PIER

PUEY UNGPHAKORN INSTITUTE
FOR ECONOMIC RESEARCH

An Estimation of Regime Switching Models with Nonlinear Endogenous Switching

by

Chotipong Charoensom

February 2024
Discussion Paper
No. 217

An estimation of regime switching models with nonlinear endogenous switching*

Chotipong Charoensom†

February 26, 2024

Abstract

This paper proposes an approach to develop regime switching models where latent process determining the switching is endogenously controlled by the model shocks with free functional forms. The linear endogeneity assumption in the conventional endogenous regime switching models can therefore be relaxed. A recursive filter technique is applied to proceed maximum likelihood estimation in order to estimate the model parameters. A nonlinear endogenous two-regime switching mean-volatility model is conducted in numerical examples to investigate the model performance. In the examples, the endogeneity in switching allows heterogeneous effects of the shock signs (asymmetric endogeneity) and of the states being before the switching determination (state-dependent endogeneity). Monte Carlo simulations show that the conventional switching model ignoring the nonlinear endogeneity leads to the volatility biases. The estimates tend to be over or under their true value depending on how the endogeneity characteristics are. In particular, the true model that accounts the nonlinear endogeneity effectively provides the more precise estimates. The same model is also applied to real data of excess returns on US stock market, and the estimation results informatively describe the effects influencing the regime shifts.

Keywords: Nonlinear endogeneity, Regime switching, Maximum likelihood estimation, Asymmetric endogeneity, State-dependent endogeneity.
JEL: C13, C32.

*The opinions, findings, conclusions or recommendations expressed in this paper are of the author and do not necessarily reflect those of the Bank of Thailand.

†Bank of Thailand, Bangkok, 10200, Thailand. Email: chotipoc@bot.or.th.

1 Introduction

Regime switching models have been widely popular for various applications in time series analysis. This approach can capture the behavior, such as in financial markets and economic cycles, suddenly changed and possibly persisted. Hamilton (1989), a prominent article, introduces an autoregressive model with discrete state of drift rate determined by a first-order Markov process. This approach is extended by Kim (1994) in a general state-space model associated with state-varying parameters. From this starting point, there is growing literature in studying available approaches to model the regime changes as well as the treatments in finance and economics theories. The discussions of which can be referred to, e.g., Ang and Timmermann (2012), and Hamilton (2010,2016).

Most studies employ regime switching models with exogenous transition probabilities independent of realizations of underlying time series. For example, Hamilton (1989) considers the fixed transition probabilities, and Diebold Lee and Weinbach (1994), Filardo and Gordon (1998), and Bazzi et al. (2017) deal with the time-varying transition probabilities explained by exogenous variables. In practice, however, the transition probabilities could be determined endogenously. In other words, the realized shocks unexplainable in underlying model trigger the state transitions of the model parameters in order to adjust the shocks, and thus it is endogeneity. This could also be viewed as the concept of equilibrium correction that incorporates the feedback effects of the shocks into the model. Another familiar phenomenon is that financial and economic crises are usually unpredictable, so, when the crises occur, the model shocks could lead to abruptly shift the states of economy as well as the model parameters. This indicates the endogeneity in the regime switching model that needs to be considered.

In earlier works of endogenous regime switching models, Kim Piger and Startz (2008) develop an estimation of endogenous Markov-switching regime where the regime shifts are controlled by the latent state variable in a probit specification. The errors, of the underlying process and of the latent process, are assumed to be jointly correlated so that the endogenous regime switching is accounted. Kang (2014) also extends this method to a general state-space model. Chib and Dueker (2004) propose a non-Markovian regime switching model in which a latent process determining the regime follows an autoregressive process. The size of autoregressive parameter represents the strength of states dependency. The endogeneity is considered in the same manner as Kim Piger and Startz (2008). Chang Choi and Park (2017) also propose an approach similar to Chib and Dueker (2004) except that the underlying model error is correlated with the error in the next period embedded in the autoregressive latent process. Chib and Dueker (2004) develop model estimation based on Bayesian procedure, whereas frequentist approach is proposed by Chang Choi and Park (2017). Extensions and applications of this approach are referred to Cheng Gao and Yan (2018), Song Ryu and Webb (2018), Chang Tan and Wei (2018), Chang Maih and Tan (2021).

Those studies model the endogeneity based on the correlation structure in joint-normal distribution, which is linear. However, there could be some situations that financial shocks are endogenous to drive states of economy in the nonlinear manners. For example, Veldkamp (2005) explains slow boom and sudden crash in asset prices due to investors' adjustment to the prices with asymmetric information uncertainties during good time and bad time. There are confirmations in several empirical works, e.g. Nelson (1991) Engle and NG (1993) Bekaert and Wu (2000) Ghysels Clara and Volkanov (2005) Ghysels Guerin and Marcellino (2014), that return volatility in equity markets asymmetrically respond to positive and negative return shocks. Zhang et al. (2022) also documents in crude oil market that impact of negative returns on volatility is more than that of positive returns. Therefore, in the context of endogenous regime switching model, the characteristics of the endogeneity could play an important role to model the regime changes. For instance, the volatility states in financial markets are changed due to the asymmetric return shocks.

To capture the nonlinear endogeneity, this study proposes an approach to model the nonlinear endogenous switching where the endogeneity effects are considered in terms of flexible functional form, which can be nonlinear and state-dependent. The model specification and estimation method are extended of Kim Piger and Startz (2008) to a more general version. A filtering technique, first proposed by Hamilton (1989), is modified in this study to construct a recursive maximum likelihood estimation for the specified models. A two-regime switching model is first to introduce the idea, and then is generalized to a multi-regime switching multivariate model. We also discuss a special case of the nonlinear endogeneity using a piecewise linear function for the two-regime model.

To evaluate the model performance, the simulations are explored based on a simple two-regime switching mean-volatility model with the nonlinear endogenous switching that consists of the asymmetric and state-dependent effects. That is, the regime changes are described by the signs of the model shock and the state being before the changes. The results of the simulations show that the typical linear endogenous switching model provides the estimated mean of each state approximately indifferent from its true value but the estimated volatility quite biased. The biases are either over or under the true values depending on the patterns of the asymmetric and state-dependent endogeneity. Particularly, we obtain the more precise estimates if this nonlinear endogeneity is allowed into the model. In the same model, it is applied to the excess returns data of the US stock market, and we find that the state transitions strongly rely on the contemporaneous positive shocks as well as the shocks associated with the previous high-volatility state.

This paper outlines the discussions into the rest five sections. Section 2 presents the nonlinear endogenous switching models for double regimes. A generalization to the multivariate version is in Section 3. Section 4 and 5 illustrate the numerical examples of the simulations and real data, respectively. This paper concludes in Section 6.

2 A two-regime switching model with nonlinear endogenous switching

This section discusses specification and estimation of two-regime switching models where the regime switching is nonlinearly endogenous. The notations for the following discussions are introduced. Let $p(\cdot)$ and $p(\cdot|\cdot)$ denote the generic notation for the unconditional and conditional probability distribution functions, respectively. In addition, denote the probability density function and the cumulative distribution function of the standard normal random variable by $\phi(\cdot)$ and $\Phi(\cdot)$. Those functions are respectively defined by $\phi_2(\cdot,\cdot;\rho)$ and $\Phi_2(\cdot,\cdot;\rho)$ to refer the bivariate standard normal distribution functions where ρ denotes the correlation parameter between the two random variables.

2.1 The model

Let x_t denote a vector of exogenous or predetermined variables which may also include their lags. Let s_t denote a discrete state process representing the regime. Consider an underlying time series y_t specified by

$$y_t = \mu_t + \sigma_t \epsilon_t, \tag{1}$$

where $\epsilon_t \sim$ i.i.d. $\mathcal{N}(0,1)$, and μ_t and σ_t are defined by

$$\mu_t = \mu(s_t, ..., s_{t-k}, y_{t-1}, ..., y_{t-k}, x_t), \tag{2}$$

$$\sigma_t = \sigma(s_t, ..., s_{t-k}, y_{t-1}, ..., y_{t-k}, x_t). \tag{3}$$

In this case, μ_t and σ_t represent the time-varying mean and volatility functions, respectively, which can be nonlinear in the input variables. This specification is extended of Kim Piger and Startz (2008) and is similar to Chang Choi and Park (2017) that y_t may depend on its lagged values and the past regimes for k periods.

For the state process s_t, let us consider the double regimes that $s_t \in \{0,1\}$ is determined by a continuous latent state process w_t. Specifically,

$$s_t = 1_{\{w_t>0\}}, \tag{4}$$

where $1_{\{\cdot\}}$ is an indicator function, and w_t follows that

$$w_t = \nu_t + \gamma_t \eta_t, \tag{5}$$

where $\eta_t \sim$ i.i.d. $\mathcal{N}(0,1)$, and ν_t and γ_t are defined by

$$\nu_t = \nu(s_{t-1}, ..., s_{t-h}, \epsilon_t, ..., \epsilon_{t-h}, z_t), \tag{6}$$

$$\gamma_t = \gamma(s_{t-1}, ..., s_{t-h}, \epsilon_t, ..., \epsilon_{t-h}, z_t). \tag{7}$$

Similarly, ν_t and γ_t respectively represent the mean and volatility functions for controlling the latent variable and thus regime shifts. z_t is a vector that consists of exogenous variables which also help the model determine the regime.[1] It can be seen that the contemporaneous shock ϵ_t and the historical shocks $(\epsilon_{t-1}, ..., \epsilon_{t-h})$ of the underlying model can endogenously determine the state variable s_t through the functions ν_t and γ_t. Apparently, since s_t may depends on s_{t-j} for $j = 1, ..., k+h$, it may not be a first-order Markov process that usually applies with traditional Markov-switching models. The functional form of ν_t and γ_t are also free and allowed to be nonlinear which can capture various relationships or patterns of the endogeneity in regime switching. This structure is more flexible than previous studies that only consider the linear endogeneity based on Gaussian approach. That is similar to setting that the random variables ϵ_t and η_t are jointly normal, and the functions ν_t and γ_t are independent of ϵ_t and its lags. Lastly, all functions above are assumed to be well defined, and σ_t and γ_t are strictly positive.

2.2 Maximum likelihood estimation

The estimation for the model parameters in (1) and (5) is discussed in this subsection. Let $\mathcal{F}_t = [y_t, y_{t-1}, ..., y_1]'$ and $\Omega_t = [x_t', x_{t-1}', ..., x_1', z_t', z_{t-1}' ..., z_1']'$ denote all historical data observed up to time t of underlying time series (y) and exogenous variables (x, z). Let θ be a vector of model parameters associated with (1) and (5). For a number of observation T, the log-likelihood function can be written as

$$\ln L(\theta) = \sum_{t=1}^{T} \ln p(y_t|\mathcal{F}_{t-1}; \Omega_t, \theta). \tag{8}$$

$\hat{\theta}$ is then chosen to maximize the log-likelihood function, i.e.

