_{t-1} and the error term v_t which is correlated with u_{t-1} and independent of w_{t-1}. This means that $p(w_t|w_{t-1},\mathcal{F}_{t-1}) = p(w_t|w_{t-1},\varepsilon_{t-1})$ and we may deduce that

$$\begin{aligned}\mathbb{P}\left\{s_t=0|s_{t-1}=0,\mathcal{F}_{t-1}\right\} &= \mathbb{P}\left\{w_t<\tau|w_{t-1}<\tau,\mathcal{F}_{t-1}\right\}\\ &= \mathbb{P}\left\{w_t<\tau|w_{t-1}\sqrt{1-\lambda^2}<\tau\sqrt{1-\lambda^2},\mathcal{F}_{t-1}\right\}\\ &= \frac{\int_{-\infty}^{\tau\sqrt{1-\lambda^2}}\Phi\left(\frac{\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}-\frac{\lambda x}{\sqrt{(1-\lambda^2)\left(1-\rho^2\pi'\Omega^{-1}\pi\right)}}\right)\varphi(x)dx}{\Phi\left(\tau\sqrt{1-\lambda^2}\right)}\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{P}\left\{s_t=0|s_{t-1}=1,\mathcal{F}_{t-1}\right\} &= \mathbb{P}\left\{w_t<\tau|w_{t-1}\geq\tau,\mathcal{F}_{t-1}\right\}\\ &= \mathbb{P}\left\{w_t<\tau|w_{t-1}\sqrt{1-\lambda^2}\geq\tau\sqrt{1-\lambda^2},\mathcal{F}_{t-1}\right\}\\ &= \frac{\int_{\tau\sqrt{1-\lambda^2}}^{\infty}\Phi\left(\frac{\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}-\frac{\lambda x}{\sqrt{(1-\lambda^2)\left(1-\rho^2\pi'\Omega^{-1}\pi\right)}}\right)\varphi(x)dx}{1-\Phi\left(\tau\sqrt{1-\lambda^2}\right)}\end{aligned}$$

By combining the above equations we may easily derive (20). The proof for the case of $\lambda = 1$ in equation (21) is very similar to what we have done here, except that we have $w_{t-1}/\sqrt{t-1} =_d \mathbb{N}(0,1)$ for $t \geq 2$ in this case instead of $w_{t-1}\sqrt{1-\lambda^2} =_d \mathbb{N}(0,1)$ when $|\lambda| < 1$. □

- ω_ρ **of** (s_t) **when** $0 < \lambda < 1$ **and** $|\rho| = 1$**:**

Note that

$$\begin{aligned}\mathbb{P}\left\{w_t < \tau | w_{t-1}, \mathcal{F}_{t-1}\right\} &= \mathbb{P}\left\{\lambda w_{t-1} + v_t < \tau | w_{t-1}, \mathcal{F}_{t-1}\right\} \\ &= \mathbb{P}\left\{\lambda w_{t-1} + \rho\pi'\Omega^{-1}\varepsilon_{t-1} < \tau | w_{t-1}, \varepsilon_{t-1}\right\} \\ &= 1\left\{\lambda w_{t-1} + \rho\pi'\Omega^{-1}\varepsilon_{t-1} < \tau\right\}.\end{aligned}$$

