Statistica Sinica **34** (2024), 1367-1389
doi:https://doi.org/10.5705/ss.202021.0336

ASYMPTOTIC BEHAVIOR OF THE MAXIMUM LIKELIHOOD ESTIMATOR FOR GENERAL MARKOV SWITCHING MODELS

Cheng-Der Fuh and Tianxiao Pang*

National Central University and Zhejiang University

Abstract: Motivated by studying the asymptotic properties of the parameter estimator in switching linear state space models, switching GARCH models, switching stochastic volatility models, and recurrent neural networks, we investigate the maximum likelihood estimator for general Markov switching models. To this end, we first propose an innovative matrix-valued Markovian iterated function system (MIFS) representation for the likelihood function. Then, we express the derivatives of the MIFS as a composition of random matrices. To the best of our knowledge, this is a new method in the literature. Using this useful device, we establish the strong consistency and asymptotic normality of the maximum likelihood estimator under some regularity conditions. Furthermore, we characterize the Fisher information as the inverse of the asymptotic variance.

Key words and phrases: Asymptotic normality, consistency, Markovian iterated function systems, recurrent neural networks, switching linear state space model.

1. Introduction

Motivated by studying the asymptotic properties of the parameter estimator in switching linear state space models, switching GARCH models, switching stochastic volatility (SV) models, and recurrent neural networks (RNNs), we investigate the maximum likelihood estimator (MLE) for general Markov switching models (GMSMs). Let $\{H_t, t \geq 0\}$ be an ergodic (aperiodic, irreducible, and positive recurrent) Markov chain on a finite state space $\mathcal{D} = \{1, \ldots, d\}$, and denote

$$Y_t = g_{H_t}(X_t, Y_{t-1}, \varepsilon_t; \theta),\ t \geq 1, \quad \text{with } Y_0 = \mathbf{0}, \tag{1.1}$$

$$X_t = f_{H_t}(X_{t-1}, \eta_t; \theta),\ t \geq 1, \quad \text{with } X_0 = \mathbf{0}, \tag{1.2}$$

where $Y_t \in \mathbf{R}^p$, for some $p \geq 1$, $X_t \in \mathbf{R}^m$, for some $m \geq 1$, $\{\varepsilon_t, t \geq 1\}$ is a sequence of independent and identically distributed (i.i.d.) $p \times 1$ random vectors, and $\{\eta_t, t \geq 1\}$ is a sequence of i.i.d. $m \times 1$ random vectors. Furthermore, we assume that $\{H_t, t \geq 0\}$ is a first-order Markov chain, and that $\{H_t, t \geq 0\}$, $\{\eta_t, t \geq 1\}$, and $\{\varepsilon_t, t \geq 1\}$ are independent. The GMSM is very flexible, and

*Corresponding author.

includes the aforementioned models as special cases. For example, if g_{H_t} and f_{H_t} are linear functions and there is no dynamic structure in the observations $\{Y_t, t \geq 0\}$, the GMSM is reduced to the following well-known switching linear state space model:

$$Y_t = B_t(H_t)X_t + \varepsilon_n,\ t \geq 1, \quad \text{with } Y_0 = \mathbf{0}, \tag{1.3}$$

$$X_t = A_t(H_t)X_{t-1} + \eta_n,\ t \geq 1, \quad \text{with } X_0 = \mathbf{0}; \tag{1.4}$$

see Kim (1994) and Ghahramani and Hinton (2000).

A GMSM is, loosely speaking, a two-layer Markov switching model (MSM) or a two-layer state space model. Specifically, let $\mathbf{Y} = \{Y_t, t \geq 0\}$ be a sequence of random variables obtained in the following way. First, a realization of a Markov chain $\mathbf{X} = \{X_t, t \geq 0\}$ is created. This chain is sometimes called the regime, and is not observed. Then, conditioned on $\mathbf{X}$, the $\mathbf{Y}$-variables are generated. Usually, the dependency of Y_t on $\mathbf{X}$ is more or less local, as when $Y_t = g(X_t, Y_{t-1}, \varepsilon_t)$, for some function g and random sequence $\{\varepsilon_t, t \geq 1\}$, independent of $\mathbf{X}$. In general, Y_t itself is not Markovian, and may in fact have a complicated dependency structure. When the state space of $\{X_t, t \geq 0\}$ is finite, it is the so-called hidden Markov model or MSM. In this paper, we consider a GMSM in which the underlying Markov chain $\mathbf{X}$ depends on a regime switching. That is, there is an extra finite state Markov chain $\mathbf{H} = \{H_t, t \geq 0\}$ such that, conditional on H_t, X_t is a general state Markov chain, for $t \geq 0$. Moreover, $\mathbf{Y}$ depends on both $\mathbf{H}$ and $\mathbf{X}$.

The purpose of this study is to provide a theoretical justification for the MLE in a GMSM. A major difficulty when analyzing the likelihood function in a GMSM is that the function can be expressed only in recursive integral form; see Equation (2.4) below, for instance. Here, we use the device in (2.5)–(2.13), to represent the probability density and the likelihood function in (2.4) as the L_1-norm of a matrix-valued Markovian iterated function system (MIFS). Then, the log likelihood function can be expressed in additive form, as in (3.7), to which we can apply the standard argument of the likelihood function for the "enlarged" Markov chain. This representation also gives a fast numerical computation algorithm of the invariant probability and the Kullback–Leibler divergence for a two-state hidden Markov model; see Fuh and Mei (2015). Furthermore, it may provide a fast algorithm for evaluating of the likelihood function using the EM algorithm. Note that the asymptotic behavior of MIFS is examined in detail by Fuh (2021). This new device enables us to apply the results of the strong law of large numbers and the central limit theorem for the asymptotic distributions of the matrix-valued MIFS, as well as to verify the strong consistency and asymptotic normality of the MLE in a GMSM.

Next, we give a brief summary of the literature on GMSMs. Note that a GMSM has two-layer hidden states $\mathbf{H}$ and $\mathbf{X}$. When there is no hidden state $\mathbf{X}$, and $\mathbf{Y}$ is conditionally independent for given $\mathbf{H}$, the GMSM is the classical hidden Markov model, and has attracted much attention because of its importance in, for example, speech recognition, signal processing, ion channels, and molecular biology. When $\mathbf{Y}$ forms an autoregression model for a given $\mathbf{H}$, the GMSM reduces to the MSM of Hamilton (1989) and the Markov switching multifractal models of, for example, Calvet and Fisher (2001). When there is only $\mathbf{X}$ and no hidden state $\mathbf{H}$, the GMSM includes the celebrated (G)ARCH models, as in Engle (1982) and Bollerslev (1986), SV models, as in Taylor (1986), and RNNs, as in Goodfellow, Bengio, and Courville (2016). Refer to Hamilton (1994) and Fan and Yao (2003) for a comprehensive summary.

When there are two-layer hidden states $\mathbf{H}$ and $\mathbf{X}$, the GMSM includes the switching linear state space model, as in Kim (1994) and Ghahramani and Hinton (2000), switching GARCH models, as in Cai (1994) and Hamilton and Susmel (1994), switching SV models, as in So, Lam, and Li (1998), and variational RNNs, as in Chung et al. (2015). When $\mathbf{H} = \{H_t, t \geq 0\}$ are i.i.d. finite-valued random variables, and $\{X_t, t \geq 0\}$ is a finite-state Markov chain for given $\mathbf{H}$, then $\{Y_t, t \geq 0\}$ is the factorial hidden Markov model, as in Ghahramani and Jordan (1997). These prior works focus on state space modeling and estimation, algorithms for fitting these models, and implementing likelihood-based methods. For instance, Kim (1994) and Ghahramani and Hinton (2000) propose a Kalman-filter-based method and a variational approximation method, respectively, to implement the MLE in switching linear state space models, and Davig and Doh (2014) estimate new Keynesian general equilibrium models using switching monetary policy rules.

RNNs are a popular modeling choice for solving sequence learning problems in machine learning (see Goodfellow, Bengio, and Courville (2016)). Early applications of RNN models in econometrics can be found in Kuan and White (1994) and White (1988), among others. Recent approaches have used artificial neural networks for auction design, as in Dütting et al. (2019), for estimating causal relationships, developing the broad idea of instrumental variables, as in Hartford et al. (2016), for portfolio theory in finance, as in Sirignano (2019) and Gu, Kelly, and Xiu (2020), and for time series, as in Verstyuk (2020). Owing to the model complexity, most econometrics and machine learning studies use the gradient descent and/or stochastic gradient descent to compute the MLE. For instance, Rumelhart, Hinton, and Williams (1987) propose a recursive algorithm (backpropagation learning) that speeds up the gradient descent method, and White (1989) establishes the consistency and asymptotic normality of the algorithm. Adaptive moment (Adam) estimation is a recent popular adaptive gradient algorithm used in machine learning, for example, in Kingma and Ba (2015).

There is extensive literature on the MLE in a special case of the GMSM in which there is only one finite hidden state **H**. When the observation is a deterministic function of the state space, Baum and Petrie (1966) establish the consistency and asymptotic normality of the MLE. When the observed random variables are conditionally independent, Leroux (1992) proves the strong consistency of the MLE, and Bickel, Ritov, and Rydén (1998) establish the asymptotic normality of the MLE, under mild conditions. By extending the inference problem to time-series analysis, where the state space is finite and the observed random variables are conditionally Markovian dependent, Goldfeld and Quandt (1973) and Hamilton (1989) use the MLE in switching autoregression with Markov regimes. Francq and Roussignol (1998) and Douc, Moulines, and Rydén (2004) study the consistency and asymptotic normality, respectively, of the MLE in Markov-switching autoregressive models, and Fuh (2004) establishes the Bahadur efficiency of the MLE in MSMs. When $\{Y_t, t \geq 0\}$ are conditionally independent given $\mathbf{X}$, Jensen and Petersen (1999) and Douc and Matias (2001) study the asymptotic properties of the MLE. Douc et al. (2011) study the consistency of the MLE for general hidden Markov models. The strong consistency and asymptotic normality of the MLE for general state hidden Markov models can be found in Fuh (2006).

