Identifiability and Inference of Hidden Markov Models

Yonghong An
U of Connecticut

Yingyao Hu
Johns Hopkins

Matt Shum
Caltech

This version: October 2013

Abstract

This paper considers the identifiability of a class of hidden Markov models where both the observed and unobserved components take values in finite spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. We prove that both the Markov transition probability and the conditional probability are identified from the joint distribution of *three* consecutive variables given that the cardinality of $\mathcal{X}$ is not greater than that of $\mathcal{Y}$. The approach of identification provides a novel methodology to estimate the hidden Markov models, and the performance of the proposed estimators is illustrated by a Monte Carlo experiment. We further extend our methodology to the Markov-switching model which generalizes the hidden Markov model, and show that the extended model can be similarly identified and estimated from the joint distribution of *four* consecutive variables if the cardinalities of $\mathcal{Y}$ and $\mathcal{X}$ are equal.

1 Introduction

A hidden Markov model is a bivariate stochastic process $\{Y_t, X_t\}_{t=1,2,...}$ where $\{X_t\}$ is an unobserved Markov chain and, conditional on $\{X_t\}$, $\{Y_t\}$ is an observed sequence of independent random variables such that the conditional distribution of Y_t only depends on X_t. The state space of $\{X_t\}$ and $\{Y_t\}$ are denoted by $\mathcal{X}$ and $\mathcal{Y}$ with cardinality $|\mathcal{X}|$ and $|\mathcal{Y}|$, respectively. When Y_t depends both on X_t and the lagged observations Y_{t-1}, it is called Markov Switching Model. Hidden Markov Models (HMMs) and Markov Switching Models (MSMs) are widely used in econometrics, finance and macroeconomics (Hamilton (1989), Hull and White (1987)). They are also found to be important in biology (Churchill (1992)), and speech recognition (Jelinek (1997), Rabiner and Juang (1993)), etc.

A hidden Markov model is fully captured by the conditional probability $f_{Y_t|X_t}$ and the transition probability of the hidden Markov chain $f_{X_t|X_{t-1}}$, which are the two objectives of identification and estimation in many existing studies of HMMs. Petrie (1969) is the first paper addresses the identifiability of a class of HMMs where both $\mathcal{X}$ and $\mathcal{Y}$ are finite state spaces. The paper proves that both the conditional and the transition probability are identified if the whole distribution of the HMM is known and satisfying a set of regularity conditions. This result is extended about three decades later: Lemma 1.2.4 of Finesso (1991) shows that the distribution of a HMM with $|\mathcal{X}|$ hidden states and $|\mathcal{Y}|$ observed states is identified by the marginal distribution of $2|\mathcal{X}|$ consecutive variables. Paz (1971) provides a stronger result in Corollary 3.4 (chapter 1): the marginal distribution of $2|\mathcal{X}|-1$ consecutive variables suffices to reconstruct the whole HMM distribution. More recently, Allman, Matias, and Rhodes (2009) (Theorem 6) provide a new result on identifiability of HMMs using the fundamental algebraic results of Kruskal (1976, 1977). They show that both the Markov transition probability and the conditional probability are uniquely determined by the joint distribution of $2k+1$ consecutive observables where k satisfies $\binom{k+|\mathcal{Y}|-1}{|\mathcal{Y}|-1} \geq |\mathcal{X}|$.

On the other hand, the existing literature on estimation of HMMs mainly focuses on the maximum-likelihood estimator (MLE) initiated by Baum and Petrie (1966) and Petrie (1969) where both the state space $\mathcal{X}$ and the observational space $\mathcal{Y}$ are finite sets. Still using MLE, Leroux (1992), Douc and Matias (2001) and Douc, Moulines, and Ryden (2004) among others investigate the consistency of MLE for general HMMs. Recently, Douc, Moulines, Olsson, and Van Handel (2011) estimate a parametric family of Hidden Markov Models and provide strong consistency under mild assumptions where both the observed and unobserved components take values in a complete separable metric space. Hsu, Kakade, and Zhang (2012) and Siddiqi, Boots, and Gordon (2010) study learning of HMMs using a singular value decomposition method. Full Bayesian approaches are also employed in the inference of parameters of HMMs (e.g. see Chapter 13 in Cappé, Moulines, and Rydén (2005)) where parameters are assigned prior distributions, and the inference on these parameters is conditional on the observations.