Consequently, we may write

$$\begin{aligned}\mathbb{P}\left\{s_t = 0 | s_{t-1} = 0, \mathcal{F}_{t-1}\right\} &= \mathbb{P}\left\{w_t < \tau | w_{t-1} < \tau, \mathcal{F}_{t-1}\right\} \\ &= \mathbb{P}\left\{\lambda w_{t-1} + \rho\pi'\Omega^{-1}\varepsilon_{t-1} < \tau | w_{t-1} < \tau, \mathcal{F}_{t-1}\right\} \\ &= \mathbb{P}\Bigg\{w_{t-1}\sqrt{1-\lambda^2} < \frac{\sqrt{1-\lambda^2}}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\Big| \\ &\quad \times w_{t-1}\sqrt{1-\lambda^2} < \tau\sqrt{1-\lambda^2}, \mathcal{F}_{t-1}\Bigg\} \\ &= \begin{cases} 1, & \text{if } \frac{1}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1}) \geq \tau, \\ \frac{\Phi\left(\frac{\sqrt{1-\lambda^2}}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)}{\Phi(\tau\sqrt{1-\lambda^2})}, & \text{otherwise.} \end{cases}\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{P}\left\{s_t = 0 | s_{t-1} = 1, \mathcal{F}_{t-1}\right\} &= \mathbb{P}\left\{w_t < \tau | w_{t-1} \geq \tau, \mathcal{F}_{t-1}\right\} \\ &= \mathbb{P}\left\{\lambda w_{t-1} + \rho\pi'\Omega^{-1}\varepsilon_{t-1} < \tau | w_{t-1} \geq \tau, \mathcal{F}_{t-1}\right\} \\ &= \mathbb{P}\Bigg\{w_{t-1}\sqrt{1-\lambda^2} < \frac{\sqrt{1-\lambda^2}}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\Big| \\ &\quad \times w_{t-1}\sqrt{1-\lambda^2} \geq \tau\sqrt{1-\lambda^2}, \mathcal{F}_{t-1}\Bigg\} \\ &= \frac{\Phi\left(\frac{\sqrt{1-\lambda^2}}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right) - \Phi(\tau\sqrt{1-\lambda^2})}{1 - \Phi(\tau\sqrt{1-\lambda^2})} \\ &\quad \times 1\left\{\frac{1}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1}) \geq \tau\right\}\end{aligned}$$

When $-1 < \lambda < 0$, with a similar approach, we may easily see that

$$\mathbb{P}\left\{s_t = 0 | s_{t-1} = 0, \mathcal{F}_{t-1}\right\} = \begin{cases} 0, & \text{if } \frac{1}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1}) \geq \tau, \\ \frac{\Phi(\tau\sqrt{1-\lambda^2}) - \Phi\left(\frac{\sqrt{1-\lambda^2}}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)}{\Phi(\tau\sqrt{1-\lambda^2})}, & \text{otherwise.} \end{cases}$$

and

$$\mathbb{P}\left\{s_t = 0 | s_{t-1} = 1, \mathcal{F}_{t-1}\right\} = \begin{cases} 1, & \text{if } \frac{1}{\lambda}(\tau - \rho\pi'\Omega^{-1}\varepsilon_{t-1}) < \tau, \\ \frac{1-\Phi\left(\frac{\sqrt{1-\lambda^2}}{\lambda}(\tau-\rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)}{1-\Phi(\tau\sqrt{1-\lambda^2})}, & \text{otherwise.} \end{cases}$$

The proof for the case of $\lambda = 0$ is trivial and the proof for the case of $\lambda = 1$ is very similar to what we did for $0 < \lambda < 1$, except that we have $w_{t-1}/\sqrt{t-1} =_d \mathbb{N}(0,1)$ for $t \geq 2$ in this case instead of $w_{t-1}\sqrt{1-\lambda^2} =_d \mathbb{N}(0,1)$ when $|\lambda| < 1$. □

- ω_ρ **of** (w_t) **when** $|\lambda| < 1$ **and** $|\rho| < 1$**:**
  Based on how we define our latent autoregressive process, we may easily see that

$$w_t | w_{t-1}, \varepsilon_{t-1} =_d \mathbb{N}(\lambda w_{t-1} + \rho\pi'\Omega^{-1}\varepsilon_{t-1}, 1 - \rho^2\pi'\Omega^{-1}\pi).$$

  It follows that

$$p(w_t | w_{t-1}, \varepsilon_{t-1}) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}} \exp\left[-\frac{1}{2}\left(\frac{w_t - \lambda w_{t-1} - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2\right].$$