This study makes three contributions to the literature. First, we provide a probability framework for the GMSM, which includes hidden Markov models, MSMs, (switching) GARCH(p, q) models, (switching) SV models, (switching) linear state space models, and variational RNNs as special cases. Moreover, we use a dynamic economic model's viewpoint to analyze the two-layer RNN model in machine learning. Second, in order to establish the strong consistency and asymptotic normality of the MLE under some regularity conditions, we first propose an innovative matrix-valued MIFS representation for the likelihood function, and then express the derivatives of the MIFS as a composition of random matrices. To the best of our knowledge, this is a new method in the literature. Moreover, we provide a weaker weighted local mean contractive condition and fill the gap in the proof of asymptotic normality in Fuh (2006). Third, we characterize the Fisher information as the inverse of the asymptotic variance by showing that the derivatives of the likelihood function still form a matrix-valued MIFS. These results can be applied to Markov switching models, nonlinear state space models, and SV models as well.

The remainder of this paper is organized as follows. In Section 2, we formally define the GMSM and represent its likelihood function as the L_1-norm of a matrix-valued MIFS. Section 3 investigates the MLE in the GMSM, and states the main results. Section 4 studies derivatives of the matrix-valued MIFS and the score function, and then characterizes the Fisher information. Section 5 concludes the paper. In Section S1 of the Supplementary Material, we consider several interesting examples, including switching linear state space models, switching

GARCH(p, q) models, switching SV models, and variational RNNs, which are popular in econometrics and machine learning. A simulation study and all technical proofs are given in Section S2 and Section S3, respectively, of the Supplementary Material.

2. GMSMs

In general, a GMSM is not Markovian. However, in this section, we provide a probability framework for the GMSM, under which it can be regarded as a Markov chain in an enlarged state space. There are two Markov chain representations for the GMSM. First, a GMSM is defined as a parameterized Markov chain in a Markovian random environment, with the underlying environmental Markov chain viewed as missing data. Specifically, let $\mathbf{H} = \{H_t, t \geq 0\}$ be an ergodic (aperiodic, irreducible, and positive recurrent) Markov chain on a finite state space $\mathcal{D} = \{1, \ldots, d\}$, with transition probability $p^\theta_{ij} = P^\theta\{H_1 = j | H_0 = i\}$ and stationary probability $\pi^\theta_H(\cdot)$. For given $\mathbf{H}$, let $\mathbf{X} = \{X_t, t \geq 0\}$ be a Markov chain on a general state space $\mathcal{X}$, with transition probability kernel $P^\theta_j(x, \cdot) = P^\theta\{X_1 \in \cdot | H_1 = j, X_0 = x\}$ and stationary probability $\pi^\theta_X(\cdot | H_0 = j)$, where $\theta \in \Theta \subseteq \mathbf{R}^q$ denotes the unknown parameter. Suppose that a random sequence $\{Y_t, t \geq 0\}$, taking values in $\mathbf{R}^p$, is adjoined to the chain such that $\{((H_t, X_t), Y_t), t \geq 0\}$ is a Markov chain on $(\mathcal{D} \times \mathcal{X}) \times \mathbf{R}^p$, such that conditioning on the full $\mathbf{H}$ sequence, $\{X_t, t \geq 0\}$ is a Markov chain with probability

$$\begin{cases} P^\theta\{X_0 \in A | H_0, H_1, \ldots, Y_0 = y\} = P^\theta\{X_0 \in A | H_0\}, \\ P^\theta\{X_1 \in A | H_0, H_1, \ldots, X_0 = x, Y_0 = y\} = P^\theta\{X_1 \in A | H_1, X_0 = x\} \ a.s., \end{cases} \tag{2.1}$$

for $A \in \mathcal{B}(\mathcal{X})$, the Borel σ-algebra of $\mathcal{X}$. Furthermore, conditioning on the full $(\mathbf{H}, \mathbf{X})$ sequence, $\{Y_t, t \geq 0\}$ is a Markov chain with probability

$$\begin{cases} P^\theta\{Y_0 \in B | H_0, H_1, \ldots, X_0, X_1, \ldots\} = P^\theta\{Y_0 \in B | H_0, X_0\}, \\ P^\theta\{Y_{t+1} \in B | H_0, H_1, \ldots, X_0, X_1, \ldots; Y_0, Y_1, \ldots, Y_t\} = \\ P^\theta\{Y_{t+1} \in B | H_{t+1}, X_{t+1}; Y_t\} \ a.s., \end{cases} \tag{2.2}$$

for each t and $B \in \mathcal{B}(\mathbf{R}^p)$, the Borel σ-algebra of $\mathbf{R}^p$. Note that in (2.2), the conditional probability of Y_{t+1} depends only on (H_{t+1}, X_{t+1}) and Y_t. Furthermore, we assume the existence of a transition probability density $p^\theta_j(x, x')$ for the Markov chain $\{X_t, t \geq 0\}$, given $H_t = j$, with respect to a σ-finite measure m on $\mathcal{X}$ such that for $i, j \in \mathcal{D}$,

$$\begin{aligned} &P^\theta\{H_1 = j, X_1 \in A, Y_1 \in B | H_0 = i, X_0 = x, Y_0 = y_0\} \\ &= \int_{x' \in A} \int_{y \in B} p^\theta_{ij} p^\theta_j(x, x') f(y; \theta | j, x', y_0) Q(dy) m(dx'), \end{aligned}$$

where $f(Y_k;\theta|H_k,X_k,Y_{k-1})$ is the conditional probability density of Y_k given $((H_k,X_k),Y_{k-1})$, with respect to a σ-finite measure Q on $\mathbf{R}^p$. We also assume that the Markov chain $\{((H_t,X_t),Y_t),t\geq 0\}$ has a stationary probability with probability density function $\pi_H^\theta(h_0)\pi_X^\theta(x_0|h_0)f(\cdot;\theta|h_0,x_0)$ with respect to $m\times Q$. In this paper, we consider $\theta=(\theta_1,\ldots,\theta_q)^\mathsf{T}\in\Theta\subseteq\mathbf{R}^q$ as the unknown parameter (here, and in what follows, T denotes the transpose of a vector or matrix), and the true parameter value is denoted by θ_0. We use $\pi_H(j)$ for $\pi_H^\theta(j)$, $\pi_X(x|j)$ for $\pi_X^\theta(x|j)$, $p_j(x,x')$ for $p_j^\theta(x,x')$, $f(y_0|H_0,X_0)$ for $f(y_0;\theta|H_0,X_0)$, and $f(y_k|H_k,X_k,Y_{k-1})$ for $f(y_k;\theta|H_k,X_k,Y_{k-1})$, depending on the context.

The following is a formal definition of the GMSM.

Definition 1. $\{Y_t,t\geq 0\}$ is called a GMSM if there is a Markov chain $\{(H_t,X_t),t\geq 0\}$ such that the process $\{((H_t,X_t),Y_t),t\geq 0\}$ is a Markov chain that satisfies (2.1) and (2.2).

For the first Markov chain representation of the likelihood function of the GMSM, recall that $\pi_H^\theta(h_0)\pi_X^\theta(x_0|h_0)f(y_0;\theta|h_0,x_0)$ is the stationary probability density, with respect to $m\times Q$, of the Markov chain $\{((H_t,X_t),Y_t),t\geq 0\}$. Note that the joint probability of $\{Y_t,t=0,\ldots,n\}$ is

$$\begin{aligned}&P\{Y_0\in B_0,Y_1\in B_1,\ldots,Y_n\in B_n\}\\&=\int_{y_0\in B_0}\int_{y_1\in B_1}\cdots\int_{y_n\in B_n}p_n(y_0,y_1,\ldots,y_n;\theta)Q(dy_n)\cdots Q(dy_1)Q(dy_0),\end{aligned}\tag{2.3}$$

where

$$\begin{aligned}p_n(y_0,y_1,\ldots,y_n;\theta)=&\sum_{h_0,\ldots,h_n=1}^{d}\int_{x_0,x_1,\ldots,x_n\in\mathcal{X}}\pi_H^\theta(h_0)\pi_X^\theta(x_0|h_0)f(y_0;\theta|h_0,x_0)\\\times\prod_{t=1}^{n}&p_{h_{t-1}h_t}^\theta p_{h_t}^\theta(x_{t-1},x_t)f(y_t;\theta|h_t,x_t,y_{t-1})m(dx_n)\cdots m(dx_0).\end{aligned}\tag{2.4}$$

To illustrate the GMSM, we use the switching linear state space model given in (1.3) and (1.4). Other examples, including the switching GARCH models, switching SV models, and variational RNNs, are provided in the Supplementary Material.

Example 1 (Switching linear state space models). Consider the model in (1.3) and (1.4), with $X_0=\mathbf{0}$ replaced with the stationary distribution π_X, where $B_t(H_t)=:B_t$ and $A_t(H_t)=:A_t$ are $p\times m$ and $m\times m$ random matrices, respectively, governed by $\{H_t,t\geq 0\}$. Let $\{(H_t,X_t),t\geq 0\}$ be a Markov chain on a general state space $\mathcal{D}\times\mathbf{R}^m$ with Borel σ-algebra $\mathcal{B}(\mathcal{D})\times\mathcal{B}(\mathbf{R}^m)$, which is irreducible with respect to a maximal irreducibility measure on $(\mathcal{D}\times\mathbf{R}^m,\mathcal{B}(\mathcal{D})\times\mathcal{B}(\mathbf{R}^m))$ and is aperiodic. With a slight abuse of notation, we still let $P(\cdot,\cdot)$ denote the transition probability kernel and assume that (H_t,X_t) has stationary measure $\pi_H(h_0)\pi_X(\cdot|h_0)$.

