Statistica Sinica Preprint No: SS-2021-0336	
Title	Asymptotic Behavior of the Maximum Likelihood Estimator for General Markov Switching Models
Manuscript ID	SS-2021-0336
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0336
Complete List of Authors	Cheng-Der Fuh and Tianxiao Pang
Corresponding Authors	Tianxiao Pang
E-mails	txpang@zju.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Asymptotic Behavior of the Maximum Likelihood Estimator for General Markov Switching Models

Cheng-Der Fuh and Tianxiao Pang

Zhejiang University City College and Zhejiang University

Abstract: Motivated by studying asymptotic properties of the parameter estimator in switching linear state space models, switching GARCH models, switching stochastic volatility models, and recurrent neural networks, in this paper, 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 above-mentioned MIFS as a composition of random matrices. To the best of our knowledge, it is a new method in the literature. Through this useful device, the strong consistency and asymptotic normality of the maximum likelihood estimator are established 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 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), in this paper, 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 $\mathcal{D} = \{1, \cdots, d\}$, 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 $\{H_t, t \geq 0\}$, $\{\eta_t, t \geq 1\}$, and $\{\varepsilon_t, t \geq 1\}$ are independent. GMSM is very flexible and includes the previously mentioned 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 $\{Y_t, t \geq 0\}$, GMSM is reduced to the following

well-known switching linear state space model:

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

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

cf. 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 $\mathbf{X}$. Y_t itself is generally 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 Markov switching model. In this paper, we consider a general Markov switching model, to 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 paper is to give a theoretical justification for the M-

LE in GMSM. It is known that a major difficulty of analyzing the likelihood function in GMSM is that the likelihood function can only be expressed in recursive integral form; see Equation (2.4) below for instance. In this paper, we use the device in (2.5)-(2.13) below, to represent the probability density, as well as the likelihood function, in (2.4) as the L_1-norm of a matrix-valued Markovian iterated function system (MIFS), and then the log likelihood function can be written as an additive form in (3.7), in which standard argument of the likelihood function for the 'enlarged' Markov chain can be applied. 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, cf. Fuh and Mei (2015). It has potential to provide a fast algorithm for the evaluation of the likelihood function via EM algorithm. Note that the asymptotic behaviors of MIFS have been studied in detail by Fuh (2021) recently. This new device enables us to apply the results of strong law of large numbers and central limit theorem for the asymptotic distributions of the matrix-valued MIFS, also to verify the strong consistency and asymptotic normality of the MLE in GMSM.

In the following paragraphs, we give a brief summary of the literature regarding GMSM. Note that the 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 a great deal of attention because of its importance in applications to speech recognition, signal processing, ion channels, molecular biology, and others. When $\mathbf{Y}$ forms an autoregression model for given $\mathbf{H}$, the GMSM reduces to the Markov switching model, cf. Hamilton (1989), and the Markov switching multifractal models, cf. Calvet and Fisher (2001). When there is only $\mathbf{X}$ and no hidden state $\mathbf{H}$, the GMSM includes the celebrated (G)ARCH models, cf. Engle (1982) and Bollerslev (1986), SV models, cf. Taylor (1986), and RNNs, cf. Goodfellow, Bengio and Courville (2016). The reader is referred to Hamilton (1994) and Fan and Yao (2003) for a comprehensive summary.

When there are two-layer hidden states $\mathbf{H}$ and $\mathbf{X}$, as mentioned before, the GMSM includes the switching linear state space model, cf. Kim (1994) and Ghahramani and Hinton (2000), switching GARCH models, cf. Cai (1994), and Hamilton and Susmel (1994), switching SV models, cf. So, Lam and Li (1998), and variational RNNs, cf. 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 factorical hidden Markov model, cf. Ghahramani and Jordan (1997). The main focus of these efforts has been state space modeling and estimation, algorithms for

fitting these models, and the implementation of likelihood-based methods. For instance, Kim (1994) and Ghahramani and Hinton (2000) propose a Kalman-filter-based method and a variational approximation method for the implementation of the MLE in switching linear state space models, respectively, and Davig and Doh (2014) estimate new Keynesian general equilibrium models with switching monetary policy rules.