Employing recently developed methodology in the literature of measurement errors (Hu (2008)), the present paper proposes a novel approach to identify and estimate HMMs. The basic idea of our identification is that we treat the hidden Markov chain $\{X_t\}$ as the unobserved or latent "true" variable while $\{Y_t\}$ is the corresponding observed variables with error. We show that when both $\mathcal{X}$ and $\mathcal{Y}$ are finite sets, a HMM is uniquely determined by the joint distribution of *three* consecutive observables given that number of unobserved states is not greater than that

of observed states, i.e., $|\mathcal{X}| \leq |\mathcal{Y}|$. The procedure of identification is constructive and it provides a novel estimating strategy for the transition and conditional probabilities. The estimators are consistent under mild conditions, and their performance is illustrated by a Monte Carlo study. Comparing with MLE, our estimators provide global solutions and they are computationally convenient. Hence our estimators can be used as initial values for MLE.

The novel approach of identification can also be employed to identify more general models that extend the basic HMMs with discrete state space. We consider Markov-switching models (MSMs) that generalize the basic HMM by allowing the dependence of the observed variable Y_t on both the hidden state X_t (as in HMMs) and the lagged variable Y_{t-1}. A Markov-switching model is characterized by the conditional probability $f_{Y_t|Y_t-1,X_t}$ and the transition probability $f_{X_t|X_{t-1}}$, which are two objectives of identification. Using the similar methodology of identification for HMMs, we show that Markov-switching models can be identified and estimated by the joint distribution of *four* consecutive observables when both $\mathcal{X}$ and $\mathcal{Y}$ are finite sets and the cardinality of $\mathcal{X}$ is equal to that of $\mathcal{Y}$.

This paper contributes to the literature of HMMs and its generalization in several ways. First, we propose a novel methodology to identify HMMs and MSMs. To the best of our knowledge, this is the first paper on identifiability of Markov Switching Models. We show that the joint distribution of *three* and *four* consecutive observables are informative enough to infer the whole HMMs and Markov-switching models, respectively, under the assumption $|\mathcal{X}| = |\mathcal{Y}|$. The number of observables required for our identification does not change with the cardinality of the state space. This is an important advantage over existing results of identification, e.g., Siddiqi, Boots, and Gordon (2010) we discussed before.

Second, instead of using the maximum likelihood estimator (EM algorithm is often implemented) as most of the existing work does, we propose a new estimating strategy for HMMs which directly mimics the identification procedure. A prominent advantage of our estimator is that it is global and does not rely on initial values. It can be easily implemented and is computationally fast. Furthermore, our estimator can be used as the initial value for MLE and this provides a guidance to choose initial values for MLE, which is an important empirical issue.

Section 2 presents the results of identification for both Hidden Markov and Markov Switching Models. Section 3 provides estimation for hidden Markov models and a Monte Carlo experiment for our proposed estimators. Section 4 illustrates our methodology by an empirical application.

Section 5 concludes. All proofs are collected in the Appendices.

2 Identification

In this section, we address the identification of HMMs and Markov Switching Models.

2.1 Case 1: Hidden Markov Models

Consider a hidden Markov model $\{X_t, Y_t\}_{t=1,2,\cdots,T}$ where $\{X_t\}$ is a Markov chain and, conditional on $\{X_t\}$, $\{Y_t\}$ is a time-series of independent random variables such that the conditional distribution of Y_t only depends on X_t. We assume that both the state space of $\{X_t\}$, $\mathcal{X}$ and the set in which $\{Y_t\}$ takes its values, $\mathcal{Y}$ are finite with equal cardinality, i.e., $|\mathcal{X}| = |\mathcal{Y}|$. Without loss of generality, we assume $\mathcal{X} = \mathcal{Y} = \{1, 2, ..., r\}$[1]. Hereafter we use capital letters to denote random variables, while lower case letters denote a particular realization. For the model described above, the conditional probability and the transition probability are

$$\begin{aligned} f_{Y_t|\{X_t\}_{t=1,2,...,T}} &= f_{Y_t|X_t}, \\ f_{X_t|\{X_{t-1}\}_{t=2,3,...,T}} &= f_{X_t|X_{t-1}}, \end{aligned}$$