When $\varepsilon_t \sim N(\mu, \sigma^2)$, $\eta_t \sim N(0,1)$, $B_t = \beta_{H_t} \in \mathbf{R}$, and $A_t = \alpha_{H_t} \in \mathbf{R}$, with $|\alpha_j| < 1$, for $j = 1, \ldots, d$, then for given $H_t = j$, $\{X_t, t \geq 0\}$ forms a Markov chain with transition probability density function

$$p_j(x_{t-1}, x_t) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-(x_t - \alpha_j x_{t-1})^2}{2}\right\}.$$

For given observations $\mathbf{y} = (y_1, \ldots, y_n)$ from the switching linear state space model (1.3) and (1.4), the likelihood function of the parameter $\theta = (\alpha_1, \ldots, \alpha_d, \beta_1, \ldots, \beta_d, \mu, \sigma^2)^\mathsf{T}$ is

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{y}) = &\sum_{h_0,h_1,\ldots,h_n=1}^{d} \int_{x_0,\ldots,x_n\in\mathcal{X}} \pi_H(h_0)\pi_X(x_0|h_0) \\ &\cdot \prod_{t=1}^{n} p_{h_{t-1}h_t} p_{h_t}(x_{t-1}, x_t)\phi_{\mu,\sigma^2}(y_t - \beta_{h_t}x_t)dx_n \cdots dx_0,\end{aligned}$$

where $\phi_{\mu,\sigma^2}(\cdot)$ is the probability density function of $N(\mu, \sigma^2)$; see Section S1 in the Supplementary Material for further details.

For the second Markov chain representation of the GMSM in (2.3) and (2.4), which we use to analyze the MLE of the GMSM, we first write the random joint probability density function $p_n(Y_0, Y_1, \ldots, Y_n; \theta)$ as the L_1-norm of a composition of Markovian random matrices, each component of which is a Markovian iterated random function. Specifically, let

$$\begin{aligned}\mathbf{M} = \Bigg\{ &g|g : \mathcal{X} \mapsto \mathbf{R} \text{ is } m\text{−measurable}, \\ &\int |g(x)|m(dx) < \infty \text{ and } \sup_{x\in\mathcal{X}} |g(x)| < \infty \Bigg\}.\end{aligned} \tag{2.5}$$

For each $t = 1, \ldots, n$ and $j = 1, \ldots, d$, define the random functions $\mathbf{P}_j^\theta(Y_0)$ and $\mathbf{P}_j^\theta(Y_t)$ on $(\mathcal{X} \times \mathbf{R}^p) \times \mathbf{M}$ as

$$\mathbf{P}_j^\theta(Y_0)[g(x)] = f(Y_0; \theta|j, x)g(x), \tag{2.6}$$

$$\mathbf{P}_j^\theta(Y_t)[g(x)] = \int_{x'\in\mathcal{X}} p_j^\theta(x', x) f(Y_t; \theta|j, x, Y_{t-1})g(x')m(dx'). \tag{2.7}$$

For the definition of $\mathbf{P}_j^\theta(Y_t)[g(x)]$ in (2.7), we consider the reverse of the transition probability density, which generalizes the corresponding result in hidden Markov models; see (1.5) in Fuh (2003). Note that, strictly speaking, the notation $\mathbf{P}_j^\theta(Y_t)[g(x)]$ in (2.7) needs to be replaced with $\mathbf{P}_j^\theta(Y_t, Y_{t-1})[g(x)]$, but we abuse the notation a bit here for convenience.

For given $i, j = 1, \ldots, d$, define the composition of two random functions as

$$\begin{aligned}
&\mathbf{P}_j^\theta(Y_{t+1}) \circ \mathbf{P}_i^\theta(Y_t)[g(x)] \\
&= \int_{x'' \in \mathcal{X}} p_j^\theta(x'', x) f(Y_{t+1}; \theta | j, x, Y_t) \\
&\quad \left(\int_{x' \in \mathcal{X}} p_i^\theta(x', x'') f(Y_t; \theta | i, x'', Y_{t-1}) g(x') m(dx') \right) m(dx''). \qquad (2.8)
\end{aligned}$$

It is straightforward to see that $\mathbf{M}$ defined in (2.5) forms a vector space with the standard scale product. *Addition* in $\mathbf{M}$ is defined as the addition of two functions. For $g \in \mathbf{M}$, denote $\|g\|_l := \int_{x \in \mathcal{X}} |g(x)| m(dx)$ as the L_1-norm on $\mathbf{M}$ with respect to m. Then, $(\mathbf{M}, \| \cdot \|_l)$ is a separable Banach space. Moreover, we define $\langle g \rangle_l := \int_{x \in \mathcal{X}} g(x) m(dx)$.

For a given vector $z = (z_1, \ldots, z_d)^\intercal \in \mathbf{R}^d$, define the L_1-norm of z as $\|z\|_d = \sum_{i=1}^d |z_i|$, and define $\langle z \rangle_d = \sum_{i=1}^d z_i$. Then, we define the L_1-norm of a $d \times d$ matrix $z = [z_{ij}]_{i,j=1,\ldots,d} \in \mathbf{R}^{d^2}$ as $\|z\|_d = \sum_{i,j=1}^d |z_{ij}|$. Denote

$$\mathbf{P}(Y_0) = \mathbf{P}^\theta(Y_0) = \operatorname{diag}(\mathbf{P}_1^\theta(Y_0), \ldots, \mathbf{P}_d^\theta(Y_0)) \qquad (2.9)$$

$$\mathbf{P}(Y_t) = \mathbf{P}^\theta(Y_t) = \begin{bmatrix} p_{11}\mathbf{P}_1^\theta(Y_t) & \cdots & p_{d1}\mathbf{P}_1^\theta(Y_t) \\ \vdots & \ddots & \vdots \\ p_{1d}\mathbf{P}_d^\theta(Y_t) & \cdots & p_{dd}\mathbf{P}_d^\theta(Y_t) \end{bmatrix}, \quad \text{for } t = 1, \ldots, n, \qquad (2.10)$$

and $\mathbf{M}^d := \{\psi = (\psi_1, \ldots, \psi_d) : \psi_j \in \mathbf{M}, \text{ for } j = 1, \ldots, d\}$. Then, $\mathbf{P}^\theta(Y_0)$ and $\mathbf{P}^\theta(Y_t)$ are random functions defined on $\mathcal{M} := (\mathcal{D} \times \mathcal{D} \times \mathcal{X} \times \mathbf{R}^p) \times \mathbf{M}^d$.

Now, for given $\mathbf{P}^\theta(Y_t)$ and $\mathbf{P}^\theta(Y_{t+1})$ in (2.10), define $\mathbf{P}^\theta(Y_{t+1}) \circ \mathbf{P}^\theta(Y_t)$ as

$$\begin{aligned}
&\mathbf{P}^\theta(Y_{t+1}) \circ \mathbf{P}^\theta(Y_t) \qquad (2.11) \\
&= \begin{bmatrix} \sum_{i=1}^d p_{i1}p_{1i}\mathbf{P}_1^\theta(Y_{t+1}) \circ \mathbf{P}_i^\theta(Y_t) & \cdots & \sum_{i=1}^d p_{i1}p_{di}\mathbf{P}_1^\theta(Y_{t+1}) \circ \mathbf{P}_i^\theta(Y_t) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^d p_{id}p_{1i}\mathbf{P}_d^\theta(Y_{t+1}) \circ \mathbf{P}_i^\theta(Y_t) & \cdots & \sum_{i=1}^d p_{id}p_{di}\mathbf{P}_d^\theta(Y_{t+1}) \circ \mathbf{P}_i^\theta(Y_t) \end{bmatrix}.
\end{aligned}$$

Note that the operation defined in (2.11) is in the domain of block operator matrices; see Tretter (2008).

Let $\pi_X(x) = (\pi_X(x|1), \ldots, \pi_X(x|d))^\intercal$. For given $t = 1, \ldots, n$, define

$$\mathbf{P}(Y_t) \circ \pi_X = \mathbf{P}(Y_t) \circ \pi_X(x) = \begin{bmatrix} p_{11}\mathbf{P}_1^\theta(Y_t)\pi_X(x|1) & \cdots & p_{d1}\mathbf{P}_1^\theta(Y_t)\pi_X(x|d) \\ \vdots & \ddots & \vdots \\ p_{1d}\mathbf{P}_d^\theta(Y_t)\pi_X(x|1) & \cdots & p_{dd}\mathbf{P}_d^\theta(Y_t)\pi_X(x|d) \end{bmatrix}, \qquad (2.12)$$

and

$$\mathbf{P}(Y_t) \circ \pi_X \circ \pi_H = \mathbf{P}(Y_t) \circ \pi_X \circ \pi_H(x) \qquad (2.13)$$

$$= \left(\sum_{i=1}^{d} \pi_H(i) p_{i1} \mathbf{P}_1^\theta(Y_t) \pi_X(x|i), \ldots, \sum_{i=1}^{d} \pi_H(i) p_{id} \mathbf{P}_d^\theta(Y_t) \pi_X(x|i) \right)^\top .$$

Define the norm $\|\cdot\|_{ld}$ of $\mathbf{P}(Y_t) \circ \pi_X \circ \pi_H$ as

$$\|\mathbf{P}(Y_t) \circ \pi_X \circ \pi_H\|_{ld} = \left\| \begin{bmatrix} \|\sum_{i=1}^{d} \pi_H(i) p_{i1} \mathbf{P}_1^\theta(Y_t) \pi_X(x|i)\|_l \\ \vdots \\ \|\sum_{i=1}^{d} \pi_H(i) p_{id} \mathbf{P}_d^\theta(Y_t) \pi_X(x|i)\|_l \end{bmatrix} \right\|_d .$$

Then, $p_n(Y_0, Y_1, \ldots, Y_n; \theta)$ in (2.4) can be represented as

$$p_n(Y_0, Y_1, \ldots, Y_n; \theta) = \|\mathbf{P}^\theta(Y_n) \circ \cdots \circ \mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}, \qquad (2.14)$$

where $\pi_H = \pi_H^\theta = (\pi_H^\theta(1), \ldots, \pi_H^\theta(d))^\top$ and $\pi_X = \pi_X^\theta = \pi_X^\theta(x)$, for $x \in \mathcal{X}$.