It is known that RNNs, cf. Goodfellow, Bengio and Courville (2016), are a popular modeling choice for solving sequence learning problems in machine learning. Early applications of RNN models in econometrics can be found in Kuan and White (1994) and White (1988), among others. For recent approaches, artificial neural networks have been used for auction design, cf. Dütting et al. (2017), for estimation of causal relationships developing the broad idea of instrumental variables, cf. Hartford et al. (2016), for portfolio theory in finance, cf. Sirignano (2019) and Gu, Kelly and Xiu (2020), and for time series, cf. Verstyuk (2020). Due to the model complexity, gradient descent and/or stochastic gradient descent are used to compute the MLE in most econometrics and machine learning literature. For instance, Rumelhart, Hinton and Williams (1987) propose a recursive algorithm (backpropagation learning) to speed up the gradient descent method, whereas White (1989) establishes the consistency and asymptotic normality

of the algorithm. Adam (adaptive moment estimation) is a recent popular adaptive gradient algorithm used in machine learning, cf. Kingma and Ba (2015).

There is an extensive literature on the MLE in a special case of 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, while 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) consider the implementation of the MLE in switching autoregression with Markov regimes. Francq and Roussignol (1998) and Douc, Moulines and Rydén (2004) study the consistency and asymptotic normality of the MLE in Markov-switching autoregressive models, respectively, and Fuh (2004) establishes the Bahadur efficiency of the MLE in Markov switching models. 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).

There are three contributions in this paper. First, we provide a probability framework for the GMSM, which includes hidden Markov models, Markov switching models, (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, then express the derivatives of the MIFS as a composition of random matrices. This is a new method in the literature as far as we know. 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. In the meantime, these results can be applied to Markov switching

models, non-linear state space models and stochastic volatility models as well.

The remainder of this paper is organized as follows. In Section 2, we define the GMSM formally and represent its likelihood function as the L_1-norm of a matrix-valued MIFS. Section 3 investigates the MLE in GMSM, and states the main results. Section 4 studies the derivatives of the matrix-valued MIFS and the score function, and then characterizes the Fisher information. Section 5 concludes. In Section S1 in the online 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 commonly used in econometrics and machine learning. A simulation study and all technical proofs are given in Section S2 and Section S3 in the online supplementary material, respectively.

2. General Markov Switching Models

It is known that a GMSM is not Markovian in general. However, in this section, we will provide a probability framework for the GMSM, under which it can be regarded as a Markov chain in an enlarged state space. Along this line, there are two Markov chain representations for the GMSM

to be described as follows. 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 $\mathcal{D} = \{1, \cdots, d\}$, with transition probability $p_{ij}^{\theta} = P^{\theta}\{H_1 = j | H_0 = i\}$ and stationary probability $\pi_H^{\theta}(\cdot)$. 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 $P_j^{\theta}(x, \cdot) = P^{\theta}\{X_1 \in \cdot | H_1 = j, X_0 = x\}$ and stationary probability $\pi_X^{\theta}(\cdot | H_0 = j)$, 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

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

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

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

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 on (H_{t+1}, X_{t+1}) and Y_t only. Furthermore, we assume the existence of a transition probability density $p_j^\theta(x, x')$ 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 $i, j \in \mathcal{D}$,

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

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 $\theta = (\theta_1, \cdots, \theta_q)^\mathsf{T} \in \Theta \subseteq \mathbf{R}^q$ 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 will 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 $f(y_k; \theta | H_k, X_k, Y_{k-1})$, respectively, here and in the sequel depending on our convenience.

Now we give a formal definition of the GMSM as follows.

Definition 1. $\{Y_t, t \geq 0\}$ is called a general Markov switching model (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 which satisfies (2.1) and (2.2).

To have the first Markov chain representation of the likelihood function for the GMSM, we first 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 $\{Y_t, t = 0, \cdots, n\}$ is

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

where

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

To illustrate the GMSM, we recall the switching linear state space model (1.3) and (1.4). Other examples, including the switching GARCH models, switching SV models, and variational RNNs, will be given in the online supplementary material.