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}} \ln L(\theta). \tag{9}$$

In order to compute $p(y_t|\mathcal{F}_{t-1}; \Omega_t, \theta)$ for each $t = 1, ..., T$, The recursive filter method proposed by Hamilton (1989) can be applied to develop the recursive estimation based on the specified models. Utilizing the Bayes' rule, it can be written that

$$p(y_t|\mathcal{F}_{t-1}; \Omega_t, \theta) = \sum_{s_t=0}^{1} \cdots \sum_{s_{t-k-h}=0}^{1} p(y_t|s_t, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) p(s_t, ..., s_{t-k-h}|\mathcal{F}_{t-1}; \Omega_t, \theta), \tag{10}$$

where

$$p(s_t, ..., s_{t-k-h}|\mathcal{F}_{t-1}; \Omega_t, \theta) = p(s_t|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) p(s_{t-1}, ..., s_{t-k-h}|\mathcal{F}_{t-1}; \Omega_{t-1}, \theta), \tag{11}$$

[1] z_t may include lags of y_t up to order $k+h$ if the equality $p(s_t|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = p(s_t|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)$, as shown in Proposition 2.1, still holds for the estimation.

$$p(s_t, ..., s_{t-k-h+1}|\mathcal{F}_t; \Omega_t, \theta) = \sum_{s_{t-k-h}=0}^{1} p(s_t, ..., s_{t-k-h}|y_t, \mathcal{F}_{t-1}; \Omega_t, \theta)$$
$$= \frac{\sum_{s_{t-k-h}=0}^{1} p(y_t|s_t, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) p(s_t, ..., s_{t-k-h}|\mathcal{F}_{t-1}; \Omega_t, \theta)}{p(y_t|\mathcal{F}_{t-1}; \Omega_t, \theta)}. \tag{12}$$

The recursion (10)-(12) is analogous to a conventional Kalman filter in which (11) and (12) could be steps of prediction and updating, respectively.

To initialize the recursion (10)-(12), it may start at $t = k + h$ and need the estimates of $p(s_{k+h}, ..., s_1|\mathcal{F}_{k+h}; \Omega_{k+h}, \theta)$. There would be a number of $2^{k+h} - 1$ estimates. They may be approximated by unconditional or stationary distribution. On the other hand, they may be assigned as additional parameters to be estimated. Applying the following results in Proposition 2.1 to compute the probability distribution functions $p(s_t|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta)$ and $p(y_t|s_t, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta)$, the recursion may be readily completed. Then, the maximum likelihood estimates could be obtained by working with numerical optimization.

Proposition 2.1. *Suppose a joint stochastic process (s_t, y_t) follows (1) and (4). Then, the following results hold.*

$$p(s_t|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = \int_{-\infty}^{\infty} \Big[(1 - s_t)\Big(1 - \Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big) + s_t\Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big]\phi(\epsilon_t)d\epsilon_t, \tag{13}$$

$$p(y_t|s_t, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = \frac{\phi\Big(\frac{y_t - \mu_t}{\sigma_t}\Big)\Big[(1 - s_t)\Big(1 - \Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big) + s_t\Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big]}{\sigma_t p(s_t|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta)}. \tag{14}$$

In addition, (s_t, y_t) is a first $(k + h)$-order Markov process associated with the transition density

$$p(s_t, y_t|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta) = \frac{1}{\sigma_t}\phi\Big(\frac{y_t - \mu_t}{\sigma_t}\Big)\Big[(1 - s_t)\Big(1 - \Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big) + s_t\Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big]. \tag{15}$$

Proof. Consider $y_t = \mu(s_t, ..., s_{t-k}, y_{t-1}, ..., y_{t-k}, x_t) + \sigma(s_t, ..., s_{t-k}, y_{t-1}, ..., y_{t-k}, x_t)\epsilon_t \equiv \mu_t + \sigma_t\epsilon_t$ and $w_t = \nu(s_{t-1}, ..., s_{t-h}, \epsilon_t, ..., \epsilon_{t-h}, z_t) + \gamma(s_{t-1}, ..., s_{t-h}, \epsilon_t, ..., \epsilon_{t-h}, z_t)\eta_t \equiv \nu_t + \gamma_t\eta_t$ where ϵ_t and η_t are independent standard normal random variables. For $s_t = 0$, it can be shown that

$$\Pr(s_t = 0|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = \Pr(w_t \leq 0|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)$$
$$= \Pr(w_t \leq 0|s_{t-1}, ..., s_{t-k-h}, \epsilon_{t-1}, ..., \epsilon_{t-h}; \Omega_t, \theta)$$
$$= \int_{-\infty}^{\infty} \Pr\Big(\eta_t \leq -\frac{\nu_t}{\gamma_t}|s_{t-1}, ..., s_{t-k-h}, \epsilon_t, \epsilon_{t-1}, ..., \epsilon_{t-h}; \Omega_t, \theta\Big)p(\epsilon_t)d\epsilon_t$$
$$= \int_{-\infty}^{\infty} \Big(1 - \Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big)\phi(\epsilon_t)d\epsilon_t.$$

Similarly, for $s_t = 1$,

$$\Pr(s_t = 1|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = \int_{-\infty}^{\infty} \Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\phi(\epsilon_t)d\epsilon_t.$$

The transition probability function (13) is already obtained. To derive (14), we first consider the case $s_t = 0$ that

$$\begin{aligned}
&\Pr(y_t \le y|s_t = 0, s_{t-1},..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta)\\
&\qquad = \frac{\Pr(y_t \le y, w_t \le 0|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}, ; \Omega_t, \theta)}{\Pr(w_t \le 0|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)}\\
&\qquad = \frac{\displaystyle\int_{-\infty}^{y} \Pr(w_t \le 0|s_{t-1}, ..., s_{t-k-h}, y_t, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)\phi\Big(\frac{y_t - \mu_t}{\sigma_t}\Big)dy_t}{\sigma_t \Pr(w_t \le 0|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)}\\
&\qquad = \frac{\displaystyle\int_{-\infty}^{y} \Big(1 - \Phi\Big(\frac{\nu_t}{\gamma_t}; \epsilon_t = \frac{y_t - \mu_t}{\sigma_t}\Big)\Big)\phi\Big(\frac{y_t - \mu_t}{\sigma_t}\Big)dy_t}{\sigma_t \displaystyle\int_{-\infty}^{\infty} \Big(1 - \Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big)\phi(\epsilon_t)d\epsilon_t}.
\end{aligned}$$

The last equality above holds by the fact that the information $(s_{t-1}, ..., s_{t-k-h}, y_t, y_{t-1}, ..., y_{t-k-h})$ is enough to deduce $(s_{t-1}, ..., s_{t-k-h}, \epsilon_t, \epsilon_{t-1}, ..., \epsilon_{t-h})$ where $\epsilon_{t-j} = (y_{t-j} - \mu_{t-j})/\sigma_{t-j}$, $j = 0, ..., h$. Therefore, we have the conditional density function of y_t evaluated at any value of y given by

$$\begin{aligned}
p(y_t = y|s_t = 0, s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) &= \frac{\partial}{\partial y}\Pr(y_t \le y|s_t = 0, s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta)\\
&= \frac{\Big(1 - \Phi\Big(\frac{\nu_t}{\gamma_t}; \epsilon_t = \frac{y - \mu_t}{\sigma_t}\Big)\Big)\phi\Big(\frac{y - \mu_t}{\sigma_t}\Big)}{\sigma_t \displaystyle\int_{-\infty}^{\infty} \Big(1 - \Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\Big)\phi(\epsilon_t)d\epsilon_t}.
\end{aligned}$$

Similarly, it can be accomplished for $s_t = 1$ where

$$p(y_t|s_t = 1, s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = \frac{\Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\phi\Big(\frac{y_t - \mu_t}{\sigma_t}\Big)}{\sigma_t \displaystyle\int_{-\infty}^{\infty} \Phi\Big(\frac{\nu_t}{\gamma_t}\Big)\phi(\epsilon_t)d\epsilon_t}.$$

Now (14) is readily obtained. Lastly, it can be seen from above that $p(s_t|s_{t-1}, ..., s_1, y_{t-1}, ..., y_1; \Omega_t, \theta)$ $= p(s_t|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)$ and $p(y_t|s_t, s_{t-1}, ..., s_1, y_{t-1}, ..., y_1; \Omega_t, \theta) = p(y_t|s_t, s_{t-1},$ $..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)$. It follows that (s_t, y_t) is a first $(k + h)$-order Markov process where

$$\begin{aligned}
&p(s_t, y_t|s_{t-1}, ..., s_1, y_{t-1}, ..., y_1; \Omega_t, \theta)\\
&\quad = p(y_t|s_t, s_{t-1}, ..., s_1, y_{t-1}, ..., y_1; \Omega_t, \theta)p(s_t|s_{t-1}, ..., s_1, y_{t-1}, ..., y_1; \Omega_t, \theta)\\
&\quad = p(y_t|s_t, s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)p(s_t|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)\\
&\quad = p(s_t, y_t|s_{t-1}, ..., s_{t-k-h}, y_{t-1}, ..., y_{t-k-h}; \Omega_t, \theta)
\end{aligned}$$

and we obtain the transition density (15). □

From the above results, the transition probability (13) is evaluated with the expected value of cumulative standard normal distribution function on the contemporaneous endogeneity effects of ϵ_t. The conditional density of y_t shown in (14) is similar to a class of skew normal density as referred by Kim Piger and Startz (2008). As modeling the endogeneity by the linear correlation structure that proposed by Kim Piger and Startz (2008), the integration in (13) averages ϵ_t out, and it can be written in terms of normal cumulative distribution function. This will be shown as a special case in the next subsection. For the nonlinear endogeneity associated with complicated correlation structures of ϵ_t on ν_t and γ_t, however, the formula for (13) may not be explicit, and we need to approximate it using numerical integration.[2] On the other hand, if ν_t and γ_t are independent of the contemporaneous shock ϵ_t, it is easy to see that the transition probability (13) becomes $(1-s_t)(1-\Phi(\nu_t/\gamma_t)) + s_t\Phi(\nu_t/\gamma_t)$, and the conditional density (14) becomes $\phi\big((y_t-\mu_t)/\sigma_t\big)/\sigma_t$. The model without the contemporaneous endogeneity then simplifies the recursion. Furthermore, we also have that (s_t, y_t) can be filtered from the information last $k+h$ periods, because, in the underlying model, the past k periods of the state s_t are determined by the past information at most h periods additional from that each of k periods.

2.3 Endogenous regime switching with piecewise linear functions

In this subsection, simple model specifications of the latent process (5) are discussed. Consider the typical linear endogenous regime switching model based on Gaussian approach that the latent process follows:

$$w_t = \alpha(s_{t-1}, z_t) + \rho\epsilon_t + \sqrt{1-\rho^2}\eta_t, \tag{16}$$

where $-1 < \rho < 1$ and

$$\begin{bmatrix} \epsilon_t \\ \rho\epsilon_t + \sqrt{1-\rho^2}\eta_t \end{bmatrix} \sim \text{i.i.d. } \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right). \tag{17}$$

In this case, we specify $\nu_t = \alpha(s_{t-1}, z_t) + \rho\epsilon_t$ and $\gamma_t = \sqrt{1-\rho^2}$, where $\alpha(s_{t-1}, z_t)$ is time-varying mean rate as a function of the previous state and exogenous variables. It can be seen that $s_t = 1_{\{w_t>0\}}$ follows a first order Markov process. This endogeneity formulation assumes that the shock of underlying model (ϵ_t) is linearly and contemporaneously correlated with the shock of the latent model ($\rho\epsilon_t + \sqrt{1-\rho^2}\eta_t$) with ρ level. Similar analyses can be found in Chib and Dueker (2004), Kim Piger and Startz (2008), Kang (2014), and others regarding applications applied with linear endogenous switching models.