Therefore, by representation (2.14), $p_n(Y_0, Y_1, \ldots, Y_n; \theta)$ is the L_1-norm of a matrix-valued MIFS. Further detailed analysis is provided in Section 3. In addition, we define $\langle \cdot \rangle_{ld}$ of $\mathbf{P}(Y_t) \circ \pi_X \circ \pi_H$ as

$$\langle \mathbf{P}(Y_t) \circ \pi_X \circ \pi_H \rangle_{ld} = \left\langle \begin{bmatrix} \langle \sum_{i=1}^{d} \pi_H(i) p_{i1} \mathbf{P}_1^\theta(Y_t) \pi_X(x|i) \rangle_l \\ \vdots \\ \langle \sum_{i=1}^{d} \pi_H(i) p_{id} \mathbf{P}_d^\theta(Y_t) \pi_X(x|i) \rangle_l \end{bmatrix} \right\rangle_d .$$

Remark 1.

(1) Note that although we assume that the initial distribution in (2.4) is stationary, it can be arbitrary. This is because we do not need this assumption in the required theorems, such as Lemma 1 in the Supplementary Material for the stability issue, the strong law of large numbers for the induced matrix-valued MIFS (Fuh (2021)), and Theorem 2 and Corollary 1 in Fuh (2006) for the central limit theorem of the induced Markov chain.

(2) For hidden Markov models, which are a special case of the GMSMs studied in this paper, the likelihood function is usually expressed as product of conditional likelihood functions, $p(y_k|y_0, \ldots, y_{k-1})$, for $k = 1, \ldots, n$. Then, use $p(y_k|y_0, \ldots, y_{-\infty})$ to approximate $p(y_k|y_0, \ldots, y_{k-1})$ under some assumptions; for example, see Bickel, Ritov, and Rydén (1998) and Yonekura, Beskos, and Singh (2021). However, this approach is difficult to be applied to more general models, such as GMSMs. For GMSMs, we show that the MIFS approach works. That is, we find that both the likelihood function and the derivatives of the likelihood function can be expressed as matrix-valued MIFS, and that the MLE of a GMSM can be examined using the asymptotic properties of MIFS established in Fuh (2021).

3. The MLE

Let $y_0, y_1, \ldots, y_n$ be the observed values from the GMSM defined in (2.1) and (2.2). Then, the likelihood function $\mathcal{L}(\theta|y_0, y_1, \ldots, y_n)$ has the form $p_n = p_n(y_0, y_1, \ldots, y_n; \theta)$, defined in (2.4). When $\partial \log \mathcal{L}(\theta|y_0, y_1, \ldots, y_n)/\partial\theta$ exists, we can seek solutions to the likelihood equations

$$\frac{\partial \log \mathcal{L}(\theta|y_0, y_1, \ldots, y_n)}{\partial\theta} = 0,$$

and obtain the MLE $\hat{\theta}_n$ in a GMSM. Note that the MLE may not be unique.

To study the asymptotic properties of the MLE in a GMSM, we first impose some suitable conditions on the underlying Markov chain. Let $Z_t := ((H_t, X_t), Y_t)$ be an aperiodic and irreducible Markov chain on a general state space $(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p$ with Borel σ-algebra $\mathcal{A} := \mathcal{B}(\mathcal{D}) \times \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathbf{R}^p)$, where irreducibility is with respect to a maximal irreducible measure on $\mathcal{A}$. For the recurrent condition on the Markov chain, we first consider that $\{Z_t, t \geq 0\}$ is *Harris recurrent*, which is defined as follows: if there exists a set $A \in \mathcal{A}$, a probability measure Γ concentrates on A and an ε with $0 < \varepsilon < 1$ such that $P_z(Z_t \in A \ i.o.) = 1$, for all $z \in (\mathcal{D} \times \mathcal{X}) \times \mathbf{R}^p$; furthermore, there exists t, such that $P^t(z, C) \geq \varepsilon\Gamma(C)$, for all $z \in A$ and all $C \in \mathcal{A}$.

Next, we consider the w-uniformly ergodic condition. Let $w : (\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p \mapsto [1, \infty)$ be a measurable function, and let $\mathbf{B}$ be the Banach space of measurable functions $h : (\mathcal{D} \times \mathcal{X}) \times \mathbf{R}^p \mapsto \mathcal{C}$ (:= set of complex numbers), with $\|h\|_w := \sup_z |h(z)|/w(z) < \infty$. We impose the following conditions on the Markov chain $\{Z_t, t \geq 0\}$.

Assume Z_t has an invariant probability measure with probability density function $\pi := \pi_H(\cdot)\pi_X(\cdot|H)f(\cdot|H, X)$, such that $\int w(z)\pi(z)dz < \infty$, and for every $h \in \mathbf{B}$ satisfying $|h| \leq w$,

$$\lim_{t\to\infty} \sup_{z\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} \left\{ \frac{|E[h((H_t, X_t), Y_t)|((H_0, X_0), Y_0) = z] - \int h(z')\pi(z')dz'|}{w(z)} \right\} = 0, \tag{3.1}$$

$$\sup_{z\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} \left\{ \frac{E\big[w((H_1, X_1), Y_1)|((H_0, X_0), Y_0) = z\big]}{w(z)} \right\} < \infty. \tag{3.2}$$

Condition (3.1) states that the chain is w-uniformly ergodic, and implies that there exist $\gamma > 0$ and $0 < \rho < 1$ such that for all $h \in \mathbf{B}$ and $n \geq 1$,

$$\sup_{z\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} \frac{|E[h((H_t, X_t), Y_t)|((H_0, X_0), Y_0) = z] - \int h(z')\pi(z')dz'|}{w(z)} \leq \gamma\rho^t\|h\|_w;$$

see pages 382–383 and Proposition 16.1.3 of Meyn and Tweedie (2009). When $w \equiv 1$, this reduces to the classical uniformly ergodic condition. Note that for an aperiodic and irreducible Markov chain $\{(H_t, X_t), Y_t), t \geq 0\}$, the w-uniformly ergodic condition (3.1) implies that the Harris recurrent condition holds; see Theorem 9.1.8 of Meyn and Tweedie (2009).

For a given nonnegative integer vector $\nu = (\nu^{(1)}, \ldots, \nu^{(q)})^{\intercal}$, write $|\nu| = \nu^{(1)} + \cdots + \nu^{(q)}$, $\nu! = \nu^{(1)}! \cdots \nu^{(q)}!$, and let $D^{\nu} = (D_1)^{\nu^{(1)}} \cdots (D_q)^{\nu^{(q)}}$ denote the νth derivative with respect to θ in $N_\delta(\theta_0) := \{\theta : \|\theta - \theta_0\| \leq \delta\}$, the δ-neighborhood of the true parameter θ_0, where $(D_l)^k$ is the kth partial derivative with respect to the lth coordinate of θ for $l = 1, \ldots, q$, and $\| \cdot \|$ denotes the L_2-norm. Here, $\nu = 0$ denotes no derivative.

The following conditions are used throughout the rest of this paper.

C1. Stationary and ergodicity conditions

For any $\theta \in \Theta \subset \mathbf{R}^q$, the Markov chain $\{((H_t, X_t), Y_t), t \geq 0\}$ defined in (2.1) and (2.2) is aperiodic, irreducible, and satisfies (3.1) and (3.2), with weight function $w(\cdot)$.

C2. Identifiability condition

The true parameter θ_0 is an interior point of Θ, and the equality $p_n(y_0, y_1, \ldots, y_n; \theta) = p_n(y_0, y_1, \ldots, y_n; \theta')$ holds P-almost surely, for all nonnegative n, if and only if $\theta = \theta'$.

C3. Conditions on the state equation functions

C3.1. For all $j \in \mathcal{D}$ and $x, x' \in \mathcal{X}$, $\theta \mapsto p_j^{\theta}(x, x')$ and $\theta \mapsto \pi_X^{\theta}(x|j)$ are continuous. Furthermore, for all $j \in \mathcal{D}$ and $x \in \mathcal{X}$, $p_j^{\theta}(x, x') \to 0$ and $\pi_X^{\theta}(x|j) \to 0$ as $\|\theta\| \to \infty$, and for all $\theta \in \Theta$ and each $j \in \mathcal{D}$, $0 < p_j^{\theta}(x, x') < \infty$, for all $x, x' \in \mathcal{X}$, and $\sup_{x \in \mathcal{X}} \int p_j^{\theta}(x', x) m(dx') < \infty$.

C3.2. For all $j \in \mathcal{D}$ and $x, x' \in \mathcal{X}$, $\theta \mapsto p_j^{\theta}(x, x')$ and $\theta \mapsto \pi_X^{\theta}(x|j)$ have twice continuous derivatives in some neighborhood $N_\delta(\theta_0)$ of θ_0.

C3.3. For any $\theta \in N_\delta(\theta_0)$ and ν with $1 \leq |\nu| \leq 2$, assume for each $j \in \mathcal{D}$, $|D^{\nu} p_j^{\theta}(x, x')| < \infty$, for all $x, x' \in \mathcal{X}$.

C3.4. For all $j \in \mathcal{D}$, $x \in \mathcal{X}$, and $k_1, k_2 = 1, \ldots, q$,

$$\int_{\mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} \left| \frac{\partial \log \pi_X^{\theta}(x|j)}{\partial \theta_{k_1}} \right|^2 m(dx) < \infty, \quad \int_{\mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} \left| \frac{\partial \log p_j^{\theta}(x, x')}{\partial \theta_{k_1}} \right|^2 m(dx') < \infty,$$
$$\int_{\mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} \left| \frac{\partial^2 \log \pi_X^{\theta}(x|j)}{\partial \theta_{k_1} \partial \theta_{k_2}} \right| m(dx) < \infty, \quad \int_{\mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} \left| \frac{\partial^2 \log p_j^{\theta}(x, x')}{\partial \theta_{k_1} \partial \theta_{k_2}} \right| m(dx') < \infty.$$

For all $j \in \mathcal{D}$, $x \in \mathcal{X}$, $l = 1, 2$, and $k_1, k_2 = 1, \ldots, q$,

$$\int_{\mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} \left| \frac{\partial^l \pi_X^{\theta}(x|j)}{\partial \theta_{k_1} \cdots \partial \theta_{k_l}} \right| m(dx) < \infty, \quad \int_{\mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} \left| \frac{\partial^l p_j^{\theta}(x, x')}{\partial \theta_{k_1} \cdots \partial \theta_{k_l}} \right| m(dx') < \infty.$$

C4. Conditions on the observation equation functions

C4.1. For all $j \in \mathcal{D}$ and $x \in \mathcal{X}$, $\theta \mapsto f(y_0;\theta|j,x)$ and $\theta \mapsto f(y_1;\theta|j,x,y_0)$ are continuous, for all $y_0, y_1 \in \mathbf{R}^p$. Furthermore, for all $j \in \mathcal{D}$, $x \in \mathcal{X}$, and $y_0, y_1 \in \mathbf{R}^p$, $f(y_0;\theta|j,x) \to 0$ and $f(y_1;\theta|j,x,y_0) \to 0$ as $\|\theta\| \to \infty$.