  Note that for the conditional transition density of the latent factor, we have

$$\begin{aligned} p(w_t | s_{t-1} = 1, \mathcal{F}_{t-1}) &= p(w_t | s_{t-1} = 1, y_{t-1}, ..., y_1) \\ &= p(w_t | w_{t-1} \geq \tau, \varepsilon_{t-1}) \\ &= \frac{\int_\tau^\infty p(w_t, w_{t-1}, \varepsilon_{t-1}) dw_{t-1}}{\int_\tau^\infty p(w_{t-1}, \varepsilon_{t-1}) dw_{t-1}} \\ &= \frac{\int_\tau^\infty p(w_t | w_{t-1}, \varepsilon_{t-1}) p(w_{t-1}, \varepsilon_{t-1}) dw_{t-1}}{\int_\tau^\infty p(w_{t-1}) p(\varepsilon_{t-1}) dw_{t-1}} \\ &= \frac{\int_\tau^\infty p(w_t | w_{t-1}, \varepsilon_{t-1}) p(w_{t-1}) dw_{t-1}}{\int_\tau^\infty p(w_{t-1}) dw_{t-1}}. \end{aligned}$$

  With a similar approach, we may derive

$$p(w_t | s_{t-1} = 0, \mathcal{F}_{t-1}) = \frac{\int_{-\infty}^\tau p(w_t | w_{t-1}, \varepsilon_{t-1}) p(w_{t-1}) dw_{t-1}}{\int_{-\infty}^\tau p(w_{t-1}) dw_{t-1}}$$

Since $w_t =_d w_{t-1} =_d \mathbb{N}(0, 1/(1-\lambda^2))$, it follows that

$$p(w_{t-1}) = \frac{\sqrt{1-\lambda^2}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} w_{t-1}^2 (1-\lambda^2)\right).$$

With a simple multiplication, we may get

$$p(w_t|w_{t-1}, \varepsilon_{t-1}) p(w_{t-1}) = \frac{\sqrt{1-\lambda^2}}{2\pi\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}} \times \exp\left[\underbrace{-\frac{1}{2}\left(\frac{w_t - \lambda w_{t-1} - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2 - \frac{1}{2} w_{t-1}^2 (1-\lambda^2)}_{-\frac{1}{2}C}\right].$$

Let us simplify C as follow

$$\begin{aligned}
C &= \left(\frac{w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2 + \frac{\lambda^2 w_{t-1}^2}{1-\rho^2\pi'\Omega^{-1}\pi} - \frac{2\lambda(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})w_{t-1}}{1-\rho^2\pi'\Omega^{-1}\pi} + w_{t-1}^2(1-\lambda^2) \\
&= \left(\frac{w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2 + \frac{(1-\rho^2\pi'\Omega^{-1}\pi + \lambda^2\rho^2\pi'\Omega^{-1}\pi) w_{t-1}^2}{1-\rho^2\pi'\Omega^{-1}\pi} - \frac{2\lambda(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})w_{t-1}}{1-\rho^2\pi'\Omega^{-1}\pi} \\
&= \left(\sqrt{\frac{1-\rho^2\pi'\Omega^{-1}\pi + \lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\rho^2\pi'\Omega^{-1}\pi}} w_{t-1} - \frac{\lambda(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})}{\sqrt{(1-\rho^2\pi'\Omega^{-1}\pi)(1-\rho^2\pi'\Omega^{-1}\pi + \lambda^2\rho^2\pi'\Omega^{-1}\pi)}}\right)^2 \\
&+ \left(\frac{w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2 - \frac{\lambda^2(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})^2}{(1-\rho^2\pi'\Omega^{-1}\pi)(1-\rho^2\pi'\Omega^{-1}\pi + \lambda^2\rho^2\pi'\Omega^{-1}\pi)} \\
&= \left(\sqrt{\frac{1-\rho^2\pi'\Omega^{-1}\pi + \lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\rho^2\pi'\Omega^{-1}\pi}} w_{t-1} - \frac{\lambda(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})}{\sqrt{(1-\rho^2\pi'\Omega^{-1}\pi)(1-\rho^2\pi'\Omega^{-1}\pi + \lambda^2\rho^2\pi'\Omega^{-1}\pi)}}\right)^2 \\
&+ \frac{(1-\lambda^2)(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})^2}{1-\rho^2\pi'\Omega^{-1}\pi + \lambda^2\rho^2\pi'\Omega^{-1}\pi}.
\end{aligned}$$