[2]The recursive estimation associated with the numerical integration may encounter computational time problem. It may be simplified by the substitution method that transforms the infinite limits of the integration to be finite.

As introduced before, the nonlinear endogeneity in regime changes could be crucial and, however, is ignored in the formulation above. It can be easily extended by applying the piecewise linear functions so that we can approximate the nonlinear effects of ϵ_t on w_t. The effects between the states can also be considered in order to account state-dependent endogeneity, so we may let the correlation level be a function of previous states. Specifically, the latent equation given by (5) can be written in terms of a finite M-piecewise linear as follows:

$$w_t = \begin{cases} \alpha_{1,t} + \rho_{1,t}\epsilon_t + \sqrt{1-\rho_{1,t}^2}\eta_t & \text{if} \quad -\infty < \epsilon_t \leq \bar{\epsilon}_{1,t}, \\ \alpha_{2,t} + \rho_{2,t}\epsilon_t + \sqrt{1-\rho_{2,t}^2}\eta_t & \text{if} \quad \bar{\epsilon}_{1,t} < \epsilon_t \leq \bar{\epsilon}_{2,t}, \\ \vdots & \\ \alpha_{M,t} + \rho_{M,t}\epsilon_t + \sqrt{1-\rho_{M,t}^2}\eta_t & \text{if} \quad \bar{\epsilon}_{M-1,t} < \epsilon_t < \infty, \end{cases} \tag{18}$$

where

$$\alpha_{i,t} = \alpha_i(s_{t-1}, ..., s_{t-h}, z_t), \tag{19}$$

$$\rho_{i,t} = \rho_i(s_{t-1}, ..., s_{t-h}, z_t), \tag{20}$$

$$\bar{\epsilon}_{i,t} = \bar{\epsilon}_i(s_{t-1}, ..., s_{t-h}, z_t), \tag{21}$$

and $-1 < \rho_{i,t} < 1$ for all $i = 1, ..., M$. We assume the thresholds $\bar{\epsilon}_{i,t}$, $i = 1, ..., M-1$, satisfies $-\infty < \bar{\epsilon}_{1,t} < \bar{\epsilon}_{2,t} < ... < \bar{\epsilon}_{M-1,t} < \infty$. Also, note that ϵ_t and η_t are independent standard normal random variables. Under the above specification, the state (s_t) is endogenously determined by the contemporary shock of the underlying model (ϵ_t), and the historical shocks are assumed to have no effect on the determination. The endogeneity effects depend on the piecewise correlation $\rho_{i,t}$, which is a function of the ϵ_t levels and the previous states. In particular, the effects are different across M disjoint intervals of ϵ_t levels and identical within each of the intervals. We may make it to be more general by letting the intervals be able to change when the lagged states or exogenous variables are changed. The number of intervals, M, can be viewed as a hyperparameter in the model to be chosen. We have the following results to calculate the recursive filter (10)-(12), and therefore the model parameters can be estimated by numerically maximizing the log-likelihood function.

Proposition 2.2. *Suppose that y_t and w_t are modeled by* (1) *and* (18)*, respectively. Then, the following results hold.*

$$p(s_t | s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = (1 - s_t)(1 - \delta_t) + s_t\delta_t, \tag{22}$$

where

$$\delta_t = -\Phi_2(-\alpha_{1,t}, \bar{\epsilon}_{1,t}; \rho_{1,t}) - \sum_{i=2}^{M-1} \Big[\Phi_2(-\alpha_{i,t}, \bar{\epsilon}_{i,t}; \rho_{i,t}) - \Phi_2(-\alpha_{i,t}, \bar{\epsilon}_{i-1,t}; \rho_{i,t})\Big] + \Phi(\alpha_{M,t})$$
$$+ \Phi_2(-\alpha_{M,t}, \bar{\epsilon}_{M-1,t}; \rho_{M,t}),$$

and

$$p(y_t|s_t,...,s_{t-k-h},\mathcal{F}_{t-1};\Omega_t,\theta)=\frac{\phi\Big(\dfrac{y_t-\mu_t}{\sigma_t}\Big)}{\sigma_t p(s_t|s_{t-1},...,s_{t-k-h},\mathcal{F}_{t-1};\Omega_t,\theta)}$$

$$\times\begin{cases}(1-s_t)(1-\Phi(\zeta_{1,t}))+s_t\Phi(\zeta_{1,t}) & \text{if} \quad -\infty<\dfrac{y_t-\mu_t}{\sigma_t}\leq\bar{\epsilon}_{1,t},\\ (1-s_t)(1-\Phi(\zeta_{2,t}))+s_t\Phi(\zeta_{2,t}) & \text{if} \quad \bar{\epsilon}_{1,t}<\dfrac{y_t-\mu_t}{\sigma_t}\leq\bar{\epsilon}_{2,t},\\ \vdots & \\ (1-s_t)(1-\Phi(\zeta_{M,t}))+s_t\Phi(\zeta_{M,t}) & \text{if} \quad \bar{\epsilon}_{M-1,t}<\dfrac{y_t-\mu_t}{\sigma_t}<\infty,\end{cases} \tag{23}$$

where

$$\zeta_{i,t}=\frac{\alpha_{i,t}\sigma_t+\rho_{i,t}(y_t-\mu_t)}{\sigma_t\sqrt{1-\rho_{i,t}^2}},\quad i=1,...,M.$$

Proof. Define $\bar{\epsilon}_{0,t}=-\infty$ and $\bar{\epsilon}_{M,t}=\infty$, and that $-\infty=\bar{\epsilon}_{0,t}<\bar{\epsilon}_{1,t}<...<\bar{\epsilon}_{M-1,t}<\bar{\epsilon}_{M,t}=\infty$. (18) can be written as

$$w_t=\sum_{i=1}^{M}\Big(\alpha_{i,t}+\rho_{i,t}\epsilon_t+\sqrt{1-\rho_{i,t}^2}\eta_t\Big)1_{\{\epsilon_t\in(\bar{\epsilon}_{i-1,t},\bar{\epsilon}_{i,t}]\}}.$$

Applying (13) in Proposition 2.1 with substituting $\nu_t=\sum_{i=1}^{M}(\alpha_{i,t}+\rho_{i,t}\epsilon_t)1_{\{\epsilon_t\in(\bar{\epsilon}_{i-1,t},\bar{\epsilon}_{i,t}]\}}$ and $\gamma_t=\sum_{i=1}^{M}\sqrt{1-\rho_{i,t}^2}1_{\{\epsilon_t\in(\bar{\epsilon}_{i-1,t},\bar{\epsilon}_{i,t}]\}}$, we can derive the case $s_t=0$:

$$\begin{aligned}\Pr(s_t=0|s_{t-1},...,s_{t-k-h},\mathcal{F}_{t-1};\Omega_t,\theta)&=\int_{-\infty}^{\infty}\Phi\Big(-\frac{\nu_t}{\gamma_t}\Big)\phi(\epsilon_t)d\epsilon_t\\ &=\sum_{i=1}^{M}\int_{\bar{\epsilon}_{i-1,t}}^{\bar{\epsilon}_{i,t}}\int_{-\infty}^{-\frac{\alpha_{i,t}+\rho_{i,t}\epsilon_t}{\sqrt{1-\rho_{i,t}^2}}}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big)dx\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{\epsilon_t^2}{2}\Big)d\epsilon_t\\ &=\sum_{i=1}^{M}\int_{\bar{\epsilon}_{i-1,t}}^{\bar{\epsilon}_{i,t}}\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi(1-\rho_{i,t}^2)}}\exp\Big(-\frac{(w-\alpha_{i,t}-\rho_{i,t}\epsilon_t)^2}{2(1-\rho_{i,t}^2)}\Big)\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{\epsilon_t^2}{2}\Big)dwd\epsilon_t\\ &=\sum_{i=1}^{M}\int_{\bar{\epsilon}_{i-1,t}}^{\bar{\epsilon}_{i,t}}\int_{-\infty}^{0}\frac{1}{2\pi\sqrt{(1-\rho_{i,t}^2)}}\exp\Big(-\frac{(w-\alpha_{i,t})^2-2\rho_{i,t}\epsilon_t(w-\alpha_{i,t})+\epsilon_t^2}{2(1-\rho_{i,t}^2)}\Big)dwd\epsilon_t\\ &=\sum_{i=1}^{M}\int_{\bar{\epsilon}_{i-1,t}}^{\bar{\epsilon}_{i,t}}\int_{-\infty}^{-\alpha_{i,t}}\frac{1}{2\pi\sqrt{(1-\rho_{i,t}^2)}}\exp\Big(-\frac{\tilde{w}^2-2\rho_{i,t}\epsilon_t\tilde{w}+\epsilon_t^2}{2(1-\rho_{i,t}^2)}\Big)d\tilde{w}d\epsilon_t\\ &=\Phi_2(-\alpha_{1,t},\bar{\epsilon}_{1,t};\rho_{1,t})+\sum_{i=2}^{M-1}\Big[\Phi_2(-\alpha_{i,t},\bar{\epsilon}_{i,t};\rho_{i,t})-\Phi_2(-\alpha_{i,t},\bar{\epsilon}_{i-1,t};\rho_{i,t})\Big]+\Phi(-\alpha_{M,t})\\ &\quad-\Phi_2(-\alpha_{M,t},\bar{\epsilon}_{M-1,t};\rho_{M,t}).\end{aligned}$$

It follows that $\Pr(s_t = 1|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = 1 - \Pr(s_t = 0|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta)$. The transition probability (22) is obtained. Similarly, substituting ν_t and γ_t into (14) easily yields the conditional density function (23). $\square$

This is an example of the latent specification that we can account for the nonlinear endogeneity effects. As shown in (22), the transition probability function can be written in terms of univariate and bivariate standard normal cumulative distribution functions represented in δ_t, and we are not necessary to evaluate numerical integration during the recursive estimation as previously mentioned. The conditional density (23) is also constructed with different structures depending on the levels of $\epsilon_t = (y_t - \mu_t)/\sigma_t$. If we let ϵ_t explain the latent process with identical structure so that $\alpha_{i,t} = \alpha_t$ and $\rho_{i,t} = \rho_t$ for all $i = 1, ..., M$, the conditional probability distribution functions (22)-(23) are given by

$$p(s_t|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = (1 - s_t)(1 - \Phi(\alpha_t)) + s_t\Phi(\alpha_t), \tag{24}$$

$$p(y_t|s_t, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta) = \frac{\phi\big(\frac{y_t-\mu_t}{\sigma_t}\big)\Big[(1-s_t)\big(1-\Phi\big(\frac{\alpha_t\sigma_t+\rho_t(y_t-\mu_t)}{\sigma_t\sqrt{1-\rho_t^2}}\big)\big) + s_t\Phi\big(\frac{\alpha_t\sigma_t+\rho_t(y_t-\mu_t)}{\sigma_t\sqrt{1-\rho_t^2}}\big)\Big]}{\sigma_t p(s_t|s_{t-1}, ..., s_{t-k-h}, \mathcal{F}_{t-1}; \Omega_t, \theta)}. \tag{25}$$

In this case, the endogeneity effect via ρ_t is symmetric across the levels of the shock, but it can still be state-varying that capture nonlinear response to the shock. The Markov regime switching model with the linear endogenous switching as, for example, proposed by Kim Piger and Startz (2008) specifies $\alpha_t = \alpha(s_{t-1}, z_t)$ and especially $\rho_t = \rho$ which is fixed. Therefore, the existence of the nonlinear endogeneity effects may lead to the model estimation biases to the conventional switching models that ignore the effects.