C4.2. For all $\theta \in \Theta$ and each $j \in \mathcal{D}$, $0 < \sup_{x\in\mathcal{X}} f(y_0;\theta|j,x) < \infty$ and $0 < \sup_{x\in\mathcal{X}} f(y_1;\theta|j,x,y_0) < \infty$, for all $y_0, y_1 \in \mathbf{R}^p$. Because m is σ-finite, there exist pairwise disjoint $\{\mathcal{X}_n, n \geq 1\}$ such that $\mathcal{X} = \cup_{n=1}^{\infty}\mathcal{X}_n$ and $0 < m(\mathcal{X}_n) < \infty$. Assume $E\big[\sum_{n=1}^{\infty}(1/2^n)\sup_{j\in\mathcal{D},x\in\mathcal{X}_n} f(Y_1;\theta|j,x,y_0)\big] < \infty$, for all $y_0 \in \mathbf{R}^p$ and $\theta \in \Theta$.

Assume that there exists $r \geq 1$ such that, for $\theta \in \Theta \subset \mathbf{R}^q$ and $g \in \mathbf{M}$,

$$\sup_{((j,x_0),y_0)\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} E^{\theta}_{(j,x_0,y_0)}\left\{ \log\left(\mathbf{P}^{\theta}_j(Y_r)\circ\cdots\circ\mathbf{P}^{\theta}_j(Y_1)\circ\mathbf{P}^{\theta}_j(y_0)[g(x_0)] \times \frac{w(H_r,X_r,Y_r)}{w(j,x_0,y_0)}\right)\right\} < 0, \tag{3.3}$$

$$\sup_{((j,x_0),y_0)\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} E^{\theta}_{(j,x_0,y_0)}\left\{\mathbf{P}^{\theta}_j(Y_1)\circ\mathbf{P}^{\theta}_j(y_0)[g(x_0)]\frac{w(H_1,X_1,Y_1)}{w(j,x_0,y_0)}\right\} < \infty. \tag{3.4}$$

C4.3. For any $\theta \in N_\delta(\theta_0)$ and ν with $1 \leq |\nu| \leq 2$, $\sup_{j\in\mathcal{D},x\in\mathcal{X}} |D^\nu f(y_1;\theta|j,x,y_0)| < \infty$, for all $y_0, y_1 \in \mathbf{R}^p$. Assume that $E\big[\sum_{n=1}^{\infty}(1/2^n)\sup_{j\in\mathcal{D},x\in\mathcal{X}_n} |D^\nu f(Y_1;\theta|j,x,y_0)|\big] < \infty$, for all $y_0 \in \mathbf{R}^p$ and $\theta \in \Theta$.

Given $1 \leq |\nu| \leq 2$, assume that there exists $r \geq 1$ such that, for all $\theta \in N_\delta(\theta_0)$ and $g \in \mathbf{M}$, $\sup_{x\in\mathcal{X}} |\partial g(x)/\partial\theta_k| < \infty$, for $k = 1, \ldots, q$, and

$$\sup_{((j,x_0),y_0)\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} E^{\theta}_{(j,x_0,y_0)}\left\{ \log\left(\left|D^\nu\left(\mathbf{P}^{\theta}_j(Y_r)\circ\cdots\circ\mathbf{P}^{\theta}_j(Y_1)\circ\mathbf{P}^{\theta}_j(y_0)[g(x_0)]\right)\right| \times \frac{w(H_r,X_r,Y_r)}{w(j,x_0,y_0)}\right)\right\} < 0, \tag{3.5}$$

$$\sup_{((j,x_0),y_0)\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} E^{\theta}_{(j,x_0,y_0)}\left\{\left|D^\nu\left(\mathbf{P}^{\theta}_j(Y_1)\circ\mathbf{P}^{\theta}_j(y_0)[g(x_0)]\right)\right|\frac{w(H_1,X_1,Y_1)}{w(j,x_0,y_0)}\right\} < \infty. \tag{3.6}$$

C4.4. For all $j \in \mathcal{D}$, $x \in \mathcal{X}$, $y_0, y_1 \in \mathbf{R}^p$, and $\theta \in \Theta \subset \mathbf{R}^q$, and for $k_1, k_2, k_3 = 1, \ldots, q$, the partial derivatives $\partial f(y_0;\theta|j,x)/\partial\theta_{k_1}$, $\partial^2 f(y_0;\theta|j,x)/\partial\theta_{k_1}\partial\theta_{k_2}$, and $\partial^3 f(y_0;\theta|j,x)/\partial\theta_{k_1}\partial\theta_{k_2}\partial\theta_{k_3}$, and $\partial f(y_1;\theta|j,x,y_0)/\partial\theta_{k_1}$, $\partial^2 f(y_1;\theta|j,x,y_0)/\partial\theta_{k_1}\partial\theta_{k_2}$, and $\partial^3 f(y_1;\theta|j,x,y_0)/\partial\theta_{k_1}\partial\theta_{k_2}\partial\theta_{k_3}$ exist.

C4.5. For all $j \in \mathcal{D}$, $x \in \mathcal{X}$, $y_0 \in \mathbf{R}^p$, and $k_1, k_2 = 1, \dots, q$,

$$E^{\theta}_{(j,x)}\left[\sup_{\theta \in N_\delta(\theta_0)} \left|\frac{\partial \log f(Y_0;\theta|j,x)}{\partial \theta_{k_1}}\right|^2\right] < \infty,$$
$$E^{\theta}_{((j,x),y_0)}\left[\sup_{\theta \in N_\delta(\theta_0)} \left|\frac{\partial \log f(Y_1;\theta|j,x,y_0)}{\partial \theta_{k_1}}\right|^2\right] < \infty,$$
$$E^{\theta}_{(j,x)}\left[\sup_{\theta \in N_\delta(\theta_0)} \left|\frac{\partial^2 \log f(Y_0;\theta|j,x)}{\partial \theta_{k_1}\partial \theta_{k_2}}\right|\right] < \infty,$$
$$E^{\theta}_{((j,x),y_0)}\left[\sup_{\theta \in N_\delta(\theta_0)} \left|\frac{\partial^2 \log f(Y_1;\theta|j,x,y_0)}{\partial \theta_{k_1}\partial \theta_{k_2}}\right|\right] < \infty.$$

For all $j \in \mathcal{D}$, $x \in \mathcal{X}$, $y_0 \in \mathbf{R}^p$, $l = 1, 2$, and $k_1, k_2 = 1, \dots, q$,

$$\int \sup_{\theta \in N_\delta(\theta_0)} \left|\frac{\partial^l f(y;\theta|j,x)}{\partial \theta_{k_1}\cdots\partial \theta_{k_l}}\right| Q(dy) < \infty,$$
$$\int \sup_{\theta \in N_\delta(\theta_0)} \left|\frac{\partial^l f(y_1;\theta|j,x,y_0)}{\partial \theta_{k_1}\cdots\partial \theta_{k_l}}\right| Q(dy_1) < \infty.$$

C4.6. $E^{\theta_0}_{((j,x),y_0)}|\log(f(y_0;\theta_0|j,x)f(Y_1;\theta_0|j,x,y_0))| < \infty$, for all $j \in \mathcal{D}$ and $x \in \mathcal{X}$.

C4.7. For each $\theta \in \Theta$, there is a $\delta > 0$ such that for all $j \in \mathcal{D}$ and $x \in \mathcal{X}$, $E^{\theta_0}_{((j,x),y_0)}(\sup_{\|\theta'-\theta\|<\delta}\left[\log(f(y_0;\theta'|j,x)f(Y_1;\theta'|j,x,y_0))\right]^+) < \infty$, where $a^+ = \max\{a, 0\}$. Furthermore, there is a $b > 0$ such that, for all $j \in \mathcal{D}$ and $x \in \mathcal{X}$, $E^{\theta_0}_{((j,x),y_0)}(\sup_{\|\theta'\|>b}\left[\log(f(y_0;\theta'|j,x)f(Y_1;\theta')|j,x,y_0))\right]^+) < \infty$.

C4.8. For $\theta \in N_\delta(\theta_0)$,

$$\sup_{((j,x),y_0)\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} E^{\theta_0}_{((j,x),y_0)}\left(\sup_{\theta \in N_\delta(\theta_0)} \sup_{x,x' \in \mathcal{X}} \frac{f(y_0;\theta|j,x)f(Y_1;\theta|j,x,y_0)}{f(y_0;\theta|j,x')f(Y_1;\theta|j,x',y_0)}\right)^2 < \infty.$$

Remark 2.

(1) Condition C1 is the stationary and w-uniform ergodicity condition for the underlying Markov chain. In practice, $\{H_t, t \geq 0\}$ is often a finite-state ergodic Markov chain, and $\{Y_t, t \geq 0\}$ are conditionally independent for given $\{H_t, t \geq 0\}$ and $\{X_t, t \geq 0\}$. Then, we need only check w-uniform ergodicity for $\{X_t, t \geq 0\}$. Note that for the switching linear state space model in Example 1, X_t is an autoregressive model with $w(x) = \|x\|^2$; see Theorem 16.5.1 of Meyn and Tweedie (2009). Additional examples are provided in the Supplementary Material.