Example 1. (Switching linear state space models). Consider the model in (1.3) and (1.4) with $X_0 = \mathbf{0}$ being replaced by 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 governed by $\{H_t, t \geq 0\}$, respectively. 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. Abuse the notation a little bit, 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 $j = 1, \cdots, d$, 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 $\mathbf{y} = (y_1, \cdots, y_n)$ from the switching linear state space model (1.3) and (1.4), the likelihood function of the parameter $\theta = (\alpha_1, \cdots, \alpha_d, \beta_1, \cdots, \beta_d, \mu, \sigma^2)^\mathsf{T}$ is

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

where $\phi_{\mu,\sigma^2}(\cdot)$ is the probability density function of $N(\mu, \sigma^2)$. More details will be given in Section S1 in the online supplementary material.

To have the second Markov chain representation for the GMSM in (2.3) and (2.4), which will be used to analyze the MLE of the GMSM, we first write the random joint probability density function $p_n(Y_0, Y_1, \cdots, Y_n; \theta)$ as the L_1-norm of a composition of Markovian random matrices, in which each component in the random matrix is a Markovian iterated random function. To be more precise, let

$$\mathbf{M} = \left\{ g | g : \mathcal{X} \mapsto \mathbf{R} \text{ is } m-\text{measurable}, \int |g(x)| m(dx) < \infty \text{ and } \sup_{x \in \mathcal{X}} |g(x)| < \infty \right\} \quad (2.5)$$

For each $t = 1, \cdots, n$ and $j = 1, \cdots, d$, 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

$$\mathbf{P}_j^\theta(Y_0)[g(x)] = \int_{x \in \mathcal{X}} f(Y_0; \theta | j, x) g(x) m(dx), \quad (2.6)$$

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

Note that $\mathbf{P}_j^\theta(Y_0)[g(x)]$ defined in (2.6) is a function which maps any $g(x)$ to a constant/random variable depending on Y_0. 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, cf. (1.5) in Fuh (2003). Note also that, strictly speaking, the no-

tation $\mathbf{P}_j^\theta(Y_t)[g(x)]$ in (2.7) needs to be replaced by $\mathbf{P}_j^\theta(Y_t, Y_{t-1})[g(x)]$, but we abuse the notation a little bit here for our convenience.

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

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

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

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

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

and $\mathbf{M}^d := \{\psi = (\psi_1, \cdots, \psi_d) : \psi_j \in \mathbf{M}, \text{ for } j = 1, \cdots, d\}$. 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, cf. Tretter (2008).

Let $\pi_X(x) = (\pi_X(x|1), \cdots, \pi_X(x|d))^\mathsf{T}$. For given $t = 1, \cdots, n$, define

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

and

$$\begin{aligned} &\mathbf{P}(Y_t) \circ \pi_X \circ \pi_H = \mathbf{P}(Y_t) \circ \pi_X \circ \pi_H(x) \qquad (2.13)\\ &= \left(\sum_{i=1}^d \pi_H(i)p_{i1}\mathbf{P}_1^\theta(Y_t)\pi_X(x|1), \cdots, \sum_{i=1}^d \pi_H(i)p_{id}\mathbf{P}_d^\theta(Y_t)\pi_X(x|d) \right)^\mathsf{T}. \end{aligned}$$

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

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

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

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

here $\pi_H=\pi_H^{\theta}=(\pi_H^{\theta}(1),\cdots,\pi_H^{\theta}(d))^{\mathsf{T}}$ and $\pi_X=\pi_X^{\theta}=\pi_X^{\theta}(x)$ for $x\in\mathcal{X}$.

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

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

Remark 1. (1) It is worth mentioning that albeit the initial distribution in (2.4) is assumed to be the stationary distribution, the initial distribution indeed can be arbitrary. This is due to that we do not need this assumption in the required theorems, such as Lemma 1 in the online supplementary material for the stability issue and the strong law of large numbers of the induced matrix-valued MIFS (cf. 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 is a special case of GMSMs studied in this paper, the likelihood function is usually expressed as product of conditional likelihood functions, $p(y_k|y_0,\cdots,y_{k-1})$ for $k=1\cdots,n$, in the literature. Then use $p(y_k|y_0,\cdots,y_{-\infty})$ to approximate $p(y_k|y_0,\cdots,y_{k-1})$ 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 approach of MIFS 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 the MLE of GMSM can be examined through the asymptotic properties of MIFS established in Fuh (2021).