By substituting the expression we got for C, we can get

$$p(w_t|w_{t-1},\varepsilon_{t-1})p(w_{t-1}) = \frac{\sqrt{1-\lambda^2}}{2\pi\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}$$
$$\times \underbrace{\exp\left[-\frac{1}{2}\left(\sqrt{\frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\rho^2\pi'\Omega^{-1}\pi}}\left(w_{t-1}-\frac{\lambda(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{(1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi)}\right)\right)^2\right]}_{K}$$
$$\times \underbrace{\exp\left[-\frac{1}{2}\left(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{\frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\lambda^2}}}\right)^2\right]}_{S}.$$

We may write

$$\begin{aligned}p(w_t|w_{t-1},u_{t-1})p(w_{t-1}) &= \frac{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}}{\sqrt{2\pi}\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}K\\ &\times\frac{\sqrt{1-\lambda^2}}{\sqrt{2\pi}\sqrt{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}}S\\ &= \bar{K}\times\bar{S}.\end{aligned}$$

Note that

$$\bar{S} = \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1},\frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\lambda^2}\right).$$

It follows that

$$\begin{aligned}p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) &= \frac{\int_\tau^\infty \bar{K}\times\bar{S}dw_{t-1}}{\int_\tau^\infty p(w_{t-1})dw_{t-1}}\\ &= \frac{\int_\tau^\infty \bar{K}dw_{t-1}}{\int_\tau^\infty p(w_{t-1})dw_{t-1}}\bar{S}\\ &= \frac{1-\int_{-\infty}^\tau \bar{K}dw_{t-1}}{\int_\tau^\infty p(w_{t-1})dw_{t-1}}\bar{S}\\ &= \frac{1-\Phi\left(\sqrt{\frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\rho^2\pi'\Omega^{-1}\pi}}\left(\tau-\frac{\lambda(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{(1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi)}\right)\right)}{1-\Phi(\tau\sqrt{1-\lambda^2})}\\ &\times\mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1},\frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\lambda^2}\right).\end{aligned}$$

Similarly, we may derive

$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\Phi\left(\sqrt{\frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\rho^2\pi'\Omega^{-1}\pi}}\left(\tau-\frac{\lambda(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{(1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi)}\right)\right)}{\Phi(\tau\sqrt{1-\lambda^2})} \times \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, \frac{1-\rho^2\pi'\Omega^{-1}\pi+\lambda^2\rho^2\pi'\Omega^{-1}\pi}{1-\lambda^2}\right).$$

- ω_ρ **of** (w_t) **when** $\lambda = 1$ **and** $|\rho| < 1$**:**
  Based on how we define our latent autoregressive process, we may easily see that

$$w_t|w_{t-1},\varepsilon_{t-1} =_d \mathbb{N}(w_{t-1}+\rho\pi'\Omega^{-1}\varepsilon_{t-1}, 1-\rho^2\pi'\Omega^{-1}\pi).$$

  It follows that

$$p(w_t|w_{t-1},\varepsilon_{t-1}) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\exp\left[-\frac{1}{2}\left(\frac{w_t-w_{t-1}-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2\right].$$