(2) Condition C2 is the identifiability condition for a GMSM. That is, the family of mixtures of $\{f(Y_1;\theta|j,x,y_0) : \theta \in \Theta\}$ is identifiable. We also use this condition to prove the strong consistency of the MLE. Although it is difficult

to check this condition in a GMSM, in many models of interest, such as a finite-state hidden Markov model with normal distributions, the parameter itself is identifiable only up to a permutation of states. A sufficient condition for the identifiability in hidden Markov models can be found in Douc et al. (2011).

(3) C3 states conditions on the state equation functions, where C3.1 is a standard continuity condition and C3.2–4 are standard smoothness conditions. These conditions are fulfilled in many practical models, such as switching linear Gaussian state space models.

(4) C4 states conditions on the observation equation functions. C4.1 is a standard continuity condition. In C4.2–3, we impose the weighted local mean contractive conditions (3.3) and (3.5) and the weighted mean moment conditions (3.4) and (3.6), to guarantee that the MIFS induced by the likelihood function of the GMSM and its derivatives, respectively, satisfy K2 and K3 in Section 4 of Fuh (2006). Note that (3.3) is a weaker condition than C1 in Fuh (2006). C4.4–5 are standard smoothness conditions. C4.6–7 are integrability conditions, which we use to prove the strong consistency of the MLE. C4.8 is a technical condition for the existence of the Fisher information to be defined in (3.11) below. In the Supplementary Material, we check that these conditions hold for several models used in practice.

Let $\{((H_t, X_t), Y_t), t \geq 0\}$ be the Markov chain defined in (2.1) and (2.2). Recall from (2.14) that the log likelihood function based on the samples $\{Y_0, Y_1, \ldots, Y_n\}$ can be written as

$$\begin{aligned} l(\theta) &= \log \mathcal{L}(\theta|Y_0, Y_1, \ldots, Y_n) = \log p_n(Y_0, Y_1, \ldots, Y_n; \theta) \\ &= \log \|\mathbf{P}^\theta(Y_n) \circ \cdots \circ \mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld} \\ &= \log \frac{\|\mathbf{P}^\theta(Y_n) \circ \cdots \circ \mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}}{\|\mathbf{P}^\theta(Y_{n-1}) \circ \cdots \circ \mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}} + \cdots \\ &\quad + \log \frac{\|\mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}}{\|\mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}} + \log \|\mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}. \end{aligned} \tag{3.7}$$

For each n, denote

$$M_n := \mathbf{P}^\theta(Y_n) \circ \cdots \circ \mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \tag{3.8}$$

as the matrix-valued MIFS on $\mathcal{M}^d$ induced from (2.5)–(2.13). Then, the log-likelihood function $l(\theta)$ based on the samples $\{Y_0, Y_1, \ldots, Y_n\}$ can be written as $S_n := \sum_{t=1}^n \phi(M_{t-1}, M_t) + \log \|\mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}$, with

$$\phi(M_{t-1}, M_t) := \log \frac{\|\mathbf{P}^\theta(Y_t) \circ \cdots \circ \mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}}{\|\mathbf{P}^\theta(Y_{t-1}) \circ \cdots \circ \mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}}. \tag{3.9}$$

To prove the strong consistency and asymptotic normality of the MLE in a GMSM under conditions C1–C4, we need to apply Lemma 1 in the Supplementary Material and Corollary 1 of Fuh (2006). For this purpose, we need to check that the induced matrix-valued MIFS satisfies the assumptions in Fuh (2006), and that the associated Markov chain is aperiodic, irreducible, and Harris recurrent.

To start with, for given $g \in \mathbf{M}$, we define the sup-norm of g as $\|g\|_\infty = \sup_{x\in\mathcal{X}} |g(x)| < \infty$. We also define the variation distance between any two elements g_1, g_2 in $\mathbf{M}$ by

$$d(g_1, g_2) = \sup_{x\in\mathcal{X}} |g_1(x) - g_2(x)|. \tag{3.10}$$

Note that $(\mathbf{M}, d)$ is a complete metric space with Borel σ-algebra $\mathcal{B}(\mathbf{M})$, but it is not separable. However, we can apply the results developed in Dudley (1966) for a nonseparable space. Therefore, Lemma 1 in the Supplementary Material and Theorems 1–4 of Fuh (2006) still hold under the regularity conditions. An alternative approach can be found in Section 7 of Diaconis and Freedman (1999), who provide a direct argument of convergence, rather than dealing with the measure-theoretic technicalities created by a nonseparable space.

Then, $\{((H_t, X_t, Y_t), M_t), t \geq 0\}$ is a Markov chain on the state space $\mathcal{M}_1 := (\mathcal{D} \times \mathcal{X} \times \mathbf{R}^p) \times \mathbf{M}^d$, with transition probability kernel $\mathbf{P}^\theta$ defined as (S3.2) in the Supplementary Material,

$$\mathbf{P}^\theta(((h_0, x_0, y_0), \psi), (A, B)) = \int_{(h_1,x_1,y_1)\in A} I_B(\mathbf{P}^\theta(y_1)\psi) P((h_0, x_0, y_0), d(h_1, x_1, y_1)),$$

for $h_0 \in \mathcal{D}$, $x_0 \in \mathcal{X}$, $y_0 \in \mathbf{R}^p$, $\psi \in \mathbf{M}^d$, $A \in \mathcal{A}$, and $B \in \mathcal{B}(\mathbf{M}^d)$, where I denotes the indicator function. In the Supplementary Material, under conditions C1–C4, we show that the stationary distribution of the Markov chain $\{((H_t, X_t, Y_t), M_t), t \geq 0\}$ exists, and is denoted as $\tilde{\Pi} := \tilde{\Pi}_\theta$.

In the following theorem, we state the strong consistency of the MLE $\hat{\theta}_n$ under some regularity conditions.

Theorem 1. *Assume conditions* C1, C2, C3.1*, and* C4.1,2,6,7 *hold. Then,* $\hat{\theta}_n \longrightarrow \theta_0$, P^{θ_0}*-a.s. as* $n \to \infty$.

To state the asymptotic normality of the MLE $\hat{\theta}_n$ in a GMSM, we need to define the Fisher information matrix

$$\begin{aligned} \mathbf{I}(\theta) &= (I_{lk}(\theta)) \\ &= \left(\mathbb{E}^\theta_{\tilde{\Pi}} \left[\left(\frac{\partial \log \|\mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}}{\partial \theta_l} \right) \right. \right. \\ &\qquad \left. \left. \left(\frac{\partial \log \|\mathbf{P}^\theta(Y_1) \circ \mathbf{P}^\theta(Y_0) \circ \pi_X \circ \pi_H\|_{ld}}{\partial \theta_k} \right) \right] \right), \end{aligned} \tag{3.11}$$

which is finite for θ in a neighborhood $N_\delta(\theta_0)$ of θ_0. Here, $\mathbb{E}^\theta_{\tilde{\Pi}}$ is the expectation under $\mathbb{P}^\theta_{\tilde{\Pi}}$, defined in (4.8) in Section 4. Furthermore, assume $\mathbf{I}(\theta_0)$ is invertible.

Theorem 2. *Assume conditions* C1–C4 *hold. Then,* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normally distributed with mean zero and variance-covariance matrix* $\mathbf{I}^{-1}(\theta_0)$.

Remark 3. In practice, although it is not easy to compute the MLE of a GMSM, we can approximate it. For example, for switching linear state space models, Kim (1994) provides a Kalman-filter-based approach for computing an approximation of the likelihood. Then, a nonlinear optimization procedure is used to compute the maximizer. This approach has been proved to perform well with a considerable advantage in terms of computation time. Ghahramani and Hinton (2000) propose a variational approximation method, similar to the EM algorithm, for computing the MLE.

4. Fisher Information and Score Function

To prove the strong consistency and asymptotic normality of the MLE $\hat{\theta}_n$ in a GMSM, we investigate the Kullback–Leibler divergence in Lemma 4 in the Supplementary Material, and the Fisher information in Theorem 3 below, which are of independent interest. The proof of the convergence of the score function and the Fisher information involves derivatives of the log likelihood function. Thus, we first show that the derivatives of the log likelihood function $l(\theta)$ in (3.7) can be expressed as an additive functional of a MIFS. Then, we can define the Fisher information and state the asymptotic normality of the score function. Note that the results in this section also fill the gap in the proofs of Lemmas 5 and 6 in Fuh (2006).

Recall $\mathbf{P}^\theta(Y_t)$ defined in (2.10) and $M_n = \mathbf{P}^\theta(Y_n)\circ\cdots\circ\mathbf{P}^\theta(Y_1)\circ\mathbf{P}^\theta(Y_0)$ defined in (3.8). For any $1 \le l \le q$ and positive integer k, recall that D_l is the partial derivative with respect to the lth coordinate of θ in a neighborhood $N_\delta(\theta_0)$ of the true parameter θ_0, and $(D_l)^k$ is the corresponding kth partial derivative. Now, for any two given random functions $\mathbf{P}^\theta_j(Y_{t+1})$ and $\mathbf{P}^\theta_j(Y_t)$, defined in (2.7), and for any given $g_\theta(\cdot) \in \mathbf{M}$, by conditions C1–C4 in Section 3 and the dominated convergence theorem, we have

$$D_l\left\{\mathbf{P}^\theta_j(Y_t)[g_\theta(x)]\right\} = D_l\left\{\int_{x'\in\mathcal{X}} p^\theta_j(x',x)f(Y_t;\theta|j,x,Y_{t-1})g_\theta(x')m(dx')\right\}$$
$$= \int_{x'\in\mathcal{X}}\Big\{f(Y_t;\theta|j,x,Y_{t-1})g_\theta(x')D_l p^\theta_j(x',x) + p^\theta_j(x',x)g_\theta(x')D_l f(Y_t;\theta|j,x,Y_{t-1})$$
$$+p^\theta_j(x',x)f(Y_t;\theta|j,x,Y_{t-1})D_l g_\theta(x')\Big\}m(dx'),$$