respectively, where $f_{Y_t|X_t}$ is a matrix Q containing r^2 conditional probability $f_{Y_t|X_t}(Y_t = i|X_t = j) \equiv P(Y_t = i|X_t = j) = Q_{ij}$. Similarly, the transition probability of the hidden Markov chain $f_{X_t|X_{t-1}}$ is also a matrix P with its r^2 elements being the transition probability, i.e., $f_{X_t|X_{t-1}}(X_t = i|X_{t-1} = j) \equiv P(X_t = i|X_{t-1} = j) = P_{ij}$. Throughout the paper, we assume that both $f_{Y_t|X_t}$ and $f_{X_t|X_{t-1}}$ are time independent. Our research problem is to identify and estimate both $f_{Y_t|X_t}$ and $f_{X_t|X_{t-1}}$ from the observed data $\{Y_t\}_{t=1,2,...,T}$.

HMMs can be generally used to describe the casual links between variables in economics. For instance, a large literature devotes to the explanation of the dependence of health on socioeconomic status, and a HMM can be a perfect tool to describe such effects, as discussed in Adams, Hurdb, McFaddena, Merrillc, and Ribeiroa (2003). In our empirical application, we analyze how the unobserved health status affects individuals' insurance status (whether an individual is insured).

[1]We use $\mathcal{X}$, $\mathcal{Y}$ and $\{1, 2, ..., r\}$ interchangeably in this paper.

Our identification procedure is based on the recently developed methodology in measurement error literature, e.g., Hu (2008). Intuitively, we treat observed variable Y_t as a variable measured with error, while X_t is the "true" or unobserved latent variable. We identify the HMMs in two steps: in the first step, we use the results in Hu (2008) to identify the conditional probability $f_{Y_t|X_t}$. In the second step, we focus on the identification of the transition probability $f_{X_t|X_{t-1}}$, which is based on the result of the first step.

Consider the observed time series data $\{Y_t\}_{t=1,2,...,T}$ which can be thought as observations of a random variable from a single agent for many time periods, i.e., $T \to \infty$. To further investigate the statistical properties of the series, we first impose a regularity condition on the model $\{X_t, Y_t\}_{t=1,2,...,T}$.

Assumption 1. *The time-series process $\{X_t, Y_t\}_{t=1}^{t=\infty}$ is strictly stationary and ergodic.*

Strict stationarity and ergodicity are commonly assumed properties for time-series data for the statistical analysis. This assumption suffices the strict stationarity and ergodicity of $\{Y_t\}$, and the time-invariant conditional probability $f_{Y_t|X_t}$. For the ergodic time-series process $\{Y_t\}$, we consider the joint distribution of four consecutive observations, $Y_{t+1}, Y_t, Y_{t-1}, Y_{t-2}$, $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$. Under Assumption 1, $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$ is uniquely determined by the observed time-series $\{Y_t\}_{t=1,2,...,T}$ where $T \to \infty$ and the result is summarized in the following lemma.

Remark. As we will show later, the assumption above can be relaxed and we may alternatively only assume the stationarity and ergodicity of $\{Y_t\}_{t=1}^{t=\infty}$. Under this alternative assumption, the identification of HMMs is still valid but we need the joint distribution of four consecutive variables $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$.

Lemma 1. *Under assumption 1, the distribution $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$ is identified from the observed time-series $\{Y_t\}_{t=1,2,...,T}$ where $T \to \infty$.*

This result is due to the ergodicity of $\{Y_t\}$, which is induced by assumption 1, Theorems 3.34 and 3.35 in White (2001). This lemma also naturally implies the identification of the joint distribution $f_{Y_{t+1},Y_t,Y_{t-1}}$, f_{Y_{t+1},Y_t} and $f_{Y_t,Y_{t-1}}$ which we use repeatedly in our identification procedure.

Remark. This identification result above is due to the asymptotic independent and identically distributed properties of the stationary and ergodic process. Therefore, this lemma also holds

for any independent and identically distributed panel data $\{Y_{it}\}_{i=1,2,...,N;t=1,2,...,T}$ where Y_{it} are independent across i. Consequently, our procedure of identification and estimation can be readily applied to panel data as a special case.