3 A multi-regime switching multivariate model with nonlinear endogenous switching

3.1 The model

Consider a $K-$state regime switching model for a system of n equations. Let $\boldsymbol{y}_t = [y_{1,t}, ..., y_{n,t}]'$ denote the vector of underlying time series associated with the the vector of corresponding errors $\boldsymbol{\epsilon}_t = [\epsilon_{1,t}, ..., \epsilon_{n,t}]'$. Let $s_{i,t} \in \{0, ..., K-1\}$ denote the state process to represent the individual regime of $y_{i,t}$, and define $\boldsymbol{s}_t = [s_{1,t}, ..., s_{n,t}]'$. We also use x_t to denote exogenous variables the same as previous section. $y_{i,t}$ is explained by individual time-varying mean $\mu_{i,t}$ and volatility $\sigma_{i,t}$ measured by the following functions:

$$\mu_{i,t} = \mu_i(s_{i,t}, ..., s_{i,t-k}, \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}, x_t), \tag{26}$$

$$\sigma_{i,t} = \sigma_i(s_{i,t}, ..., s_{i,t-k}, \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}, x_t), \tag{27}$$

where $\sigma_{i,t}$ is strictly positive. In this multivariate version, both $\mu_{i,t}$ and $\sigma_{i,t}$ can be varied depending on the current and past states of its individual regime, and the lagged underlying time series in the system. In addition, $y_{i,t}$ is allowed to be correlated with $y_{j,t}$ via their error components associated with the correlation function

$$q_{ij,t} = q_{ij}(s_{i,t}, ..., s_{i,t-k}, s_{j,t}, ..., s_{j,t-k}, \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}, x_t), \tag{28}$$

where $-1 < q_{ij,t} = q_{ji,t} < 1$ if $i \neq j$, and $q_{ij,t} = 1$ if $i = j$. Let $\boldsymbol{Q}_t$ denote a time-varying correlation matrix whose entries are $q_{ij,t}$. We can apply a Cholesky decomposition to find a lower triangular matrix $\boldsymbol{L}_t$ whose entries are $l_{ij,t}$ so that

$$\boldsymbol{Q}_t = \boldsymbol{L}_t \boldsymbol{L}_t', \tag{29}$$

where

$$\boldsymbol{L}_t = \begin{bmatrix} l_{11,t} & 0 & \cdots & 0 \\ l_{21,t} & l_{22,t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1,t} & l_{n2,t} & \cdots & l_{nn,t} \end{bmatrix} \tag{30}$$

with

$$l_{11,t} = 1, \quad l_{ii,t} = \sqrt{1 - \sum_{k=1}^{i-1} l_{ik,t}^2} \text{ for } 1 < i \leq n, \tag{31}$$

$$l_{i1,t} = q_{i1,t} \text{ for } 1 \leq i \leq n, \quad l_{ij,t} = \frac{1}{l_{jj,t}}\Big(q_{ij,t} - \sum_{k=1}^{j-1} l_{ik,t} l_{jk,t}\Big) \text{ for } 1 < j < i \leq n. \tag{32}$$

We also assume that $\sum_{k=1}^{i-1} l_{ik,t}^2 < 1$, for $i = 2, ..., n$, to guarantee the real positive diagonal entries of $\boldsymbol{L}_t$. Therefore, we write

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{n,t} \end{bmatrix} = \begin{bmatrix} \mu_{1,t} \\ \mu_{2,t} \\ \vdots \\ \mu_{n,t} \end{bmatrix} + \begin{bmatrix} \sigma_{1,t} & 0 & \cdots & 0 \\ 0 & \sigma_{2,t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{n,t} \end{bmatrix} \begin{bmatrix} l_{11,t} & 0 & \cdots & 0 \\ l_{21,t} & l_{22,t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1,t} & l_{n2,t} & \cdots & l_{nn,t} \end{bmatrix} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \vdots \\ \epsilon_{n,t} \end{bmatrix} \tag{33}$$

where $\boldsymbol{\epsilon}_t = [\epsilon_{1,t}, ..., \epsilon_{n,t}]'$ is i.i.d. $n-$dimensional standard normal random vector. It can be seen that $\mathbb{E}[\boldsymbol{L}_t \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t' \boldsymbol{L}_t' | \boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k}, \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}, x_t] = \boldsymbol{Q}_t$, and the entries (i, j) of $\text{Cov}(\boldsymbol{y}_t | \boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k}, \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}, x_t)$ are $\sigma_{i,t}\sigma_{j,t}q_{ij,t}$.

In addition, similar to the system of the underlying processes, we characterize the system of latent processes $w_{i,t}$, $i = 1, ..., n$, that their individual mean and volatility functions are respectively given by

$$\nu_{i,t} = \nu_i(s_{i,t-1}, ..., s_{i,t-h}, \boldsymbol{\epsilon}_t, ..., \boldsymbol{\epsilon}_{t-h}, z_t), \tag{34}$$

$$\gamma_{i,t} = \gamma_i(s_{i,t-1}, ..., s_{i,t-h}, \boldsymbol{\epsilon}_t, ..., \boldsymbol{\epsilon}_{t-h}, z_t), \tag{35}$$

where $\gamma_{i,t} > 0$ and z_t is exogenous. The correlations between $w_{i,t}$ and $w_{j,t}$ via their error terms can also be accounted with the quantity

$$r_{ij,t} = r_{ij}(s_{i,t-1}, ..., s_{i,t-h}, s_{j,t-1}, ..., s_{j,t-h}, \boldsymbol{\epsilon}_t, ..., \boldsymbol{\epsilon}_{t-h}, z_t), \tag{36}$$

where $-1 < r_{ij,t} = r_{ji,t} < 1$ if $i \neq j$ and $r_{ij,t} = 1$ if $i = j$. A correlation matrix whose entries (i, j) are $r_{ij,t}$ is denoted by $\boldsymbol{R}_t$, and it is associated with the lower triangular matrix $\boldsymbol{M}_t$ whose entries are $m_{ij,t}$ so that $\boldsymbol{R}_t = \boldsymbol{M}_t\boldsymbol{M}_t'$. Thus, $m_{ij,t}$, $1 \leq j \leq i \leq n$, can be written in terms of $r_{ij,t}$ by the same recursive formulas of (31)-(32) with assuming $\sum_{k=1}^{i-1} m_{ik,t}^2 < 1$ for $i = 2, ..., n$. Then, we model the system of the latent processes as

$$\begin{bmatrix} w_{1,t} \\ w_{2,t} \\ \vdots \\ w_{n,t} \end{bmatrix} = \begin{bmatrix} \nu_{1,t} \\ \nu_{2,t} \\ \vdots \\ \nu_{n,t} \end{bmatrix} + \begin{bmatrix} \gamma_{1,t} & 0 & \cdots & 0 \\ 0 & \gamma_{2,t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{n,t} \end{bmatrix} \begin{bmatrix} m_{11,t} & 0 & \cdots & 0 \\ m_{21,t} & m_{22,t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1,t} & m_{n2,t} & \cdots & m_{nn,t} \end{bmatrix} \begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \\ \vdots \\ \eta_{n,t} \end{bmatrix}, \tag{37}$$

where $[\eta_{1,t}, ..., \eta_{n,t}]'$ is i.i.d. $n-$dimensional standard normal random vector. Let $[\tau_0, \tau_1, ..., \tau_{K-1}, \tau_K]'$ denote the fixed thresholds, where $\tau_0 = -\infty$, $\tau_K = \infty$, and $-\infty = \tau_0 < \tau_1 < ... < \tau_{K-1} < \tau_K = \infty$. The value of $s_{i,t}$ is determined by which the threshold interval taken by the latent factor $w_{i,t}$, as follows:

$$s_{i,t} = \begin{cases} 0 & \text{if} \quad \tau_0 < w_{i,t} \leq \tau_1, \\ 1 & \text{if} \quad \tau_1 < w_{i,t} \leq \tau_2, \\ \vdots & \\ K-1 & \text{if} \quad \tau_{K-1} < w_{i,t} < \tau_K, \qquad i = 1, ..., n. \end{cases} \tag{38}$$

From the construction above, the nonlinear endogenous switching on the system of regimes is described by the current and past shocks from underlying equations that control the latent variables. It is important to note that we may encounter the identification problem for some model specifications of the latent variables if $[\tau_1, ..., \tau_{K-1}]'$ are model parameters to be estimated. For instance, $w_{i,t}$ has an intercept in $\nu_{i,t}$, so there could be multiple values of $[\tau_1, ..., \tau_{K-1}]'$ consistent with corresponding multiple values of intercept, i.e. adding any constant to $[\tau_1, ..., \tau_{K-1}]'$ and the intercept equally. The intercept could also play a role of benchmark level for the fixed state transition probabilities independent of the endogenous shocks and exogenous variables. Therefore, to avoid this problem, we may let $w_{i,t}, i = 1, ..., n$, have the (regime-dependent) intercept and set $[\tau_1, ..., \tau_{K-1}]' = [\Phi^{-1}(\frac{1}{K}), ..., \Phi^{-1}(\frac{K-1}{K})]'$ as a choice of hyperparameters, where $\Phi^{-1}(p)$ is a quantile function of standard normal distribution to each probability p. This choice is to control the estimates of $\nu_{i,t}$ and $\gamma_{i,t}$ so that $w_{i,t}$ evolves relative to standard normal distribution.

Lastly, we write the vectors $\boldsymbol{w}_t = [w_{1,t}, ..., w_{n,t}]'$, $\boldsymbol{\eta}_t = [\eta_{1,t}, ..., \eta_{n,t}]'$, $\boldsymbol{\mu}_t = [\mu_{1,t}, ..., \mu_{n,t}]'$, $\boldsymbol{\Sigma}_t = \text{diag}(\sigma_{1,t}, ..., \sigma_{n,t})$, $\boldsymbol{\nu}_t = [\nu_{1,t}, ..., \nu_{n,t}]'$, and $\boldsymbol{\Gamma}_t = \text{diag}(\gamma_{1,t}, ..., \gamma_{n,t})$. These, as well as $\boldsymbol{L}_t$ and

$\boldsymbol{M}_t$, can be viewed as the functions of all individual states $\boldsymbol{s}_t$ and their lags. The multivariate regime switching model given by (33) and (37) can be written in a following compact form:

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\Sigma}_t \boldsymbol{L}_t \boldsymbol{\epsilon}_t, \tag{39}$$

$$\boldsymbol{w}_t = \boldsymbol{\nu}_t + \boldsymbol{\Gamma}_t \boldsymbol{M}_t \boldsymbol{\eta}_t, \tag{40}$$

where $s_{i,t} \in \boldsymbol{s}_t$ for each i and t follows (38).