3. Maximum Likelihood Estimator

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

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

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

To study the asymptotic properties of the MLE in GMSM, we first impose some suitable conditions for 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* to be 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 $z \in (\mathcal{D} \times \mathcal{X}) \times \mathbf{R}^p$ and 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 shall 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 $\int w(z)d\pi(z) < \infty$, and for every $h \in \mathbf{B}$ satisfying $|h| \leq w$,

$$\lim_{t\to\infty} \sup_{z\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p} \left\{ \left| E[h((H_t, X_t), Y_t)|((H_0, X_0), Y_0) = z] - \int h(y)d\pi(y) \right| /w(z) \right\} = 0 \quad (3.1)$$

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

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

$$\sup_{z\in(\mathcal{D}\times\mathcal{X})\times\mathbf{R}^p}\left|E[h((H_t,X_t),Y_t)|((H_0,X_0),Y_0)=z]-\int h(y)d\pi(y)\right|/w(z)\leq\gamma\rho^t\|h\|_w,$$

cf. 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 $\{(H_t,X_t),Y_t),t\geq 0\}$, w-uniformly ergodic condition (3.1) implies that Harris recurrent condition holds, cf. Theorem 9.1.8 of Meyn and Tweedie (2009).

For a given non-negative integer vector $\nu=(\nu^{(1)},\cdots,\nu^{(q)})^{\mathsf{T}}$, 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 ν-th derivative with respect to θ in $N_\delta(\theta_0):=\{\theta:\|\theta-\theta_0\|<\delta\}$, the δ-neighbourhood of the true parameter θ_0, where $(D_l)^k$ is the k-th partial derivative with respect to the l-th coordinate of θ for $l=1,\cdots,q$, and $\|\cdot\|$ stands for the L_2-norm. Here $\nu=0$ denotes no derivative.

The following conditions will be 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 $p_n(y_0, y_1, \cdots, y_n; \theta) = p_n(y_0, y_1, \cdots, y_n; \theta')$ holds P-almost surely, for all non-negative n, if and only if $\theta = \theta'$.

C3: Conditions on the state equation functions

C3(1). For all $j \in \mathcal{D}$, $x, x' \in \mathcal{X}$, $\theta \mapsto p_j^\theta(x, x')$ 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 $j \in \mathcal{D}$, $x, x' \in \mathcal{X}$, $\theta \mapsto p_j^\theta(x, x')$ 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 $1 \le |\nu| \le 2$, 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 $k_1, k_2 = 1, \cdots, q$,

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

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

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

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

C4: Conditions on the observation equation functions

C4(1). For all $j \in \mathcal{D}$, $x \in \mathcal{X}$, $\theta \mapsto f(y_0;\theta|j,x)$ as well as $\theta \mapsto f(y_1;\theta|j,x,y_0)$ 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 $y_0, y_1 \in \mathbf{R}^p$. Since 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 $E\big[\sum_{n=1}^{\infty}\frac{1}{2^n}\sup_{j\in\mathcal{D},x\in\mathcal{X}_n} f(Y_1;\theta|j,x,y_0)\big] < \infty$ 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 $E\big[\sum_{n=1}^{\infty}\frac{1}{2^n}\sup_{j\in\mathcal{D},x\in\mathcal{X}_n} |D^\nu f(Y_1;\theta|j,x,y_0)|\big] < \infty$ 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 $g \in \mathbf{M}$, $\sup_{x\in\mathcal{X}} \left|\frac{\partial g(x)}{\partial \theta_k}\right| < \infty$ for $k = 1, \cdots, q$, 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 $j \in \mathcal{D}$, $x \in \mathcal{X}$, $y_0, y_1 \in \mathbf{R}^p$, $\theta \in \Theta \subset \mathbf{R}^q$, and for $k_1, k_2, k_3 = 1, \cdots, q$, the partial derivatives $\frac{\partial f(y_0;\theta|j,x)}{\partial\theta_{k_1}}$, $\frac{\partial^2 f(y_0;\theta|j,x)}{\partial\theta_{k_1}\partial\theta_{k_2}}$ and $\frac{\partial^3 f(y_0;\theta|j,x)}{\partial\theta_{k_1}\partial\theta_{k_2}\partial\theta_{k_3}}$ exist, as well as the partial derivatives $\frac{\partial f(y_1;\theta|j,x,y_0)}{\partial\theta_{k_1}}$, $\frac{\partial^2 f(y_1;\theta|j,x,y_0)}{\partial\theta_{k_1}\partial\theta_{k_2}}$ and $\frac{\partial^3 f(y_1;\theta|j,x,y_0)}{\partial\theta_{k_1}\partial\theta_{k_2}\partial\theta_{k_3}}$ exist.