  Note that for the conditional transition density of the latent factor, we have

$$\begin{aligned}
p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) &= p(w_t|s_{t-1}=1,y_{t-1},...,y_1)\\
&= p(w_t|w_{t-1}\geq\tau,\varepsilon_{t-1})\\
&= \frac{\int_\tau^\infty p(w_t,w_{t-1},\varepsilon_{t-1})dw_{t-1}}{\int_\tau^\infty p(w_{t-1},\varepsilon_{t-1})dw_{t-1}}\\
&= \frac{\int_\tau^\infty p(w_t|w_{t-1},\varepsilon_{t-1})p(w_{t-1},\varepsilon_{t-1})dw_{t-1}}{\int_\tau^\infty p(w_{t-1})p(\varepsilon_{t-1})dw_{t-1}}\\
&= \frac{\int_\tau^\infty p(w_t|w_{t-1},\varepsilon_{t-1})p(w_{t-1})dw_{t-1}}{\int_\tau^\infty p(w_{t-1})dw_{t-1}}
\end{aligned}$$

  With a similar approach, we may derive

$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\int_{-\infty}^\tau p(w_t|w_{t-1},\varepsilon_{t-1})p(w_{t-1})dw_{t-1}}{\int_{-\infty}^\tau p(w_{t-1})dw_{t-1}}$$

  Since $w_{t-1} =_d \mathbb{N}(0,t-1)$, it follows that

$$p(w_{t-1}) = \frac{1}{\sqrt{2\pi}\sqrt{t-1}}\exp\left(-\frac{w_{t-1}^2}{2(t-1)}\right).$$

With a simple multiplication, we may get

$$p(w_t|w_{t-1},\varepsilon_{t-1})p(w_{t-1}) = \frac{1}{2\pi\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}\sqrt{t-1}} \times \exp\left[\underbrace{-\frac{1}{2}\left(\frac{w_t - w_{t-1} - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2 - \frac{w_{t-1}^2}{2(t-1)}}_{-\frac{1}{2}C}\right].$$

Let us simply C as follow

$$\begin{aligned}
C &= \left(\frac{w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2 + \frac{w_{t-1}^2}{1-\rho^2\pi'\Omega^{-1}\pi} - \frac{2(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})w_{t-1}}{1-\rho^2\pi'\Omega^{-1}\pi} + \frac{w_{t-1}^2}{t-1} \\
&= \left(\frac{w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2 + \frac{(t-\rho^2\pi'\Omega^{-1}\pi)w_{t-1}^2}{(t-1)(1-\rho^2\pi'\Omega^{-1}\pi)} - \frac{2(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})w_{t-1}}{1-\rho^2\pi'\Omega^{-1}\pi} \\
&= \left(\sqrt{\frac{t-\rho^2\pi'\Omega^{-1}\pi}{(t-1)(1-\rho^2\pi'\Omega^{-1}\pi)}}w_{t-1} - \sqrt{\frac{t-1}{(1-\rho^2\pi'\Omega^{-1}\pi)(t-\rho^2\pi'\Omega^{-1}\pi)}}(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)^2 \\
&+ \left(\frac{w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}}\right)^2 - \frac{t-1}{(1-\rho^2\pi'\Omega^{-1}\pi)(t-\rho^2\pi'\Omega^{-1}\pi)}(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})^2 \\
&= \left(\sqrt{\frac{t-\rho^2\pi'\Omega^{-1}\pi}{(t-1)(1-\rho^2\pi'\Omega^{-1}\pi)}}w_{t-1} - \sqrt{\frac{t-1}{(1-\rho^2\pi'\Omega^{-1}\pi)(t-\rho^2\pi'\Omega^{-1}\pi)}}(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})\right)^2 \\
&+ \frac{(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})^2}{t-\rho^2\pi'\Omega^{-1}\pi}.
\end{aligned}$$