Considering that both Y_t and X_t are discrete variables, we define a square matrix S_{r_1,R_2,R_3} with its (i,j)-th element being the joint probability $P(R_1 = r_1, R_2 = i, R_3 = j)$, square matrices $S_{R_1|r_2,R_3}$ and S_{R_2,R_3} are similarly defined. A diagonal matrix D_{r_1,R_3} is defined such that its (i,i)-th element on diagonal is $P(R_1 = r_1, R_3 = i)$, $i, j, r_1 = 1, 2, ..., r$. In what follows, our assumptions on the model will be imposed on the matrices describing the joint (conditional) distributions.

Assumption 2. *The matrix $S_{Y_t,Y_{t-1}}$ that describes the observed joint distribution of Y_t, Y_{t-1} has full rank, i.e., rank$(S_{Y_t,Y_{t-1}}) = |\mathcal{X}| = |\mathcal{Y}|$.*

Remark 1. The implications of this assumption can be further observed from the following relationship

$$S_{Y_t,Y_{t-1}} = S_{Y_t|X_t} S_{X_t,Y_{t-1}}. \tag{1}$$

The three matrices $S_{Y_t,Y_{t-1}}, S_{Y_t|X_t}, S_{X_t,Y_{t-1}}$ are of the same dimension $r \times r$. Therefore, the full rank of $S_{Y_t,Y_{t-1}}$ implies that both the matrices on the R.H.S. have full rank, too. The full rank condition on the matrix $S_{Y_t|X_t}$ requires that the probability $P(Y_t|X_t = j)$ is not a linear combination of $P(Y_t|X_t = k), k \neq j, k, j \in \mathcal{Y}$. The full rank of $S_{X_t,Y_{t-1}}$ imposes similar restrictions on the joint probability distribution $P(X_t = i, Y_{t-1} = j), i, j \in \mathcal{Y}$.

Remark 2. The advantage of imposing this assumption is that it can be directly verified from data by testing the rank of $S_{Y_t,Y_{t-1}}$.

Under the assumption of full rank above, we are able to obtain an eigen-decomposition of an observed matrix involving our identification objectives

$$S_{y_{t+1},Y_t,Y_{t-1}} S^{-1}_{Y_t,Y_{t-1}} = S_{Y_t|X_t} D_{y_{t+1}|X_t} S^{-1}_{Y_t|X_t}, \text{ for all } y_{t+1} \in \mathcal{Y}. \tag{2}$$

This eigen-decomposition allows us to identify $S_{Y_t|X_t}$ as the eigenvector matrix of the L.H.S., $S_{Y_{t+1},y_t,Y_{t-1}} S^{-1}_{Y_t,Y_{t-1}}$ which can be recovered from data directly. To ensure the uniqueness of the decomposition, we impose two more assumptions on the model.

Assumption 3. *$D_{y_{t+1}|X_t=k} \neq D_{y_{t+1}|X_t=j}$ for any given $y_{t+1} \in \mathcal{Y}$ whenever $k \neq j, k, j \in \mathcal{Y}$.*

With this assumption, we can exclude the possibility of degenerate eigenvalues, it is still remaining to correctly order the eigenvalues (eigenvectors) which is guaranteed by the next assumption. To further investigate the restrictions this assumption imposes to the model, we consider the following equality

$$D_{y_{t+1}|X_t=j} = P(y_{t+1}|X_t = j) = \sum_{k\in\mathcal{X}} P(y_{t+1}|X_{t+1} = k)P(X_{t+1} = k|X_t = j).$$

It is clear from this equation that assumption 3 does not restrict $P(Y_t|X_t)$ and $P(X_{t+1}|X_t)$ directly rather implicitly imposes restrictions on the combination of them. This is in contrast to what assumed in the literature, e.g., one of the requirements of identifiability in Petrie (1969) is that there exists a $j \in \mathcal{X} = \{1, 2, ..., r\}$ such that all the $P(Y_t = y_t|X_t = j)$ are distinct.