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

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

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

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

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 $x \in \mathcal{X}$, $E^{\theta_0}_{((j,x),y_0)}\left(\sup_{\|\theta'-\theta\|<\delta}\left[\log(f(y_0;\theta'|j,x)f(Y_1;\theta'|j,x,y_0))\right]^+\right) < \infty$,

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

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. For many practical examples, $\{H_t, t \geq 0\}$ is a finite state ergodic Markov chain, $\{Y_t, t \geq 0\}$ are conditionally independent for given $\{H_t, t \geq 0\}$ and $\{X_t, t \geq 0\}$. Then we only need to 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 $w(x) = \|x\|^2$, cf. Theorem 16.5.1 of Meyn and Tweedie (2009). Other examples will be given in the online supplementary material.

(2) Condition C2 is the identifiability condition for GMSM. That is, the family of mixtures of $\{f(Y_1;\theta|j,x,y_0) : \theta \in \Theta\}$ is identifiable. This condition will also be used to prove the strong consistency of the MLE. Although it is difficult to check this condition in 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 is the 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 practically used models such as switching linear Gaussian state space models.

(4) C4 is the 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), respectively, to guarantee that the MIFS induced by the likelihood function of the GMSM and its derivatives 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 will be used 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. We will check these conditions hold for several practically used models in the online supplementary material.

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 $\{Y_0, Y_1, \cdots, Y_n\}$ can be written as

$$l(\theta) \quad = \quad \log \mathcal{L}(\theta | Y_0, Y_1, \cdots, Y_n) = \log p_n(Y_0, Y_1, \cdots, Y_n; \theta) \tag{3.7}$$

$$
\begin{aligned}
&= \log \|\mathbf{P}^{\theta}(Y_n) \circ \cdots \circ \mathbf{P}^{\theta}(Y_1) \circ \mathbf{P}^{\theta}(Y_0) \circ \pi_X \circ \pi_H\|_{ld} \\
&= \log \frac{\|\mathbf{P}^{\theta}(Y_n) \circ \cdots \circ \mathbf{P}^{\theta}(Y_1) \circ \mathbf{P}^{\theta}(Y_0) \circ \pi_X \circ \pi_H\|_{ld}}{\|\mathbf{P}^{\theta}(Y_{n-1}) \circ \cdots \circ \mathbf{P}^{\theta}(Y_1) \circ \mathbf{P}^{\theta}(Y_0) \circ \pi_X \circ \pi_H\|_{ld}} + \cdots \\
&\quad + \log \frac{\|\mathbf{P}^{\theta}(Y_1) \circ \mathbf{P}^{\theta}(Y_0) \circ \pi_X \circ \pi_H\|_{ld}}{\|\mathbf{P}^{\theta}(Y_0) \circ \pi_X \circ \pi_H\|_{ld}} + \log \|\mathbf{P}^{\theta}(Y_0) \circ \pi_X \circ \pi_H\|_{ld}.
\end{aligned}
$$