By substituting the expression we got for C, we can get

$$\begin{aligned}
p(w_t|w_{t-1},\varepsilon_{t-1})p(w_{t-1}) &= \frac{1}{2\pi\sqrt{1-\rho^2\pi'\Omega^{-1}\pi}\sqrt{t-1}} \\
&\times \underbrace{\exp\left[-\frac{1}{2}\left(\sqrt{\frac{t-\rho^2\pi'\Omega^{-1}\pi}{(t-1)(1-\rho^2\pi'\Omega^{-1}\pi)}}\left(w_{t-1} - \frac{(t-1)(w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1})}{t-\rho^2\pi'\Omega^{-1}\pi}\right)\right)^2\right]}_{K} \\
&\times \underbrace{\exp\left[-\frac{1}{2}\left(\frac{w_t - \rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\sqrt{t-\rho^2\pi'\Omega^{-1}\pi}}\right)^2\right]}_{S}.
\end{aligned}$$

We may write

$$\begin{aligned} p(w_t|w_{t-1},\varepsilon_{t-1})p(w_{t-1}) &= \frac{\sqrt{t-\rho^2\pi'\Omega^{-1}\pi}}{\sqrt{2\pi}\sqrt{(1-\rho^2\pi'\Omega^{-1}\pi)(t-1)}}K \\ &\times \frac{1}{\sqrt{2\pi}\sqrt{t-\rho^2\pi'\Omega^{-1}\pi}}S \\ &= \bar{K}\times\bar{S}. \end{aligned}$$

Note that

$$\bar{S} = \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, t-\rho^2\pi'\Omega^{-1}\pi\right).$$

It follows that

$$\begin{aligned} p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) &= \frac{\int_\tau^\infty \bar{K}\times\bar{S}dw_{t-1}}{\int_\tau^\infty p(w_{t-1})dw_{t-1}} \\ &= \frac{\int_\tau^\infty \bar{K}dw_{t-1}}{\int_\tau^\infty p(w_{t-1})dw_{t-1}}\bar{S} \\ &= \frac{1-\int_{-\infty}^\tau \bar{K}dw_{t-1}}{\int_\tau^\infty p(w_{t-1})dw_{t-1}}\bar{S} \\ &= \frac{1-\Phi\left(\sqrt{\frac{t-\rho^2\pi'\Omega^{-1}\pi}{(t-1)(1-\rho^2\pi'\Omega^{-1}\pi)}}\left(\tau-\frac{(t-1)(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{t-\rho^2\pi'\Omega^{-1}\pi}\right)\right)}{1-\Phi(\tau/\sqrt{t-1})} \\ &\times \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, t-\rho^2\pi'\Omega^{-1}\pi\right). \end{aligned}$$

Similarly, we may derive

$$\begin{aligned} p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) &= \frac{\Phi\left(\sqrt{\frac{t-\rho^2\pi'\Omega^{-1}\pi}{(t-1)(1-\rho^2\pi'\Omega^{-1}\pi)}}\left(\tau-\frac{(t-1)(w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1})}{t-\rho^2\pi'\Omega^{-1}\pi}\right)\right)}{\Phi(\tau/\sqrt{t-1})} \\ &\times \mathbb{N}\left(\rho\pi'\Omega^{-1}\varepsilon_{t-1}, t-\rho^2\pi'\Omega^{-1}\pi\right). \end{aligned}$$

- ω_ρ **of** (w_t) **when** $0<\lambda<1$ **and** $|\rho|=1$**:**
  We have

$$\begin{aligned} p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) &= p(w_t|s_{t-1}=1,y_{t-1},...,y_1) \\ &= p(w_t|w_{t-1}\geq\tau,\varepsilon_{t-1}) \\ &= p(\lambda w_{t-1}+v_t|w_{t-1}\geq\tau,\mathcal{F}_{t-1}). \end{aligned}$$

Note that

$$p(w_{t-1}|w_{t-1} \geq \tau) = \frac{\sqrt{1-\lambda^2}\varphi(w_{t-1}\sqrt{1-\lambda^2})}{1-\Phi(\tau\sqrt{1-\lambda^2})} 1\{w_{t-1} \geq \tau\}.$$