3.2 Maximum likelihood estimation

Let $\boldsymbol{\mathcal{F}}_t = [\boldsymbol{y}_t', \boldsymbol{y}_{t-1}', ..., \boldsymbol{y}_1']'$ denote all historical information of the observed underlying time series up to period t. Let Ω_t denote all available information of exogenous variables up to period t. Let θ denote the model parameters. We write the log-likelihood function associated with the recursive formulation as follows:

$$\ln L(\theta) = \sum_{t=1}^{T} \ln p(\boldsymbol{y}_t | \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta) \tag{41}$$

where

$$p(\boldsymbol{y}_t | \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta) \tag{42}$$
$$= \sum_{\boldsymbol{s}_t \in \{0,...,K-1\}^n} \cdots \sum_{\boldsymbol{s}_{t-k-h} \in \{0,...,K-1\}^n} p(\boldsymbol{y}_t | \boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k-h}, \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta) p(\boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k-h} | \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta),$$

$$p(\boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k-h} | \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta) = p(\boldsymbol{s}_t | \boldsymbol{s}_{t-1}, ..., \boldsymbol{s}_{t-k-h}, \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta) p(\boldsymbol{s}_{t-1}, ..., \boldsymbol{s}_{t-k-h} | \boldsymbol{\mathcal{F}}_{t-1}; \Omega_{t-1}, \theta), \tag{43}$$

$$p(\boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k-h+1} | \boldsymbol{\mathcal{F}}_t; \Omega_t, \theta) = \sum_{\boldsymbol{s}_{t-k-h} \in \{0,...,K-1\}^n} p(\boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k-h} | \boldsymbol{y}_t, \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta)$$
$$= \frac{\sum_{\boldsymbol{s}_{t-k-h} \in \{0,...,K-1\}^n} p(\boldsymbol{y}_t | \boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k-h}, \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta) p(\boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k-h} | \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta)}{p(\boldsymbol{y}_t | \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta)}. \tag{44}$$

The recursion above is similar to that expressed in the two-regime case. The model parameters θ will be chosen with using the numerical optimization in order to maximize the log-likelihood function.

To proceed the above recursion, we may need to estimate $p(\boldsymbol{s}_t | \boldsymbol{s}_{t-1}, ..., \boldsymbol{s}_{t-k-h}, \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta)$ and $p(\boldsymbol{y}_t | \boldsymbol{s}_t, ..., \boldsymbol{s}_{t-k-h}, \boldsymbol{\mathcal{F}}_{t-1}; \Omega_t, \theta)$. Let $\pi_n(\boldsymbol{x}; \boldsymbol{v}, \boldsymbol{V})$ denote the probability density function of $n-$dimensional normal random vector evaluated at $\boldsymbol{x}$ with the parameters represented by the mean vector $\boldsymbol{v}$ and the covariance matrix $\boldsymbol{V}$. Lastly, recall that ϕ and Φ denote the probability density function and the cumulative distribution function of the standard normal random variable. We have the following results.

Proposition 3.1. *Suppose the multivariate regime switching model is modeled by* (38)*,* (39) *and* (40)*. Then, the following results hold.*

$$p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{\mathcal{F}}_{t-1};\Omega_t,\theta)$$
$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i=1}^{n}[\Phi(c_{i,t}(\boldsymbol{s}_t+\boldsymbol{1}))-\Phi(c_{i,t}(\boldsymbol{s}_t))]\phi(\epsilon_{1,t})\cdots\phi(\epsilon_{n,t})d\epsilon_{1,t}\cdots d\epsilon_{n,t}, \tag{45}$$

$$p(\boldsymbol{y}_t|\boldsymbol{s}_t,...,\boldsymbol{s}_{t-k-h},\boldsymbol{\mathcal{F}}_{t-1};\Omega_t,\theta) = \frac{\pi_n(\boldsymbol{y}_t;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t\boldsymbol{Q}_t\boldsymbol{\Sigma}_t)\prod_{i=1}^{n}[\Phi(c_{i,t}(\boldsymbol{s}_t+\boldsymbol{1}))-\Phi(c_{i,t}(\boldsymbol{s}_t))]}{p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{\mathcal{F}}_{t-1};\Omega_t,\theta)}, \tag{46}$$

where $[c_{1,t}(\boldsymbol{s}_t),...,c_{n,t}(\boldsymbol{s}_t)]' = \boldsymbol{M}_t^{-1}\boldsymbol{\Gamma}_t^{-1}(\boldsymbol{\tau}(\boldsymbol{s}_t)-\boldsymbol{\nu}_t)$*,* $\boldsymbol{\tau}(\boldsymbol{s}_t) = [\tau_{s_{1,t}},...,\tau_{s_{n,t}}]'$*, and* $\boldsymbol{1}$ *is* $n-$*dimensional unit vector.*

Proof. Let $\bar{\boldsymbol{s}} = [\bar{s}_1,...,\bar{s}_n]' \in \{0,...,K-1\}^n$ denote possible representatives of the states. Define $\boldsymbol{\tau}(\boldsymbol{s}_t) = [\tau_{s_{1,t}},...,\tau_{s_{n,t}}]'$ and let $\boldsymbol{1}$ be $n-$dimensional unit vector. Similar to the derivations in Proposition 2.1, it can be shown that

$$\begin{aligned}
&\Pr(\boldsymbol{s}_t=\bar{\boldsymbol{s}}|\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{\mathcal{F}}_{t-1};\Omega_t,\theta)\\
&\quad= \Pr(\boldsymbol{\tau}(\bar{\boldsymbol{s}})<\boldsymbol{w}_t\le\boldsymbol{\tau}(\bar{\boldsymbol{s}}+\boldsymbol{1})|\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{y}_{t-1},...,\boldsymbol{y}_{t-k-h};\Omega_t,\theta)\\
&\quad= \Pr(\boldsymbol{\tau}(\bar{\boldsymbol{s}})<\boldsymbol{w}_t\le\boldsymbol{\tau}(\bar{\boldsymbol{s}}+\boldsymbol{1})|\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{\epsilon}_{t-1},...,\boldsymbol{\epsilon}_{t-h};\Omega_t,\theta)\\
&\quad= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\Pr(\boldsymbol{M}_t^{-1}\boldsymbol{\Gamma}_t^{-1}(\boldsymbol{\tau}(\bar{\boldsymbol{s}})-\boldsymbol{\nu}_t)<\boldsymbol{\eta}_t\le\boldsymbol{M}_t^{-1}\boldsymbol{\Gamma}_t^{-1}(\boldsymbol{\tau}(\bar{\boldsymbol{s}}+\boldsymbol{1})-\boldsymbol{\nu}_t)|\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{\epsilon}_t,\\
&\qquad\qquad\boldsymbol{\epsilon}_{t-1},...,\boldsymbol{\epsilon}_{t-h};\Omega_t,\theta)p(\boldsymbol{\epsilon}_t)d\boldsymbol{\epsilon}_t\\
&\quad= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i=1}^{n}[\Phi(c_{i,t}(\bar{\boldsymbol{s}}+\boldsymbol{1}))-\Phi(c_{i,t}(\bar{\boldsymbol{s}}))]\phi(\epsilon_{1,t})\cdots\phi(\epsilon_{n,t})d\epsilon_{1,t}\cdots d\epsilon_{n,t}
\end{aligned}$$

where $[c_{1,t}(\boldsymbol{s}_t),...,c_{n,t}(\boldsymbol{s}_t)]' = \boldsymbol{M}_t^{-1}\boldsymbol{\Gamma}_t^{-1}(\boldsymbol{\tau}(\boldsymbol{s}_t)-\boldsymbol{\nu}_t)$. The last equality is obtained by the fact that the joint probability distribution function of the independent random variables is the product of their individual probability distribution function. Then, (45) is obtained. In addition, for $\boldsymbol{s}_t=\bar{\boldsymbol{s}}$, we can write

$$\begin{aligned}
&p(\boldsymbol{y}_t|\boldsymbol{s}_t=\bar{\boldsymbol{s}},...,\boldsymbol{s}_{t-k-h},\boldsymbol{\mathcal{F}}_{t-1};\Omega_t,\theta)\\
&= \frac{\Pr(\boldsymbol{\tau}(\bar{\boldsymbol{s}})<\boldsymbol{w}_t\le\boldsymbol{\tau}(\bar{\boldsymbol{s}}+\boldsymbol{1})|\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{y}_t,\boldsymbol{y}_{t-1},...,\boldsymbol{y}_{t-k-h};\Omega_t,\theta)\pi_n(\boldsymbol{y}_t;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t\boldsymbol{Q}_t\boldsymbol{\Sigma}_t)}{\Pr(\boldsymbol{\tau}(\bar{\boldsymbol{s}})<\boldsymbol{w}_t\le\boldsymbol{\tau}(\bar{\boldsymbol{s}}+\boldsymbol{1})|\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{y}_{t-1},...,\boldsymbol{y}_{t-k-h};\Omega_t,\theta)}.
\end{aligned}$$

Note that $\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t$ and $\boldsymbol{Q}_t$ in the $\pi_n(\boldsymbol{y}_t;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t\boldsymbol{Q}_t\boldsymbol{\Sigma}_t)$ above depend on $\boldsymbol{s}_t$ evaluated at $\bar{\boldsymbol{s}}$. To determine $\boldsymbol{w}_t$, the information $(\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{\epsilon}_t,\boldsymbol{\epsilon}_{t-1},...,\boldsymbol{\epsilon}_{t-h})$ is deduced from the information $(\boldsymbol{s}_{t-1},...,\boldsymbol{s}_{t-k-h},\boldsymbol{y}_t,\boldsymbol{y}_{t-1},...,\boldsymbol{y}_{t-k-h})$ where $\boldsymbol{\epsilon}_{t-j} = \boldsymbol{L}_{t-j}^{-1}\boldsymbol{\Sigma}_{t-j}^{-1}(\boldsymbol{y}_{t-j}-\boldsymbol{\mu}_{t-j})$, $j=0,...,h$. Thus, we obtain (46). □

4 Monte Carlo experiments

This section provides Monte Carlo experiments to evaluate models performance. Various simulations are conducted with data generated by the nonlinear endogenous switching model. The estimation results based on the linear endogenous switching model (i.e. restricted) is set to be benchmark in order to compare with the nonlinear endogenous switching model (i.e. unrestricted).

4.1 Model parameters and simulations

Consider a data-generating process given by

$$y_t = \mu(s_t) + \sigma(s_t)\epsilon_t, \quad s_t = 1_{\{w_t>0\}}, \tag{47}$$

$$w_t = \begin{cases} \alpha(s_{t-1}) + \rho_+(s_{t-1})\epsilon_t + \sqrt{1-\rho_+^2(s_{t-1})}\eta_t & \text{if } \epsilon_t > 0, \\ \alpha(s_{t-1}) + \rho_-(s_{t-1})\epsilon_t + \sqrt{1-\rho_-^2(s_{t-1})}\eta_t & \text{if } \epsilon_t \leq 0, \end{cases} \tag{48}$$

where ϵ_t and η_t are i.i.d. standard normal random variables, and $-1 < \rho_+(0), \rho_-(0), \rho_+(1), \rho_-(1) < 1$. The model above is a Markov two-regime switching mean and volatility model with constant transition probabilities. The model also allows the regime switching associated with the asymmetric and state-dependent endogeneity. In other words, the state transition probability endogenously and asymmetrically responds to the contemporary shock of the underlying model, and the degree of the responses also depends on the state before the transition. The same model will be used in the next section for an empirical analysis on stock market return data. The parameters in the experiments are set to be approximately consistent with those obtained in the empirical section.