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 $\{Y_0, Y_1, \cdots, Y_n\}$ 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 GMSM under conditions C1-C4, we need to apply Lemma 1 in the online supplementary material and Corollary 1 of Fuh (2006). For this purpose, we need to check that the induced matrix-valued MIFS satisfies assumptions in Fuh (2006), and 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 non-separable space. Therefore, Lemma 1 in the online supplementary material and Theorems 1-4 of Fuh (2006) still hold under the regularity conditions. An alternative approach for this issue can be found in Section 7 of Diaconis and Freedman (1999), in which they provide a direct argument of convergence rather than dealing with the measure-theoretic technicalities created by a non-separable 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 online 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 online supplementary material, under conditions C1-C4, we will 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 GMSM, we need to defined the Fisher information matrix

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

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 $\mathbb{P}^{\theta}_{\tilde{\Pi}}$ to be 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, it is not easy to compute the MLE of GMSM. However, it is possible to have an approximate MLE for GMSM. For example, for the 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, and this approach has been proved to perform an excellent job with a considerable advantage in computation time, while Ghahramani and Hinton (2000) propose a variational approximation method, which is somewhat 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 GMSM, we shall investigate the Kullback-Leibler divergence in Lemma 4 in the online supplementary material, and Fisher information in Theorem 3 below, which are of independent interest. Since the proof of convergence of the score function and Fisher information involves derivatives of the log likelihood function, we first show that derivatives of the log likelihood function $l(\theta)$ in (3.7) can also be written 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 l-th coordinate of θ in a neighborhood $N_\delta(\theta_0)$ of the true parameter θ_0, and $(D_l)^k$ is the corresponding k-th

partial derivative. Now for any two given random functions $\mathbf{P}_j^\theta(Y_{t+1})$ and $\mathbf{P}_j^\theta(Y_t)$ 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

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

and

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

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

Although only the first two derivatives of the MIFS are used in this paper, 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 non-negative integer vector $\nu = (\nu^{(1)}, \cdots, \nu^{(q)})^\mathsf{T}$, 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 ν-th 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 $\nu_1, \nu_2, \cdots, \nu_K$, and let $W_n = (W_n^{\nu_1}, W_n^{\nu_2}, \cdots, W_n^{\nu_K})^\mathsf{T}$. Recall $\mathcal{M} = (\mathcal{D} \times \mathcal{D} \times \mathcal{X} \times \mathbf{R}^p) \times \mathbf{M}^d$. Then $W_n \in \mathcal{M}^K := \{v = (m_1, \cdots, m_K)^\mathsf{T} : m_k \in \mathcal{M}, 1 \leq k \leq K\}$. 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

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

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

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

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

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

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 $V=(V_1,V_2,\cdots,V_K)^\mathsf{T}\in\mathcal{M}^K$, we define

$$B\circ V:=\left(\sum_{j=1}^K b_{1j}\circ V_j,\sum_{j=1}^K b_{2j}\circ V_j,\cdots,\sum_{j=1}^K b_{Kj}\circ V_j\right)^\mathsf{T}. \qquad (4.5)$$

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

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

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 $q = 1$, i.e., 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 $W_n = (W_n^0, W_n^1, \cdots, W_n^r)^\mathsf{T}$, the vector of the first r-th derivatives. Then for any $0 \le k \le r$, we have

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

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

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

where 0 denotes the zero function in $\mathcal{M}$. Note that W_n forms a MIFS on $\mathcal{M}^K$ and 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. And for $k = 0, 1. \cdots, r$, the component $D^k\mathbf{P}^\theta(Y_n)$ in A_n is a $d \times d$ $\mathbf{M}^d$-valued matrix, other 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 $\mathbf{M}^d$.

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 0 being a $d \times d$ zero matrix), which can be regarded as a block matrix or partioned matrix, cf. Zhang (2011). In the same manner, although the operator defined in (4.5) looks like a traditional matrix multiplication, it is different by having the multiplication within each component replaced by $\circ$. Nevertheless, the essential idea is to have a matrix form for W_n, by which it constitutes a MIFS via (4.6).