Assumption 4. *There exists a functional $\mathcal{F}(\cdot)$ such that $\mathcal{F}\left(f_{Y_t|X_t}(\cdot|\cdot)\right)$ is monotonic in x_t or y_t.*

Remark. The monotonicity imposed above is not as restrictive as it looks for the following two reasons: first, the functional $\mathcal{F}(\cdot)$ could take any reasonable form. For example, in the Monte Carlo study, the conditional probability matrix $S_{Y_t|X_t}$ is strictly diagonally dominant, hence the columns can be ordered according to the position of the maximal entry of each column. Alternatively, the ordering condition may be derived from some known properties of the distribution $f_{Y_{t+1}|X_t}$. For instance, if it is known that $E(Y_{t+1}|X_t = x_t)$ is monotonic in x_t, we are able to correctly order all the columns of $S_{Y_t|X_t}$. Second, in empirical applications, this assumption is model specific and oftentimes it is implied by the model. For instance, An (2010) and An, Hu, and Shum (2010) show that such an assumption satisfies naturally in auction models. In the empirical application of the present paper, Y_t is whether a patient is insured while X_t is the unobserved health status of this patient. A reasonable assumption is that given all other factors, a more healthy patient is insured with a higher probability, i.e., $\Pr(Y_t = 1|X_t = 1) > \Pr(Y_t = 1|X_t = 0)$. Similarly, $\Pr(Y_t = 0|X_t = 0) > \Pr(Y_t = 0|X_t = 1)$ also holds. For instance, in a simple HMM with two hidden states being "Low" and "High" atmospheric pressure and two observations being "Rain" and "Dry". Then the monotonicity is natural: $P(Y_t = Rain|X_t = Low) > P(Y_t = Dry|X_t = Low)$ and $P(Y_t = Rain|X_t = High) < P(Y_t = Dry|X_t = High)$, i.e., it is more likely to rain (be dry) when atmospheric pressure is low (high). While in a theoretical model, noisy observations within the framework of continuous state spaces are more possible, this assumption oftentimes can be justified in specific economic problems. Of course, we may

still have the risk that this assumption is difficult to justify for the qualitative data, e.g., in the classical case of DNA segmentation.

The decomposition in equation (2) together with assumptions 2, 3 and 4 guarantees the identification of $S_{Y_t|X_t}$. To identify the transition probability $S_{X_t|X_{t-1}}$, we decompose the joint distribution of Y_{t+1} and Y_t,

$$\begin{aligned} S_{Y_{t+1},Y_t} &= S_{Y_{t+1}|X_{t+1}} S_{X_{t+1},X_t} S_{Y_t|X_t}^T \\ &= S_{Y_t|X_t} S_{X_{t+1},X_t} S_{Y_t|X_t}^T, \end{aligned} \tag{3}$$

where the second equality is due to the stationarity of the HMMs. The identification of S_{X_{t+1},X_t} follows directly from this relationship and the identified $S_{Y_t|X_t}$.

Theorem 1. *Suppose a class of Hidden Markov Models $\{X_t, Y_t\}_{t=1,2,...,T}$ with $|\mathcal{X}| = |\mathcal{Y}|$ satisfy assumption 1, 2, 3, and 4, then the conditional probability matrix $S_{Y_t|X_t}$ and the transition probability matrix $S_{X_t|X_{t-1}}$ are identified from the observed joint probability $f_{Y_{t+1},Y_t,Y_{t-1}}$ for $t \in \{2, ..., T-1\}$. More specifically, the conditional probability matrix $S_{Y_t|X_t}$ is identified as the eigenvector matrix of $S_{Y_{t+1},Y_t,Y_{t-1}}\ S_{Y_t,Y_{t-1}}^{-1}$.*

$$S_{y_{t+1},Y_t,Y_{t-1}} S_{Y_t,Y_{t-1}}^{-1} = S_{Y_t|X_t} D_{y_{t+1}|X_t} S_{Y_t|X_t}^{-1}. \tag{4}$$

The transition matrix of the Markov chain $S_{X_t|X_{t-1}}$ is identified as

$$(S_{X_{t+1}|X_t})_{i,j} = \frac{(S_{X_{t+1},X_t})_{i,j}}{\sum_j (S_{X_{t+1},X_t})_{i,j}}, \tag{5}$$

where

$$S_{X_{t+1},X_t} = S_{Y_t|X_t}^{-1} S_{Y_{t+1},Y_t} \left(S_{Y_t|X_t}^T\right)^{-1}. \tag{6}$$

Remark 1. A novelty of the identification results is that given $|\mathcal{X}| = |\mathcal{Y}|$, HMMs are identified from the distribution of *three* consecutive observations, which does not vary with the cardinality of $\mathcal{X}$, $\mathcal{Y}$, and this is in contrast to most of the existing results: in Paz (1971), the marginal distribution of $2|\mathcal{X}| - 1$ consecutive variables is needed, while in Allman, Matias, and Rhodes (2009), $2k + 1$ consecutive observables are needed where k satisfies $\binom{k+|\mathcal{Y}|-1}{|\mathcal{Y}|-1} \geq |\mathcal{X}|$.