The states may be viewed as the economic conditions such that the lower (higher) volatility and the higher (lower) mean of stock return represent the lower (higher) financial stress regime. We refer $s_t = 0$ and $s_t = 1$ to denote the states of low-volatility and high-volatility, respectively. For the simulations model, we consider the parameters of each regime to represent the conditions. The true parameters of the underlying process are set as $(\mu(0), \mu(1)) = (0.7, 0.0)$ and $(\sigma(0), \sigma(1)) = (2.5, 6.5)$ under several characteristics of the endogenous regime switching. When the financial stress is high, the return mean is zero and the return volatility exceeds the level doubled from the lower regime.

For the latent process, there are three sets of simulations that rely on different degrees and features of the endogeneity. First, the asymmetric endogeneity is individually considered so that the latent process (48) is set as $\rho_+(0) = \rho_+(1) = \rho_+$ and $\rho_-(0) = \rho_-(1) = \rho_-$ with $\rho_+ \neq \rho_-$. Specifically,

$$w_t = \begin{cases} \alpha(s_{t-1}) + \rho_+\epsilon_t + \sqrt{1-\rho_+^2}\eta_t & \text{if } \epsilon_t > 0, \\ \alpha(s_{t-1}) + \rho_-\epsilon_t + \sqrt{1-\rho_-^2}\eta_t & \text{if } \epsilon_t \leq 0. \end{cases} \tag{49}$$

This is to study the nonlinear endogeneity bias when the positive and negative shocks have distinct effects on changing the regime. They are varied as (ρ_+, ρ_-) = (-0.3,-0.7), (-0.1,-0.9), (-0.7,-0.3), (-0.9,-0.1). Second, another side is explored by only allowing the state-dependent endogeneity so that $\rho_+(0) = \rho_-(0) = \rho(0)$ and $\rho_+(1) = \rho_-(1) = \rho(1)$ with $\rho(0) \neq \rho(1)$. (48) can be written as

$$w_t = \alpha(s_{t-1}) + \rho(s_{t-1})\epsilon_t + \sqrt{1-\rho^2(s_{t-1})}\eta_t. \tag{50}$$

In other words, the regime transition probability symmetrically responds to the shock signs but it is different based on the previous regime.[3] They are characterized by the same degrees as $(\rho(0), \rho(1))$ = (-0.3,-0.7), (-0.1,-0.9), (-0.7,-0.3), (-0.9,-0.1). Third, the combined effects between the asymmetric and state-dependent endogeneity are considered in the simulations. The endogeneity parameters in (48) are chosen to be $(\rho_+(0), \rho_+(1), \rho_-(0), \rho_-(1))$ = (-0.1,-0.5,-0.5,-0.9), (-0.5,-0.9,-0.1,-0.5), (-0.5,-0.1,-0.9,-0.5), (-0.9,-0.5,-0.5,-0.1). For this first two subcases, the effects are relatively strong when the previous state is in the high-volatility, and rely more on negative and positive shocks respectively. For this last two subcases, the more effects are set to be in the previous regime being low-volatility, and the relative effects to the shock signs are the same as the first two subcases.

In addition to considering the correlation function ρ, $\alpha(0)$ and $\alpha(1)$ are chosen to control the degrees of state persistence. We consider two degrees by choosing the probabilities being the same state that $\big(p(s_t = 0|s_{t-1} = 0), p(s_t = 1|s_{t-1} = 1)\big)$ equals (0.95,0.90) for the high persistence, and equals (0.85,0.70) for the low persistence. Their stationary distributions are equivalent, and they have the unconditional probabilities 2/3 for $s_t = 0$ and 1/3 for $s_t = 1$.[4]

4.2 Monte Carlo results

For the Monte Carlo simulations described above, the models (47)-(48) are simulated with generations of ϵ_t and η_t.[5] This study considers 300 replications, and each replication is conducted with the sample of 500 periods ($T = 500$). Tables 1-3 below report the mean and the root mean squared error from the true value (RMSE) calculated from the 300 maximum likelihood estimates under the various characteristics of regime switching and the model restrictions.[6] The conventional regime switching model that restricts to the linear endogenous switching, i.e. $w_t =$

[3] The state-dependent endogeneity is also investigated in Cheng Gao and Yan (2018). But their latent model follows a first-order autoregressive latent process associated with one-period lag effect of the endogeneity, which is different from this study.

[4] Note that a Markov chain associated with a fixed transition probability matrix (P) has the stationary distribution (π) such that $\pi P = \pi$.

[5] To complete the data generation, we may need to initialize s_0. For each iteration of the simulations, it is drawn from the stationary distribution of the specified two-state Markov process.

[6] In the simulations, the correlation functions, ρ, and the transition probabilities are also estimated from the maximum likelihood estimation. The mean of the estimates are mostly consistent with the true values but are not reported here.

$\alpha(s_{t-1})+\rho\epsilon_t+\sqrt{1-\rho^2}\eta_t$, is referred to the restricted model. On the other hand, the unrestricted model is the model with the true endogeneity specification. It allows the endogeneous switching to be nonlinear so that the correlation ρ can be the function of the shock asymmetry (asymmetric endogeneity) or the previous state (state-dependent endogeneity) instead of a constant. The functions (24)-(25) where $\rho_t = \rho$ are applied to the recursive filter (10)-(12) for calculating the log-likelihood function of the restricted model. As the latent process (48) is a special case of (18), the functions (22)-(23) given in Proposition 2.2 with the two partitions and the shock threshold at zero are instead applied for the calculation of the unrestricted model. Then, the parameters of each model can be estimated by the numerical optimization applying with the Broyden–Fletcher–Goldfarb–Shanno algorithm.

Table 1 presents the Monte Carlo results based on the true model associated with the only asymmetric endogeneity in regime switching so that $\rho_+ \neq \rho_-$. The mean estimates of $\mu(0)$ are approximately close to the true value under the estimation of restricted and unrestricted models. For $\mu(1)$, the mean estimates are slightly biased from zero (or the true value) despite unrestricted model estimation, but they tend to be insignificant as shown in the high levels of RMSE. However, we can observe the biased characteristics of the volatility parameters. As shown in the restricted model, the mean estimates of $\sigma(0)$ are underestimated from the true value when the true model has the endogeneity effect highly on the negative shock ($|\rho_+| < |\rho_-|$). In this case, it can also be seen in the overestimation to the mean estimates of $\sigma(1)$. In contrast, when the endogeneity effect is highly on the positive shock ($|\rho_+| > |\rho_-|$), the mean estimates of $\sigma(0)$ and $\sigma(1)$ are biased from their true values in the opposite direction accordingly. The degrees of the biases become more obvious when the asymmetry level increases. Moreover, if we compare them between the state persistence levels, the biases tend to expand when the state persistence is relatively weak (or the regime is likely switched more frequently). For the results under the unrestricted model that the asymmetric endogeneity is accounted, the model performance improves as the nonlinear endogeneity biases decrease, especially in the volatility parameters. In addition, based on the levels of RMSE reported in all cases (also including the cases in Table 2-3), the estimators under the stronger state persistence tend to be relatively more efficient. This may be because of the higher frequency of the state transition that leads to more variation of the estimated parameters.

For the cases that the state-dependent endogeneity ($\rho(0) \neq \rho(1)$) is embedded in the regime shifts, the Monte Carlo results are reported in Table 2. The mean estimates of $\mu(0)$ are mostly close to its true value analogous to previous simulations. For the restricted model, the estimated $\mu(1)$, especially in the low persistence of state, seems to be biased to the negative value when $|\rho(0)| < |\rho(1)|$ and positive value when $|\rho(0)| > |\rho(1)|$. But the samples of their estimates are still highly distributed as represented in RMSE levels. Particularly, the overestimation and underestimation of the volatility estimates occur. The mean estimates of $\sigma(0)$ and $\sigma(1)$ tend to be over and under their true value, respectively, when the endogeneity effect depends more

(ρ_+, ρ_-)	Restricted model $\mu(0)$	$\mu(1)$	$\sigma(0)$	$\sigma(1)$	Unrestricted model $\mu(0)$	$\mu(1)$	$\sigma(0)$	$\sigma(1)$
	0.70	0.00	2.50	6.50	0.70	0.00	2.50	6.50
High persistence: $p(s_t = 0 \vert s_{t-1} = 0) = 0.95$, $p(s_t = 1 \vert s_{t-1} = 1) = 0.90$								
(-0.3,-0.7)	0.69	0.13	2.46	6.72	0.71	-0.04	2.49	6.52
	(0.17)	(0.74)	(0.14)	(0.51)	(0.17)	(0.70)	(0.15)	(0.54)
(-0.1,-0.9)	0.69	0.32	2.42	7.07	0.74	-0.18	2.49	6.46
	(0.16)	(0.85)	(0.14)	(0.75)	(0.16)	(0.64)	(0.12)	(0.43)
(-0.7,-0.3)	0.67	0.01	2.53	6.19	0.72	-0.15	2.47	6.45
	(0.19)	(0.71)	(0.15)	(0.51)	(0.20)	(0.75)	(0.16)	(0.55)
(-0.9,-0.1)	0.64	0.04	2.62	5.92	0.75	-0.21	2.49	6.38
	(0.20)	(0.74)	(0.21)	(0.69)	(0.20)	(0.75)	(0.18)	(0.54)
Low persistence: $p(s_t = 0 \vert s_{t-1} = 0) = 0.85$, $p(s_t = 1 \vert s_{t-1} = 1) = 0.70$								
(-0.3,-0.7)	0.73	0.12	2.40	6.85	0.67	0.05	2.47	6.61
	(0.24)	(1.29)	(0.22)	(0.61)	(0.24)	(0.96)	(0.20)	(0.56)
(-0.1,-0.9)	0.74	0.45	2.26	7.33	0.69	-0.12	2.41	6.58
	(0.24)	(1.50)	(0.30)	(0.99)	(0.16)	(0.54)	(0.17)	(0.38)
(-0.7,-0.3)	0.64	-0.04	2.61	6.15	0.73	-0.22	2.48	6.41
	(0.26)	(1.27)	(0.23)	(0.62)	(0.26)	(0.95)	(0.26)	(0.60)
(-0.9,-0.1)	0.60	-0.51	2.73	5.63	0.75	-0.29	2.51	6.24
	(0.33)	(1.93)	(0.30)	(1.07)	(0.27)	(0.71)	(0.26)	(0.75)

Table 1

Monte Carlo results under the regime switching with the asymmetric endogeneity, i.e. $\rho_+(0) = \rho_+(1) = \rho_+$ and $\rho_-(0) = \rho_-(1) = \rho_-$. Notes: The true values of (ρ_+, ρ_-) are presented in the first column and are associated with two degrees of persistence. The true value of $(\mu(0), \mu(1), \sigma(0), \sigma(1))$ are given in the column heading. Each cell contains the mean of the maximum likelihood estimates and the corresponding root mean squared error in the parenthesis. The restricted model is referred to the linear endogeneity with $\rho_+ = \rho_-$. The unrestricted model is referred to the nonlinear endogeneity allowing $\rho_+ \neq \rho_-$.

on the previous high-volatility regime that $|\rho(0)| < |\rho(1)|$. The bias directions of the mean estimates are also converse for the cases that $|\rho(0)| > |\rho(1)|$. Similarly, the degree of the biases is likely to increase in the level of state-dependent endogeneity and decrease in the level of state persistence. Overall, the biases of the mean estimates are significantly reduced if the state-dependent endogeneity is allowed in the model estimation, and the results are shown in the unrestricted model.