Lemma: Given probability density function, $p_X(x)$, the probability density function, $p_Y(y)$, for $Y = \alpha + \beta X$ with $\beta \neq 0$ is given by

$$p_Y(y) = \frac{1}{|\beta|} p_X\left(x = \frac{y-\alpha}{\beta}\right).$$

Since $w_t = \lambda w_{t-1} + \rho\pi'\Omega^{-1}\varepsilon_{t-1}$, by choosing $\alpha = \rho\pi'\Omega^{-1}\varepsilon_{t-1}$ and $\beta = \lambda$, we may derive

$$p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) = \frac{\frac{\sqrt{1-\lambda^2}}{\lambda}\varphi(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\lambda}\sqrt{1-\lambda^2})}{1-\Phi(\tau\sqrt{1-\lambda^2})} 1\{w_t \geq \lambda\tau + \rho\pi'\Omega^{-1}\varepsilon_{t-1}\}$$
$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\frac{\sqrt{1-\lambda^2}}{\lambda}\varphi(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\lambda}\sqrt{1-\lambda^2})}{\Phi(\tau\sqrt{1-\lambda^2})} 1\{w_t < \lambda\tau + \rho\pi'\Omega^{-1}\varepsilon_{t-1}\}.$$

Similarly, if we let $-1 < \lambda < 0$, then the conditional transition density of w_t can be obtained using

$$p(w_t|s_{t-1}=1,\mathcal{F}_{t-1}) = \frac{\frac{\sqrt{1-\lambda^2}}{\lambda}\varphi(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\lambda}\sqrt{1-\lambda^2})}{1-\Phi(\tau\sqrt{1-\lambda^2})} 1\{w_t \leq \lambda\tau + \rho\pi'\Omega^{-1}\varepsilon_{t-1}\}$$
$$p(w_t|s_{t-1}=0,\mathcal{F}_{t-1}) = \frac{\frac{\sqrt{1-\lambda^2}}{\lambda}\varphi(\frac{w_t-\rho\pi'\Omega^{-1}\varepsilon_{t-1}}{\lambda}\sqrt{1-\lambda^2})}{\Phi(\tau\sqrt{1-\lambda^2})} 1\{w_t > \lambda\tau + \rho\pi'\Omega^{-1}\varepsilon_{t-1}\}.$$

The proof for the case of $\lambda = 1$ is very similar to what we have done for $0 < \lambda < 1$, except that we have $w_{t-1}/\sqrt{t-1} =_d \mathbb{N}(0,1)$ for $t \geq 2$ in this case instead of $w_{t-1}\sqrt{1-\lambda^2} =_d \mathbb{N}(0,1)$ when $|\lambda| < 1$. □

C Additional Figures and Tables

Table III: Estimated parameters of the regime-dependent CAPM (3 Portfolios Sorted on BE/ME)