Remark 2. In the proof of Theorem 1, we imposed the restriction $|\mathcal{X}| = |\mathcal{Y}|$. Nevertheless, the identification results can be extended to the case where $|\mathcal{X}| < |\mathcal{Y}|$ with additional assumptions.

When $|\mathcal{X}| < |\mathcal{Y}|$, it is convenient for us to combine observations of Y_t such that the matrix $S_{Y_t,Y_{t-1}}$ is still invertible with its rank being $|\mathcal{X}|$. To fix ideas, we consider a toy model where $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, 2\}$, then we may combine the observations such that

$$Y_t^* = \begin{cases} 0 & \text{if } Y_t = 0, \\ 1 & \text{if } Y_t = 1, 2. \end{cases}$$

Then the proposed identification procedure allows us to identify $S_{Y_t^*|X_t}$ (correspondingly, the assumptions will be imposed on the new model), and consequently we can identify $S_{Y_t|X_t}$. For the purpose of identification, it is necessary to change assumptions 2, 3, and 4 to accommodate the model with observations $\{Y_t^*\}$.

Assumption 2'. *The matrix $S_{Y_t,Y_{t-1}}$ that describes the observed joint distribution of Y_t, Y_{t-1} has a rank $|\mathcal{X}|$, i.e., rank$(S_{Y_t,Y_{t-1}}) = |\mathcal{X}| < |\mathcal{Y}|$.*

The new assumptions that correspond to Assumptions 3 and 4 are neither sufficient nor necessary to assumptions 3 and 4. We will not explore further the identification in this case, the purpose of this part is only to provide a possible approach to deal with the case where $|\mathcal{X}| < |\mathcal{Y}|$.

2.2 Case 2: Markov Switching Models

The methodology presented in the previous section can be readily applied to identification of extended models of HMMs. In this section, we consider the identification of Markov switching models.

Generalized from hidden Markov models with discrete state space, a Markov switching model[2] (MSM) allows dependence of the observed random variable Y_t on both the unobserved variable X_t and the lagged observation Y_{t-1}. Naturally the identification objectives are the transition probability $f_{X_t|X_{t-1}}$ and the conditional probability $f_{Y_t|Y_{t-1},X_t}$. The MSMs allow more general interactions between Y_t and X_t than that in HMMs. Therefore, it can be similarly used to model the more complicated causal links between economic variables. For example, in our empirical illustration we may assume the current insurance status is determined not only by the current health status of a patient, but also the insurance status of last period for this patient.

Markov-switching models have much in common with basic HMMs. However, identification

[2] It is also called Markov jump systems when the hidden state space is finite.

of Markov-switching models is much more intricate than that for HMMs due to the fact that the properties of the observed process Y_t are controlled not only by those of the unobservable chain X_t (as is the case in HMMs) but also by the lagged observations of Y_{t-1}. Hence we employ a slightly different identification approach from we introduced previously to identify MSMs.

First, we start our identification argument from Lemma 1, which provides identification of the joint distribution $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$ from the observed data $\{Y_t\}$. Then we define a "kernel" of the Markov switching models,

Definition 1. *The kernel of a Markov-switching model is defined as* $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$.

It can be shown that (in Appendix) the kernel $f_{Y_t,Y_t|Y_{t-1},X_{t-1}}$ can be decomposed into our two identification objectives $f_{Y_t|Y_{t-1},X_t}$ and $f_{X_t|X_{t-1}}$, i.e.,

$$f_{Y_t,X_t|Y_{t-1},X_{t-1}} = f_{Y_t|Y_{t-1},X_t} f_{X_t|X_{t-1}}. \tag{7}$$

According to this decomposition, it is sufficient to identify the kernel and one of the identification objective in order to identify a Markov-switching model. We will prove that the kernel and the conditional probability $f_{Y_t|Y_{t-1},X_t}$ can both be identified. To prove the identifiability of $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$, we impose several regularity conditions to the model as follows.