From the results discussed above, due to the restricted model, we observe that there are interesting bias features on the volatility parameters depending on either asymmetric or state-dependent endogeneity effects. For the last cases, we consider the regime switching model that the asymmetric and state-dependent endogeneity effects are combined together. The results reported in Table 3 are consistent with those obtained from the simulations based on the individual endogeneity effect. Specifically, $(\rho_+(0), \rho_+(1), \rho_-(0), \rho_-(1)) = (-0.1, -0.5, -0.5, -0.9)$ is associated with the cases of the relatively high effects on the negative shock ($|\rho_+| \leq |\rho_-|$) and on the previous

$(\rho(0),\rho(1))$	Restricted model $\mu(0)$	$\mu(1)$	$\sigma(0)$	$\sigma(1)$	Unrestricted model $\mu(0)$	$\mu(1)$	$\sigma(0)$	$\sigma(1)$
	0.70	0.00	2.50	6.50	0.70	0.00	2.50	6.50
High persistence: $p(s_t=0\|s_{t-1}=0)=0.95$, $p(s_t=1\|s_{t-1}=1)=0.90$								
(-0.3,-0.7)	0.69	-0.10	2.57	6.14	0.68	0.02	2.51	6.43
	(0.19)	(0.76)	(0.16)	(0.53)	(0.18)	(0.74)	(0.15)	(0.55)
(-0.1,-0.9)	0.68	-0.13	2.65	5.92	0.67	0.05	2.52	6.43
	(0.19)	(0.86)	(0.24)	(0.73)	(0.18)	(0.79)	(0.16)	(0.54)
(-0.7,-0.3)	0.69	0.21	2.43	6.75	0.69	0.00	2.48	6.53
	(0.17)	(0.70)	(0.14)	(0.50)	(0.16)	(0.69)	(0.14)	(0.46)
(-0.9,-0.1)	0.68	0.56	2.37	7.14	0.69	0.03	2.49	6.54
	(0.16)	(0.90)	(0.17)	(0.77)	(0.15)	(0.66)	(0.12)	(0.39)
Low persistence: $p(s_t=0\|s_{t-1}=0)=0.85$, $p(s_t=1\|s_{t-1}=1)=0.70$								
(-0.3,-0.7)	0.71	-0.41	2.61	6.13	0.69	-0.15	2.52	6.38
	(0.30)	(1.45)	(0.25)	(0.62)	(0.30)	(1.45)	(0.24)	(0.62)
(-0.1,-0.9)	0.85	-1.35	2.67	5.81	0.71	-0.18	2.50	6.30
	(0.40)	(2.07)	(0.31)	(0.87)	(0.30)	(1.24)	(0.22)	(0.65)
(-0.7,-0.3)	0.66	0.51	2.38	6.87	0.69	0.06	2.49	6.56
	(0.23)	(1.31)	(0.21)	(0.66)	(0.22)	(1.27)	(0.21)	(0.51)
(-0.9,-0.1)	0.69	0.92	2.21	7.22	0.69	0.05	2.49	6.54
	(0.20)	(1.40)	(0.33)	(0.90)	(0.17)	(0.90)	(0.16)	(0.44)

Table 2

Monte Carlo results under the regime switching with the state-dependent endogeneity, i.e. $\rho_+(0)=\rho_-(0)=\rho(0)$ and $\rho_+(1)=\rho_-(1)=\rho(1)$. Notes: The true values of $(\rho(0),\rho(1))$ are presented in the first column and are associated with two degrees of persistence. The true value of $(\mu(0),\mu(1),\sigma(0),\sigma(1))$ are given in the column heading. Each cell contains the mean of the maximum likelihood estimates and the corresponding root mean squared error in the parenthesis. The restricted model is referred to the linear endogeneity with $\rho(0)=\rho(1)$. The unrestricted model is referred to the nonlinear endogeneity allowing $\rho(0)\neq\rho(1)$.

regime being high-volatility ($|\rho(0)|\le|\rho(1)|$). In this case, the restricted model provides the appropriate results as the biased estimates produced from the individuals are possibly offset. It also holds for the cases $(\rho_+(0),\rho_+(1),\rho_-(0),\rho_-(1))=(-0.9,-0.5,-0.5,-0.1)$. On the other hand, the results under the cases that $(\rho_+(0),\rho_+(1),\rho_-(0),\rho_-(1))=(-0.5,-0.9,-0.1,-0.5)$ are also consistent with that individuals $|\rho_+|\ge|\rho_-|$ and $|\rho(0)|\le|\rho(1)|$. That is, the mean estimates of $\sigma(0)$ and $\sigma(1)$ under the restricted model are respectively above and below their true values. It can be explained in the same way to the cases that $(\rho_+(0),\rho_+(1),\rho_-(0),\rho_-(1))=(-0.5,-0.1,-0.9,-0.5)$. As the results, the nonlinear endogenous switching, if taken into account, provides more accurate estimates.

Lastly, the power of the statistical test is also explored to evaluate a significance of model improvement. Let us consider the likelihood ratio test statistic given by

$$2(\ln L(\hat{\theta}_{\mathrm{UR}})-\ln L(\hat{\theta}_{\mathrm{R}})) \tag{51}$$

where $\hat{\theta}_{\mathrm{UR}}$ and $\hat{\theta}_{\mathrm{R}}$ represent a set of parameters estimated from the unrestricted and restricted

$(\rho_+(0), \rho_+(1), \rho_-(0), \rho_-(1))$	Restricted model				Unrestricted model			
	$\mu(0)$	$\mu(1)$	$\sigma(0)$	$\sigma(1)$	$\mu(0)$	$\mu(1)$	$\sigma(0)$	$\sigma(1)$
	0.70	0.00	2.50	6.50	0.70	0.00	2.50	6.50
High persistence: $p(s_t = 0\|s_{t-1} = 0) = 0.95$, $p(s_t = 1\|s_{t-1} = 1) = 0.90$								
(-0.1,-0.5,-0.5,-0.9)	0.70	-0.07	2.51	6.39	0.74	-0.21	2.49	6.47
	(0.18)	(0.75)	(0.14)	(0.43)	(0.19)	(0.85)	(0.16)	(0.59)
(-0.5,-0.9-0.1,-0.5)	0.65	-0.04	2.67	5.89	0.70	-0.08	2.50	6.45
	(0.19)	(0.73)	(0.24)	(0.73)	(0.19)	(0.71)	(0.17)	(0.58)
(-0.5,-0.1,-0.9,-0.5)	0.69	0.45	2.39	7.16	0.69	0.00	2.49	6.55
	(0.16)	(0.86)	(0.16)	(0.79)	(0.16)	(0.67)	(0.13)	(0.46)
(-0.9,-0.5,-0.5,-0.1)	0.67	0.14	2.48	6.47	0.71	-0.09	2.47	6.52
	(0.18)	(0.72)	(0.14)	(0.43)	(0.18)	(0.71)	(0.15)	(0.56)
Low persistence: $p(s_t = 0\|s_{t-1} = 0) = 0.85$, $p(s_t = 1\|s_{t-1} = 1) = 0.70$								
(-0.1,-0.5,-0.5,-0.9)	0.77	-0.44	2.47	6.41	0.78	-0.43	2.45	6.46
	(0.30)	(1.45)	(0.22)	(0.48)	(0.31)	(1.45)	(0.24)	(0.6)
(-0.5,-0.9-0.1,-0.5)	0.69	-0.93	2.75	5.63	0.73	-0.25	2.48	6.27
	(0.30)	(2.16)	(0.32)	(1.04)	(0.28)	(0.95)	(0.23)	(0.68)
(-0.5,-0.1,-0.9,-0.5)	0.71	0.85	2.28	7.40	0.65	0.04	2.43	6.65
	(0.22)	(1.62)	(0.27)	(1.09)	(0.17)	(0.57)	(0.17)	(0.41)
(-0.9,-0.5,-0.5,-0.1)	0.63	0.33	2.47	6.49	0.78	-0.34	2.41	6.47
	(0.25)	(1.14)	(0.18)	(0.53)	(0.25)	(0.93)	(0.23)	(0.54)

Table 3

Monte Carlo results under the regime switching with the asymmetric and state-dependent endogeneity, i.e. unrestricted values of $(\rho_+(0), \rho_+(1), \rho_-(0), \rho_-(1))$. Notes: The true values of $(\rho_+(0), \rho_+(1), \rho_-(0), \rho_-(1))$ are presented in the first column and are associated with two degrees of persistence. The true value of $(\mu(0), \mu(1), \sigma(0), \sigma(1))$ are given in the column heading. Each cell contains the mean of the maximum likelihood estimates and the corresponding root mean squared error in the parenthesis. The restricted model is referred to the linear endogeneity with $\rho_+(0) = \rho_+(1) = \rho_-(0) = \rho_-(1)$. The unrestricted model is referred to the nonlinear edogeneity allowing unrestricted $(\rho_+(0), \rho_+(1), \rho_-(0), \rho_-(1))$.

models, respectively. It is asymptotically chi-squared distributed with the degree of freedom that equals to the difference number of parameters between $\hat{\theta}_{\mathrm{UR}}$ and $\hat{\theta}_{\mathrm{R}}$. Table 4 reports the power as the percentage of the Monte Carlo simulations rejected the null hypothesis that the endogeneity is linear at 5% significance level. For the cases that the endogeneity is individually asymmetric or state-dependent shown in the first two panels, intuitively, the power increases in the nonlinearity size. For the results from the combined effects of asymmetric and state-dependent endogeneity shown in the last panel, the power is relatively high when both effects individually lead to the same biased characteristics under the restricted model. The low level of the state persistence tends to have higher level of the power as more biased levels of the estimated parameters appear in the restricted model. This is mostly consistent with the results discussed above.

	Asymmetric endogeneity with (ρ_+, ρ_-) to be			
	(-0.3,-0.7)	(-0.1,-0.9)	(-0.7,-0.3)	(-0.9,-0.1)
High persistence	21.7	76.3	16.0	58.0
Low persistence	19.3	84.0	20.7	66.0
	State-dependent endogeneity with $(\rho(0), \rho(1))$ to be			
	(-0.3,-0.7)	(-0.1,-0.9)	(-0.7,-0.3)	(-0.9,-0.1)
High persistence	21.0	77.7	24.0	87.7
Low persistence	33.3	85.7	38.3	96.3
	Asymmetric and state-dependent endogeneity with $(\rho_+(0), \rho_+(1), \rho_-(0), \rho_-(1))$ to be			
	(-0.1,-0.5,-0.5,-0.9)	(-0.5,-0.9,-0.1,-0.5)	(-0.5,-0.1,-0.9,-0.5)	(-0.9,-0.5,-0.5,-0.1)
High persistence	12.0	44.3	60.3	6.0
Low persistence	19.3	36.3	73.7	10.3

Table 4

The power of the likelihood ratio test. Note: Each cell containing the number reports the percentage of the Monte Carlo simulations rejected the null hypothesis that the endogeneity is linear at the 5% significance level. The results in the first two and last panels are based on the tests associated with one and a three degrees of freedom, respectively.