and

$$
\begin{aligned}
&D_l\left\{\mathbf{P}_j^\theta(Y_{t+1})\circ\mathbf{P}_i^\theta(Y_t)[g_\theta(x)]\right\}\\
&= D_l\biggl\{\int_{x''\in\mathcal{X}} p_j^\theta(x'',x)f(Y_{t+1};\theta|j,x,Y_t)\\
&\quad\left(\int_{x'\in\mathcal{X}} p_i^\theta(x',x'')f(Y_t;\theta|i,x'',Y_{t-1})g_\theta(x')m(dx')\right)m(dx'')\biggr\}\\
&= \int_{x''\in\mathcal{X}} D_l\left\{p_j^\theta(x'',x)f(Y_{t+1};\theta|j,x,Y_t)\right\}\\
&\quad\left(\int_{x'\in\mathcal{X}} p_i^\theta(x',x'')f(Y_t;\theta|i,x'',Y_{t-1})g_\theta(x')m(dx')\right)m(dx'')\\
&\quad+\int_{x''\in\mathcal{X}} p_j^\theta(x'',x)f(Y_{t+1};\theta|j,x,Y_t)\\
&\quad\left(\int_{x'\in\mathcal{X}} D_l\left\{p_i^\theta(x',x'')f(Y_t;\theta|i,x'',Y_{t-1})g_\theta(x')\right\}m(dx')\right)m(dx'')\\
&= \left\{D_l\mathbf{P}_j^\theta(Y_{t+1})\right\}\circ\mathbf{P}_i^\theta(Y_t)[g_\theta(x)]+\mathbf{P}_j^\theta(Y_{t+1})\circ\left\{D_l(\mathbf{P}_i^\theta(Y_t)[g_\theta(x)])\right\}.
\end{aligned}
$$

Denote

$$
D_l\mathbf{P}(Y_t) := D_l\mathbf{P}^\theta(Y_t) = \begin{bmatrix} D_l(p_{11}\mathbf{P}_1^\theta(Y_t)) & \cdots & D_l(p_{d1}\mathbf{P}_1^\theta(Y_t)) \\ \vdots & \ddots & \vdots \\ D_l(p_{1d}\mathbf{P}_d^\theta(Y_t)) & \cdots & D_l(p_{dd}\mathbf{P}_d^\theta(Y_t)) \end{bmatrix} \tag{4.1}
$$

$$
= \begin{bmatrix} D_l(p_{11})\mathbf{P}_1^\theta(Y_t) & \cdots & D_l(p_{d1})\mathbf{P}_1^\theta(Y_t) \\ \vdots & \ddots & \vdots \\ D_l(p_{1d})\mathbf{P}_d^\theta(Y_t) & \cdots & D_l(p_{dd})\mathbf{P}_d^\theta(Y_t)) \end{bmatrix} + \begin{bmatrix} p_{11}D_l(\mathbf{P}_1^\theta(Y_t)) & \cdots & p_{d1}D_l(\mathbf{P}_1^\theta(Y_t)) \\ \vdots & \ddots & \vdots \\ p_{1d}D_l(\mathbf{P}_d^\theta(Y_t)) & \cdots & p_{dd}D_l(\mathbf{P}_d^\theta(Y_t)) \end{bmatrix},
$$

for $t = 1,\ldots,n$. Note that p_{ij} may depend on θ, for $i,j = 1,\ldots,d$.

Although we use only the first two derivatives of the MIFS, we consider a general setting in the following arguments. For higher derivatives, we assume the corresponding assumptions in C3.2–4 and C4.3–5 hold, without specification. Recall that, for a given nonnegative integer vector $\nu = (\nu^{(1)},\ldots,\nu^{(q)})^\mathsf{T}$, we write $|\nu| = \nu^{(1)}+\cdots+\nu^{(q)}$ and $\nu! = \nu^{(1)}!\cdots\nu^{(q)}!$, and let $D^\nu = (D_1)^{\nu^{(1)}}\cdots(D_q)^{\nu^{(q)}}$ denote the νth derivative with respect to θ in $N_\delta(\theta_0)$. For any ν, define $W_n^\nu = D^\nu M_n = (D_1)^{\nu^{(1)}}\cdots(D_q)^{\nu^{(q)}}(M_n)$. Then, by conditions C1–C4 and the dominated convergence theorem, we have $D^\nu\|(M_n\circ\pi_X\circ\pi_H)\|_{ld} = \langle D^\nu(M_n\circ\pi_X\circ\pi_H)\rangle_{ld}$.

Now, let us consider all derivatives with order r or less. Note that for a fixed integer $r\geq 1$, there are exactly $K = (r+q)!/r!q!$ different ν satisfying $|\nu|\leq r$. Label all such ν by $\nu_1,\nu_2,\ldots,\nu_K$, and let $W_n = (W_n^{\nu_1},W_n^{\nu_2},\ldots,W_n^{\nu_K})^\mathsf{T}$. Recall $\mathcal{M} = (\mathcal{D}\times\mathcal{D}\times\mathcal{X}\times\mathbf{R}^p)\times\mathbf{M}^d$. Then, $W_n\in\mathcal{M}^K := \{v = (m_1,\ldots,m_K)^\mathsf{T} : m_k\in\mathcal{M}, 1\leq k\leq K\}$. Moreover, for given ν_l and ν_k, let $\nu_l+\nu_k$ denote componentwise addition in the vector.

To investigate the dynamic of W_n, note that for any ν_l, we have

$$\begin{aligned}
&W_n^{\nu_l} \\
&= D^{\nu_l}\big(\mathbf{P}^\theta(Y_n)\circ\cdots\circ\mathbf{P}^\theta(Y_1)\circ\mathbf{P}^\theta(Y_0)\big) \\
&= \sum_{\substack{1\le k\le m\le K\\ \nu_l=\nu_k+\nu_m}} \left\{\frac{(\nu_l)!}{(\nu_k)!(\nu_m)!} D^{\nu_m}\mathbf{P}^\theta(Y_n)\circ D^{\nu_k}\Big(\mathbf{P}^\theta(Y_{n-1})\circ\cdots\circ\mathbf{P}^\theta(Y_1)\circ\mathbf{P}^\theta(Y_0)\Big)\right\} \\
&= \sum_{\substack{1\le k\le m\le K\\ \nu_l=\nu_k+\nu_m}} \frac{(\nu_l)!}{(\nu_k)!(\nu_m)!}\left\{D^{\nu_m}\mathbf{P}^\theta(Y_n)\circ W_{n-1}^{\nu_k}\right\}.
\end{aligned} \tag{4.2}$$

Hence, we can denote a $K\times K$ matrix

$$A_n = [a_{lk}^n]_{1\le l,k\le K}, \tag{4.3}$$

with each $a_{lk}^n\in\mathcal{M}$ defined as

$$a_{lk}^n = \begin{cases} \frac{(\nu_l)!}{(\nu_k)!(\nu_m)!}D^{\nu_m}\mathbf{P}^\theta(Y_n), & \text{if exists } 1\le m\le K \text{ such that } \nu_l=\nu_k+\nu_m, \\ 0, & \text{otherwise.}\end{cases} \tag{4.4}$$

In addition, for each $K\times K$ $\mathcal{M}$-valued matrix $B=[b_{lk}]_{1\le l,k\le K}$ and each K-dimensional $\mathcal{M}$-valued vector $V=(V_1,V_2,\ldots,V_K)^\intercal\in\mathcal{M}^K$, we define

$$B\circ V := \left(\sum_{j=1}^K b_{1j}\circ V_j, \sum_{j=1}^K b_{2j}\circ V_j,\ldots,\sum_{j=1}^K b_{Kj}\circ V_j\right)^\intercal. \tag{4.5}$$

Then, by (4.2), we have $W_n = A_n\circ W_{n-1}$, and thus

$$W_n = A_n\circ A_{n-1}\circ\cdots\circ A_1\circ W_0, \tag{4.6}$$

where $W_0=\{W_0^\nu : |\nu|\le r\}$, with $W_0^\nu = D^\nu\mathbf{P}^\theta(Y_0)$.

Remark 4. To illustrate (4.6), let $q=1$, that is, θ is a one-dimensional parameter. In this case, $\nu\in\mathbf{R}$ and we can simply label all $|\nu|\le r$ by natural order so that $W_n=(W_n^0,W_n^1,\ldots,W_n^r)^\intercal$, the vector of the first rth derivatives. Then, for any $0\le k\le r$, we have

$$\begin{aligned}
W_n^k &= D^k(\mathbf{P}^\theta(Y_n)\circ\cdots\circ\mathbf{P}^\theta(Y_1)\circ\mathbf{P}^\theta(Y_0)) \\
&= \sum_{0\le k_1\le k}\left\{\frac{k!}{(k_1)!(k-k_1)!}D^{k_1}\mathbf{P}^\theta(Y_n)\right. \\
&\qquad \left.\circ D^{k-k_1}\Big(\mathbf{P}^\theta(Y_{n-1})\circ\cdots\circ\mathbf{P}^\theta(Y_1)\circ\mathbf{P}^\theta(Y_0)\Big)\right\} \\
&= \sum_{0\le k_1\le k} C_{k_1}^k\left\{D^{k_1}\mathbf{P}^\theta(Y_n)\circ W_{n-1}^{k-k_1}\right\},
\end{aligned}$$

where $C_a^b = b!/(a!(b-a)!)$. Therefore, $W_n = A_n \circ W_{n-1}$, with

$$A_n = \begin{bmatrix} \mathbf{P}^\theta(Y_n) & 0 & \cdots & 0 \\ C_1^1 D^1 \mathbf{P}^\theta(Y_n) & \mathbf{P}^\theta(Y_n) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_r^r D^r \mathbf{P}^\theta(Y_n) & C_{r-1}^r D^{r-1} \mathbf{P}^\theta(Y_n) & \cdots & \mathbf{P}^\theta(Y_n) \end{bmatrix}, \tag{4.7}$$

where zero denotes the zero function in $\mathcal{M}$. Note that W_n forms a MIFS on $\mathcal{M}^K$, and that the components in W_n can be different.