Assumption 5. *For all possible values of* $(y_t, y_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$*, the matrix* $S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}}$ *has full rank, i.e., rank(*$S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}}$*)=*$|\mathcal{X}|$=$|\mathcal{Y}|$.

This assumption is similar to assumption 2, and its implications can be seen from

$$S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}} = S_{Y_{t+1}|y_t,X_{t+1}} S_{X_{t+1}|X_t} D_{y_t|y_{t-1},X_t} S_{X_t|X_{t-1}} S_{X_{t-1}|y_{t-1},Y_{t-2}} D_{y_{t-1},Y_{t-2}}.$$

Since all the matrices in the equation above are $r \times r$, $S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}}$ has full rank implies that all the matrices on the R.H.S. have full rank, too. The full rank condition on the two diagonal matrices $D_{y_t|y_{t-1},X_t}$ and $D_{y_{t-1},Y_{t-2}}$ requires that for any given value y_t and y_{t-1} the probability $P(y_t|y_{t-1}, X_t = k)$ and $P(y_{t-1}, Y_{t-2} = k)$ are positive for all $k = \{1, 2, ..., r\}$, respectively. $S_{X_{t+1}|X_t}$ and $S_{X_t|X_{t-1}}$ are transition matrices of the Markov chain $\{X_t\}$, on which the restriction of full rank requires that the probability $P(X_{t+1}|X_t = j)$ is not a linear combination of $P(X_{t+1}|X_t = k), k \neq j$. Full rank of the matrices $S_{Y_{t+1}|y_t,X_{t+1}}$ and $S_{X_{t-1}|y_{t-1},Y_{t-2}}$ imposes the similar restriction on the conditional probability $P(Y_{t+1} = k|y_t, X_{t+1} = j)$.

Remark. For MSMs with both the observed and unobserved components taking values in a finite space, this crucial invertibility assumption can be directly verified from the data, thus making the assumption testable.

We denote the kernel $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$ as $S_{y_t,X_t|y_{t-1},X_{t-1}}$ in matrix form. The decomposition result in Eq.(7), together with assumption 5 guarantees that there exists a representation of the kernel of MSM, $S_{y_t,X_t|y_{t-1},X_{t-1}}$.

$$S_{y_t,X_t|y_{t-1},X_{t-1}} = S^{-1}_{Y_{t+1}|y_t,X_t} S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}} S^{-1}_{Y_t,y_{t-1},Y_{t-2}} S_{Y_t|y_{t-1},X_{t-1}}. \tag{8}$$

Remark. The representation of $S_{y_t,X_t|y_{t-1},X_{t-1}}$ implies that identification of $f_{Y_{t+1}|Y_t,X_t}$ and $f_{Y_t|Y_{t-1},X_{t-1}}$ is sufficient to identify the kernel of the MSM, $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$ since the joint distributions $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$ and $f_{Y_t,Y_{t-1},Y_{t-2}}$ are both observed from data. Under the stationary assumption, the kernel $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$ is time-invariant. Therefore, $f_{Y_{t+1}|Y_t,X_t} = f_{Y_t|Y_{t-1},X_{t-1}}$, we only need to identify $f_{Y_{t+1}|Y_t,X_t}$ in order to identify the kernel.

Next we impose two additional assumptions under which the distribution $f_{Y_{t+1}|Y_t,X_t}$ is identified by the joint distribution of four consecutive variables $\{Y_{t+1}, Y_t, Y_{t-1}, Y_{t-2}\}$. Recall that $(y_t, y_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$. We may choose another two realizations $\widetilde{y}_t$ and $\widetilde{y}_{t-1}$ for Y_t and Y_{t-1} satisfying $(\widetilde{y}_t, \widetilde{y}_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$. Consequently, $(y_t, \widetilde{y}_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$ and $(\widetilde{y}_t, y_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$. We define a diagonal matrix $\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}$ which describes the probability of X_t for given realizations $\{y_t, \widetilde{y}_t, y_{t-1}, \widetilde{y}_{t-1}\}$.