5 An empirical example

This section applies the nonlinear endogenous regime switching model to the monthly excess returns on the US stock market. The value-weighted stock returns data on the NYSE index is considered in the excess of the US three-month T-bill rates as the risk-free rates. The data is collected from January 1966 to December 2023 in the daily basis. The monthly stock returns are computed by $r_t = (\frac{\text{NYSE}_t}{\text{NYSE}_{t-1}} - 1) \times 100$ where NYSE_t is the value of the index at the end of month t. For the monthly risk-free rates, the data is obtained in terms of continuously compounded rates, so they are constructed by $r_t^{\text{f}} = [\exp(\frac{\text{T-bill rate}_{t-1}}{365} \times N_t) - 1] \times 100$ where T-bill rate_t is the rate quoted at the end of month t and N_t is the number of days in month t. Then, define the time series data $y_t = r_t - r_t^{\text{f}}$ as the monthly excess returns.

The data is applied to estimate the two-regime switching mean-volatility model that allows the endogenous switching with asymmetric and state-dependent effects given by (47)-(48). Table 5 reports the estimation results under different endogeneity specifications based on two sample periods given by the full sample period (Jan 1966 - Dec 2023) and the more recent sample period (Jan 2000 - Dec 2023). We obtain $\hat{\mu}(0) > 0$ and $\hat{\sigma}(0) < \hat{\sigma}(1)$. For the estimation results on the full sample period, $\hat{\mu}(0)$ are around 0.5% representing positive excess return during the low-volatility regime. The estimates of $\hat{\mu}(1)$, except the exogenous switching case, are also positive but seem to be insignificant due to the high level of standard errors relative to the estimates. Ghysels Guerin Marcellino (2014), in a context of regime switching conditional variance MIDAS model, interpret the negative return during the high-volatility regime by the flight-to-quality effect. This is con-

Endogeneity Spec.	Parameter										Log-likelihood
	$\mu(0)$	$\mu(1)$	$\sigma(0)$	$\sigma(1)$	$\alpha(0)$	$\alpha(1)$	Endogeneity				
Sample period: Jan 1966 - Dec 2023											
Exogenous	0.70	-0.84	3.20	6.20	-1.50	1.05					-1985.9
	(0.18)	(0.54)	(0.19)	(0.36)	(0.25)	(0.28)					
							ρ				
Linear	0.34	0.89	3.33	6.43	-1.64	1.15	-0.65				-1977.7
	(0.19)	(0.73)	(0.18)	(0.38)	(0.19)	(0.25)	(0.16)				
							ρ_+		ρ_-		
Asymmetric	0.55	0.18	3.17	6.72	-1.34	1.29	-0.77		-0.18		-1976.7
	(0.17)	(0.70)	(0.24)	(0.57)	(0.47)	(0.24)	(0.17)		(0.50)		
							$\rho(0)$		$\rho(1)$		
State-dependent	0.36	0.92	3.29	6.81	-1.64	1.14	-0.51		-0.79		-1976.8
	(0.19)	(0.77)	(0.23)	(0.58)	(0.21)	(0.23)	(0.34)		(0.13)		
							$\rho_+(0)$	$\rho_+(1)$	$\rho_-(0)$	$\rho_-(1)$	
Asymmetric and	0.58	0.20	3.25	7.05	-1.33	1.14	-0.76	-0.74	-0.02	-0.61	-1975.7
State-dependent	(0.14)	(0.49)	(0.24)	(0.66)	(0.60)	(0.30)	(0.77)	(0.19)	(0.61)	(0.84)	
Sample period: Jan 2000 - Dec 2023											
Exogenous	0.98	-0.23	2.33	5.51	-1.71	1.83					-810.8
	(0.25)	(0.47)	(0.21)	(0.30)	(0.30)	(0.26)					
							ρ				
Linear	0.77	0.24	2.45	5.70	-1.63	1.72	-0.62				-808.0
	(0.28)	(0.52)	(0.24)	(0.34)	(0.25)	(0.29)	(0.32)				
							ρ_+		ρ_-		
Asymmetric	0.91	0.03	2.34	6.19	-1.16	1.71	-0.89		-0.08		-804.2
	(0.28)	(0.36)	(0.27)	(0.43)	(0.68)	(0.24)	(0.14)		(0.74)		
							$\rho(0)$		$\rho(1)$		
State-dependent	0.82	0.28	2.34	6.10	-1.63	1.71	-0.37		-0.93		-805.1
	(0.29)	(0.55)	(0.29)	(0.38)	(0.25)	(0.21)	(0.91)		(0.08)		
							$\rho_+(0)$	$\rho_+(1)$	$\rho_-(0)$	$\rho_-(1)$	
Asymmetric and	0.93	0.11	2.31	6.33	-1.12	1.66	-0.85	-0.94	0.06	-0.64	-803.3
State-dependent	(0.15)	(0.60)	(0.30)	(0.44)	(0.94)	(0.24)	(0.99)	(0.08)	(1.18)	(1.08)	

Table 5

Maximum likelihood estimates for the model (47)-(48) with different endogeneity specifications. Note: The parameters are estimated on the monthly excess returns data. The estimated parameters are reported corresponding to the column heading and endogeneity specifications. The standard errors of the estimators reported in the parentheses are based on the outer product of the gradient of the log-likelihood function.

sistent with them for the case that the regime switching is exogenous, but we find that the effect tends to disappear when the endogenous regime switching is allowed. For the model accounting the linear endogeneity, the return shock is negatively correlated to the probability of state being the high-volatility regime (i.e. $\hat{\rho} < 0$). This is also consistent with Kim Piger and Startz (2008) and Chang Choi and Park (2017) who propose the linear endogenous switching models and applies their model to the stock return data as well. However, the model accounting the asymmetric and state-dependent endogeneity effects informatively provides that the regime changes are strongly affected by the contemporaneous positive shocks, especially during the previous regime being high-volatility. In addition, the model ignoring the nonlinear endogeneity effects could lead to the biased volatility estimation. Namely, it overestimates $\hat{\sigma}(0)$ and in particular underestimates $\hat{\sigma}(1)$. For the estimation results on the more recent sample period, the estimates $\hat{\mu}(0)$ relatively increase, and $\hat{\sigma}(0)$ and $\hat{\sigma}(1)$ relatively decrease for all endogeneity cases. In particular, the results present the increases in the nonlinear endogeneity effects, so it may cause the volatility biases to be more serious. Based on this example, the model that allows the nonlinear endogenous switching as proposed could be more effective as well as give us additional information on the characteristics in transferring states.

6 Conclusion

This paper proposes the regime switching models in which the switching can be nonlinearly endogenous. The latent process controlling the state transition is endogenously explained by the shocks of the underlying model. The endogenous effects of the shocks are characterized by the free functional form which allows nonlinearities. This approach is mainly an extension of the endogenous switching model proposed by Kim Piger and Startz (2008) to a more general version. We also refer to the traditional endogenous switching models that usually assume the shocks between the underlying and latent equations to be jointly normal and correlated in the linear manner. In this paper, the model parameters estimation is provided by applying the recursive filer algorithm to conduct the maximum likelihood estimation. A two-regime switching model is first discussed, and then extended to a multivariate model with multiple regimes switching.

The two-regime switching mean-volatility model is considered for the numerical examples to explore the advantage of allowing the nonlinear endogenous switching characterized by the asymmetric and state-dependent effects. The Monte Carlo studies show that the presence of either two effects, if ignored, could noticeably lead to the volatility biases, but it is not obvious for the mean. Under the linear endogenous switching model, the estimated volatility during the high-volatility regime tends to be overestimated (underestimated) when the asymmetric effect is more negative (positive) and/or the state-dependent effect relies more on the low-volatility (high-volatility) regime. It is conversely true for the estimated volatility during the low-volatility regime. The proposed model and estimation that allows the nonlinear endogeneity in regime switching provides the more accurate estimates. In addition, the same model is analyzed to the monthly excess returns on the US stock market. The results show that the regime shifts are essentially controlled by the contemporaneous positive shocks, and especially during the previous regime being high-volatility. Under the linear endogenous switching model, we obtain the biased patterns of the volatility consistent with that obtained from the Monte Carlo simulations. This paper only analyzes the empirical example by the simple model with the contemporary endogeneity effects.

The lagged endogeneity effects are not yet explored here. It would also be interesting to apply the proposed approach to other applications, for instance, multivariate analyses using the multivariate regime switching models taken into account the nonlinear endogenous switching.

References

[1] Ang, A., Timmermann, A., 2012. Regime Changes and Financial Markets. Annual Review of Financial Economics 4, 313-337.

[2] Bazzi, M., Blasques, F., Koopman, S. J., Lucas, A., 2017. Time-Varying Transition Probabilities for Markov Regime Switching Models. Journal of Time Series Analysis 38, 458-478.

[3] Bekaert, G., Wu, G., 2000. Asymmetric Volatility and Risk in Equity Markets. The Review of Financial Studies 13, 1-42.

[4] Cenesizoglu, T., 2015. The Reaction of Stock Returns to News About Fundamentals. Management Science 61, 931-1196.

[5] Chang, Y., Choi, Y., Park, J. Y., 2017. A new approach to model regime switching. Journal of Econometrics 196, 127–143.

[6] Chang, Y., Maih, J., Tan, F., 2021. Origins of monetary policy shifts: A New approach to regime switching in DSGE models. Journal of Economic Dynamics & Control 133, 104235.

[7] Chang, Y., Tan, F., Wei, X., 2018. State Space Models with Endogenous Regime Switching. CAEPR working paper

[8] Cheng, T., Gao, J., Yan, Y., 2018. A new regime switching model with state–varying endogeneity. Journal of Management Science and Engineering 3, 214–231.

[9] Chib, S., Dueker, M., 2004. Non-Markovian Regime Switching with Endogenous States and Time-Varying State Strengths. Federal Reserve Bank of St. Louis working paper 2004-030A.

[10] Diebold, F. X., Lee, J.-H., Weinbach, G. C., 1994. Regime Switching with Time-Varying Transition Probabilities, in Hargreaves, C., ed., Time Series Analysis and Cointegration, Oxford University Press.

[11] Engle, R. F., NG, V. K., 1993. Measuring and testing the impact of news on volatility. Journal of Finance 48, 1749–1778.

[12] Filardo, A. J., Gordon, S. F., 1998. Business cycle durations. Journal of Econometrics 85, 99-123.

[13] Ghysels, E., Clara, P. S-., Volkanov, R., 2005. There is a risk-return trade-off after all. Journal of Financial Economics 76, 509-548.

[14] Ghysels, E., Guerin, P. Marcellino, M., 2014. Regime switches in the risk–return trade-off. Journal of Empirical Finance 28, 118-138.

[15] Hamilton, J. D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357-384.

[16] Hamilton, J. D., 2010. Regime switching models. In: Durlauf, S.N., Blume, L.E. (eds) Macroeconometrics and Time Series Analysis, 202-209.

[17] Hamilton, J. D., 2016. Macroeconomic Regimes and Regime Shifts. Handbook of Macroeconomics 2, 163-201.

[18] Kang, K. H., 2014. Estimation of state-space models with endogenous Markov regime-switching parameters. The Econometrics Journal 17, 56-82.

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

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

[21] Nelson, D. B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59, 347–370.

[22] Song, W., Ryu, D., Webb, R. I., 2018. Volatility dynamics under an endogenous Markov-switching framework: A cross-market approach. Quantitative Finance 18, 1559–1571.

[23] Veldkamp, L. L., 2005. Slow boom, sudden crash. Journal of Economic Theory 124, 230-257.

[24] Zhang, X., Bouri, E., Xu, Y., Zhang, G., 2022. The asymmetric relationship between returns and implied higher moments: Evidence from the crude oil market. Energy Economics 109, 105950.