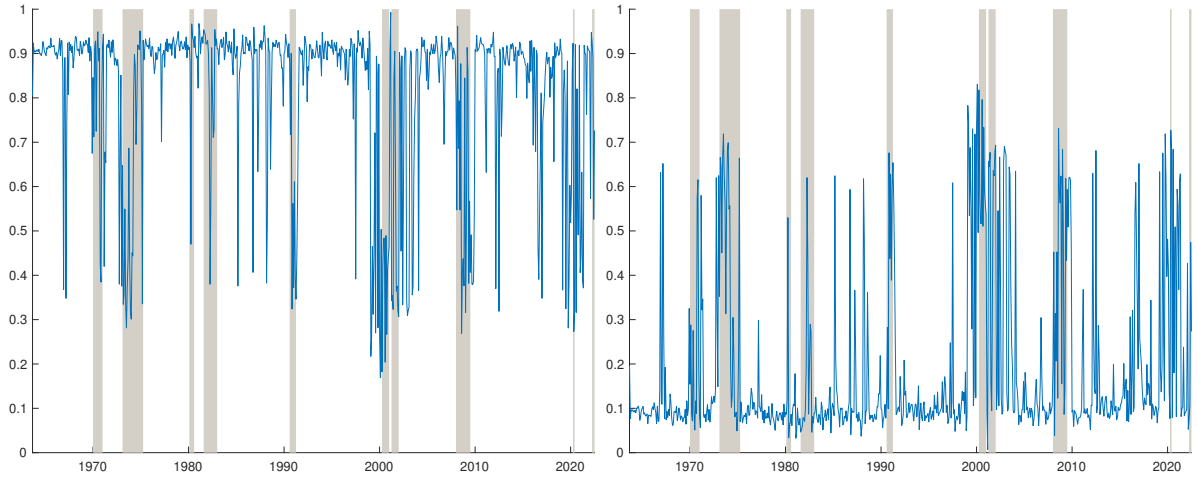
		Low	2	High
Low Volatility	α	0.3746***	0.3763***	0.5629***
		(0.504)	(0.4957)	(0.3671)
	β	1.0261***	0.958***	0.9689***
		(0.1798)	(0.1231)	(0.1459)
	π	0***	0.4622***	0.5768***
		(0.8727)	(0.799)	(0.7993)
High Volatility	α	0.2272	0.5604	0.9046
		(0.8453)	(0.602)	(0.9884)
	β	1.0049***	0.838***	1.0302***
		(0.1937)	(0.1415)	(0.2524)
	π	0***	2.3943***	3.2998***
		(0.8727)	(2.0942)	(2.8033)
	σ	1.0522***	0.8375***	1.7862***
		(0.1836)	(0.2776)	(0.9524)
	ρ		-0.6727	
			(0.8666)	
	λ		0.9977**	
			(0.3554)	
	τ		11.0506***	
			(0.6395)	

Notes: The standard errors are calculated using bootstrap method and reported in parenthesis. Significance is computed using bootstrap confidence intervals. *p<0.1; **p<0.05; ***p<0.01

Table IV: Estimated parameters for the 2 factor model (3 Portfolios Sorted on BE/ME)

		Low	2	High
High Volatility	α	0.3694***	0.4949***	1.0761
		(1.5415)	(1.7451)	(0.6655)
	β	1.1629***	0.8868***	0.7912***
		(1.0691)	(0.5067)	(0.9238)
	h	0.0251***	0.5892***	0.427***
		(0.7879)	(0.1736)	(0.7845)
	π	0.0651***	0***	1.8824***
		(0.4660)	(0.3184)	(0.4151)
Low Volatility	α	0.4482***	0.3101***	0.3141***
		(0.1892)	(0.1796)	(0.1383)
	β	0.9786***	0.9877***	1.1126***
		(0.8000)	(0.7421)	(1.2156)
	h	-0.3102***	0.2363***	0.7505***
		(0.8583)	(0.2052)	(0.1775)
	π	0.0651***	0***	0.9801***
		(0.5268)	(0.5489)	(3263)
	σ	0.6696***	0.9853***	0.0005***
		(0.2290)	(0.2126)	(0.5488)
	ρ		-0.7450***	
			(0.5872)	
	λ		0.9837***	
			(0.4719)	
	τ		-8.8408***	
			(1.2369)	

Notes: The standard errors are calculated using bootstrap method and reported in parenthesis. Significance is computed using bootstrap confidence intervals. *p<0.1; **p<0.05; ***p<0.01



(a) Low Volatility Regime (b) High Volatility Regime

Figure VIII: Smoothed high and low State Probabilities. *Notes*: This figure presents the time series of the probabilities of being in the high and low volatility regimes (solid blue line) along with the NBER recession periods (grey shaded area). The left panel plots the low volatility probability series and the right panel plots the high volatility probability series obtained from the endogenous volatility switching model