It is worth mentioning that the feature of getting 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 r-th 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 as (S3.2) in the online supplementary material,

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

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 will show in the online supplementary material that, under condi-

tions C1-C4, for $\theta \in N_\delta(\theta_0)$ the MIFS W_n in (4.6) satisfies Assumption K in Fuh (2006). By using this result and the result that the ν-th 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 online 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 1*, 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* $n \to \infty$*,*

$$\frac{1}{\sqrt{n}}(l'_1(\theta_0), \cdots, l'_q(\theta_0))^\tau \longrightarrow N(0, \mathbf{I}(\theta_0)) \quad \textit{in distribution.}$$

5. Conclusions

In this paper, we provide a general Markov switching model, which includes many practically used models as special cases. In this framework, the hidden unit can be either one or two layers, and can be linear (or nonlinear) predictable (or stochastic) function of past information. It can be regarded as a Markov model if we include all hidden units. Furthermore, by making use of a matrix-valued MIFS representation of the likelihood function, we prove the strong consistency and asymptotic normality of the MLE in 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.

By using this framework, it is interesting to explore the asymptotic properties, including the strong consistency, asymptotic normality and even high order asymptotics, of other commonly used estimators such as GMM, Bayesian estimators and generalized empirical likelihood (GEL) estimator.

Supplementary Materials

The online supplementary material includes some examples of GMSM, a simulation study as well as the proofs of Theorems.

Acknowledgements

The authors would like to thank the editor, professor Rong Chen, and an anonymous referee for their constructive and illuminating comments that have significantly improved the present article. The research of Tianxiao Pang is 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**(6), 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**(4), 1614–1635.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics* **31**(3), 307–327.

Cai, J. (1994). A Markov model of switching-regime ARCH, *J. Business & Econom. Statist.* **12**(3), 309–316.

Calvet, L. E. and Fisher, A. J. (2001). Forecasting multifractal volatility, *Journal of Econometrics* **105**(1), 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**(5), 862–875.

Diaconis, P. and Freedman, D. (1999). Iterated random functions, *SIAM Review* **41**(1), 45–76.

Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models, *Bernoulli* **7**(3), 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**(1), 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**(5), 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**(1), 109–126.

Dütting, P., Feng, Z., Narasimhan, H. and Parkes, D. C. (2017). Optimal auctions through deep learning, *arXiv:1706.03459*.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation, *Econometrica* **50**(4), 987–1008.

REFERENCES

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

Francq, C. and Roussignol, M. (1998). Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum likelihood estimator, *Statistics* **32**(2), 151–173.

Fuh, C.-D. (2003). SPRT and CUSUM in hidden Markov models, *Ann. Statist.* **31**(3), 942–977.

Fuh, C.-D. (2004). On Bahadur efficiency of the maximum likelihood estimator in hidden Markov models, *Statistica Sinica* **14**(1), 127–144.

Fuh, C.-D. (2006). Efficient likelihood estimation in state space models, *Ann. Statist.* **34**(4), 2026–2068. Corrigendum in **38**(2), 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**(18), 4866-4878.

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models, *Neural Computation* **12**(4), 831–864.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models, *Machine Learning* **29**(2-3), 245–273.

Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions,

Journal of Econometrics **1**(1), 3–16.

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

Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning, *The Review of Financial Studies* **33**(5), 2223–2273.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* **57**(2), 357–384.

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

Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime, *Journal of Econometrics* **64**(1-2), 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**(2), 514–535.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching, *Journal of Econometrics* **60**(1-2), 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), 1–91.

Leroux, B. G. (1992). Maximum likelihood estimation for hidden Markov models, *Stoch. Proc. Appl.* **40**(1), 127–143.

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

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**(4), 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**(2), 244–253.

Stenflo, D. (2012). A survey of average contractive iterated function systems, *J. Difference Eq. Appl.* **18**(8), 1355–1380.

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

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

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, *Journal of the American Statistical Association* **84**(408), 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**(408), 164–191.

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

Department of Statistics, Zhejiang University City College, Hangzhou 310015, China

E-mail: cdffuh@gmail.com

School of Mathematical Sciences, Zhejiang University, Hangzhou 310058, China

E-mail: txpang@zju.edu.cn