Note that W_n in (4.6) and A_n in (4.7) are $K \times K$ random matrices. In addition, for $k = 0, 1, \ldots, r$, the component $D^k \mathbf{P}^\theta(Y_n)$ in A_n is a $d \times d$ $\mathcal{M}$-valued matrix, rather than the traditional $\mathbf{R}$-valued vector and matrix. That is, $D^k \mathbf{P}^\theta(Y_n)$ is a $d \times d$ $\mathcal{M}$-valued random matrix in which each component is a random functional defined on $\mathbf{M}$.

To illustrate this phenomenon, we consider H_t as a finite d-state Markov chain and there is no X_t. Let θ be a one-dimensional parameter. Then, A_n in (4.7) is a $K \times K$ matrix, with each element being a $d \times d$ matrix (with zero being a $d \times d$ zero matrix), which can be regarded as a block matrix or partioned matrix; see Zhang (2011). In the same manner, although the operator defined in (4.5) looks like a traditional matrix multiplication, it replaces the multiplication within each component with $\circ$. Nevertheless, the essential idea is to have a matrix form for W_n, by which it constitutes a MIFS, from (4.6).

Note that obtaining a neat form in (4.6) is based on a matrix representation in (4.3) and (4.4), for all partial derivatives up to the rth order. Then, $\{((H_t, X_t, Y_t), W_t), t \geq 0\}$ is a Markov chain on the state space $\mathcal{M}_1^K := (\mathcal{D} \times \mathcal{X} \times \mathbf{R}^p) \times (\mathbf{M}^d)^K$, with transition probability kernel $\mathbb{P}^\theta$, defined in (S3.2) in the Supplementary Material,

$$\begin{aligned} &\mathbb{P}^\theta_{\tilde{\Pi}}(((h_0, x_0, y_0), \psi), (A, B)) \\ &= \int_{(h_1, x_1, y_1) \in A} I_B(W_1(\psi)) \; P((h_0, x_0, y_0), d(h_1, x_1, y_1)), \end{aligned} \tag{4.8}$$

for $h_0 \in \mathcal{D}$, $x_0 \in \mathcal{X}$, $y_0 \in \mathbf{R}^p$, $\psi \in (\mathbf{M}^d)^K$, $A \in \mathcal{B}(\mathcal{D}) \times \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathbf{R}^p)$, and $B \in \mathcal{B}((\mathbf{M}^d)^K)$.

We show in the Supplementary Material that, under conditions C1–C4, for $\theta \in N_\delta(\theta_0)$, the MIFS W_n in (4.6) satisfies Assumption K in Fuh (2006). Using this result and the result that the νth derivatives of the log likelihood function can be written as an additive functional of the Markov chain $\{((H_t, X_t, Y_t), W_t), t \geq 0\}$ in Lemma 5 in the Supplementary Material, we have the strong law of large numbers for the observed Fisher information. Then, we characterize the Fisher information matrix in Theorem 3, and state the asymptotic normality of the score function in Theorem 4.

Theorem 3. *Assume conditions* C1–C4 *hold. Then, for $\theta \in N_\delta(\theta_0)$, we have that as $n \to \infty$,*

$$\frac{1}{n}\frac{\partial^2}{\partial\theta_l\partial\theta_k}\log\|\mathbf{P}^\theta(Y_n)\circ\cdots\circ\mathbf{P}^\theta(Y_1)\circ\mathbf{P}^\theta(Y_0)\circ\pi_X\circ\pi_H\|_{ld}\to -I_{lk}(\theta), \quad (4.9)$$

with probability one, where $I_{lk}(\theta)$ is defined in (3.11) *and is finite for θ in a neighborhood $N_\delta(\theta_0)$ of θ_0. Recall that $\mathbf{I}(\theta) = (I_{lk}(\theta))$ is the Fisher information matrix.*

Theorem 4. *Assume conditions* C1–C4 *hold. Let $l'_k(\theta_0) = \partial l(\theta)/\partial\theta_k|_{\theta=\theta_0}$. Then, as $n \to \infty$,*

$$\frac{1}{\sqrt{n}}(l'_1(\theta_0),\ldots,l'_q(\theta_0))^T \longrightarrow N(0,\mathbf{I}(\theta_0)) \quad \textit{in distribution.}$$

5. Conclusion

We provide a GMSM, which includes many practically used models as special cases. In this framework, the hidden unit can be one or two layers, and can be a linear (or nonlinear) predictable (or stochastic) function of past information. This can be viewed as a Markov model if we include all hidden units. Furthermore, by using a matrix-valued MIFS representation of the likelihood function, we prove the strong consistency and asymptotic normality of the MLE in a GMSM under a weighted local mean contractive property. It is easy to check that the (switching) linear state space models, (switching) GARCH(p, q) models, (switching) SV models, and variational RNNs satisfy these conditions under some commonly used assumptions.

Using this framework, it would be interesting to explore the asymptotic properties, including the strong consistency, asymptotic normality, and even high-order asymptotics, of other commonly used estimators, such as the GMM, Bayesian estimators, and generalized empirical likelihood estimator.

Supplementary Material

The Supplementary Material includes examples of GMSM, a simulation study, and proofs for the theorems presented here.

Acknowledgments

The authors would like to thank the editor, professor Rong Chen, and the anonymous referee for their constructive and helpful comments and suggestions. The research of Tianxiao Pang was partially supported by the National Social Science Foundation of China (No. 21BTJ067).

References

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37**, 1554–1563.

Bickel, P., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.* **26**, 1614–1635.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econometrics* **31**, 307–327.

Cai, J. (1994). A Markov model of switching-regime ARCH. *J. Business & Econom. Statist.* **12**, 309–316.

Calvet, L. E. and Fisher, A. J. (2001). Forecasting multifractal volatility. *J. Econometrics* **105**, 27–58.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C. and Bengio, Y. A. (2015). A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, 2980–2988.

Davig, T. and Doh, T. (2014). Monetary policy regime shifts and inflation persistence. *The Review of Economics and Statistics* **96**, 862–875.

Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM Review* **41**, 45–76.

Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli* **7**, 381–420.

Douc, R., Moulines, É., Losson, J. and Handel, R. V. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.* **39**, 474–513.

Douc, R., Moulines, É. and Rydén, T. (2004). Asymptotics properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32**, 2254–2304.

Dudley, R. M. (1966). Weak convergence of probabilities on non-separable metric spaces and empirical measures on Euclidean spaces. *Illinois J. Math.* **10**, 109–126.

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D. C. and Ravindranath, S. S. (2019). Optimal auctions through deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, **PMLR 97**. Long Beach.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50**, 987–1008.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

Francq, C. and Roussignol, M. (1998). Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum likelihood estimator. *Statistics* **32**, 151–173.

Fuh, C.-D. (2003). SPRT and CUSUM in hidden Markov models. *Ann. Statist.* **31**, 942–977.

Fuh, C.-D. (2004). On Bahadur efficiency of the maximum likelihood estimator in hidden Markov models. *Statist. Sinica* **14**, 127–144.

Fuh, C.-D. (2006). Efficient likelihood estimation in state space models. *Ann. Statist.* **34**, 2026–2068. Corrigendum in **38**, 1279–1285.

Fuh, C.-D. (2021). Asymptotic behavior for Markovian iterated function systems. *Stoch. Proc. Appl.* **138**, 186–211.

Fuh, C. D. and Mei, Y. (2015). Quickest change detection and Kullback-Leibler divergence for two-state hidden Markov models. *IEEE Transactions on Signal Processing* **63**, 4866-4878.

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural Comput.* **12**, 831–864.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Mach. Learn.* **29**, 245–273.

Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions. *J. Econometrics* **1**, 3–16.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning.* MIT Press, Cambridge.

Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies* **33**, 2223–2273.

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

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton University Press, New Jersey.

Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *J. Econometrics* **64**, 307–333.

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. *arXiv:1612.09596.*

Jensen, J. L. and Petersen, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist.* **27**, 514–535.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *J. Econometrics* **60**, 1–22.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of 3rd International Conference on Learning Representations.*

Kuan, C.-M. and White, H. (1994). Artificial neural networks: An econometric perspective (with discussions). *Econometric Reviews* **13**, 1–91.

Leroux, B. G. (1992). Maximum likelihood estimation for hidden Markov models. *Stoch. Proc. Appl.* **40**, 127–143.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability.* 2nd Edition. Cambridge University Press, Cambridge.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1987). Learning internal representation by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, 318–362.

Sirignano, J. A. (2019). Deep learning for limit order books. *Quantitative Finance* **19**, 549–570.

So, M. K. P., Lam, K. and Li, W. K. (1998). A stochastic volatility model with Markov switching. *J. Business & Econom. Statist.* **16**, 244–253.

Taylor, S. J. (1986). *Modeling Financial Time Series.* John Wiley & Sons, Chichester.

Tretter, C. (2008). *Spectral Theory of Block Operator Matrices and Applications.* Imperial College Press, London.

Verstyuk, S. (2020). Modeling multivariate time series in economics: From auto-regressions to recurrent neural networks. Retrieved from `http://www.verstyuk.net/papers/VARMRNN.pdf`.

White, H. (1988). Economic prediction using neural networks: The case of IBM stock prices. In *Proceedings of the Second Annual IEEE Conference on Neural Networks II*, 451–458.

White, H. (1989). Some asymptotic results for learning in single hidden layer feedforward network models. *J. Amer. Statist. Assoc.* **84**, 1003–1013.

Yonekura, S., Beskos A. and Singh, A. A. (2021). Asymptotic analysis of model selection criteria for general hidden Markov models. *Stochastic Processes and their Applications* **132**, 164–191.

Zhang, F. (2011). *Matrix Theory: Basic Results and Techniques.* 2nd Edition. Springer-Verlag, New York.

Cheng-Der Fuh
Graduate Institute of Statistics, National Central University, Taoyuan 320317, Taiwan.
E-mail: cdffuh@gmail.com
Tianxiao Pang
School of Mathematical Sciences, Zhejiang University, Hangzhou 310058, China.
E-mail: txpang@zju.edu.cn

(Received September 2021; accepted October 2022)