Definition 2. *A diagonal matrix $\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}$ is defined as*

$$\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t} \equiv D_{y_t|y_{t-1},X_t} D^{-1}_{\widetilde{y}_t|y_{t-1},X_t} D_{\widetilde{y}_t|\widetilde{y}_{t-1},X_t} D^{-1}_{y_t|\widetilde{y}_{t-1},X_t},$$

with its (k,k)-th element being

$$\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}(X_t = k) = \frac{P(y_t|y_{t-1}, X_t = k) P(\widetilde{y}_t|\widetilde{y}_{t-1}, X_t = k)}{P(\widetilde{y}_t|y_{t-1}, X_t = k) P(y_t|\widetilde{y}_{t-1}, X_t = k)}.$$

Our next assumption requires that any two diagonal elements of the matrix $\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}$ are distinct.

Assumption 6. $\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}(X_t = k) \neq \Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}(X_t = j)$ *whenever* $k \neq j, k, j \in \mathcal{Y}$.

This assumption imposes restrictions on the conditional probability of the MSM, $P(y_t|X_t)$, i.e.,

$$\frac{P(y_t|y_{t-1},X_t=k)P(\widetilde{y}_t|\widetilde{y}_{t-1},X_t=k)}{P(\widetilde{y}_t|y_{t-1},X_t=k)P(y_t|\widetilde{y}_{t-1},X_t=k)} \neq \frac{P(y_t|y_{t-1},X_t=j)P(\widetilde{y}_t|\widetilde{y}_{t-1},X_t=j)}{P(\widetilde{y}_t|y_{t-1},X_t=j)P(y_t|\widetilde{y}_{t-1},X_t=j)}.$$

This condition is less restrictive than it looks: as we show in the proof of identification, the identifiability of $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$ only requires that the assumption holds for any $y_t \neq \widetilde{y}_t$ and $y_{t-1} \neq \widetilde{y}_{t-1}$. This property provides flexibility for us to choose $\widetilde{y}_t$ and $\widetilde{y}_{t-1}$.

The next assumption is similar to assumption 4 for HMMs, it enables us to order the eigenvalues/eigenvectors of the matrix decomposition, which we employ in our identification of the objective $f_{Y_{t+1}|y_t,X_t}$,

Assumption 7. *There exists a functional $\mathcal{F}(\cdot)$ such that $\mathcal{F}\left(f_{Y_{t+1}|y_t,X_t}(\cdot|y_t,\cdot)\right)$ is monotonic.*

Again, the assumption of monotonicity above leave us with the flexibility. It can be monotonic in y_{t+1} given (y_t, x_t) or in x_t for given (y_{t+1}, y_t). Under assumptions 1, 5, 6, and 7, the probability distribution $f_{Y_{t+1}|y_t,X_t}$ is identified as the eigenvector matrix of an observed matrix. Consequently, the kernel $f_{Y_{t+1},X_{t+1}|Y_t,X_t}$ is identified according to Eq.(8) from the observed joint probability $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$.

Up to now, it is straightforward to characterize the identification of the transition probability $f_{X_t|X_{t-1}}$ in matrix form:

$$S_{X_t|X_{t-1}} = D^{-1}_{y_t|y_{t-1},X_t} S_{y_t,X_t|y_{t-1},X_{t-1}},$$

which is due to the decomposition of the kernel and the fact that each element of $D_{y_t|y_{t-1},X_t}$ is greater than zero hence $D^{-1}_{y_t|y_{t-1},X_t}$ exists.

Theorem 2. *Suppose a class of Markov-switching models $\{X_t, Y_t\}_{t=3}^{\infty}$ with $|\mathcal{X}| = |\mathcal{Y}|$ satisfy assumptions 1, 5, 6, and 7, then both the conditional probability $f_{Y_t|Y_{t-1},X_t}$ and the transition probability of the hidden Markov process $f_{X_t|X_{t-1}}$ are identified from the observed joint probability $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$ for any $t \in \{3, ..., T-1\}$.*

Remark 1. Similar to the identification results of HMMs, here we impose the stationarity that is guaranteed by assumption 1. If we only assumption the stationarity and ergodicity of $\{Y_t\}$ instead, Markov-switching models can still be identified but from the joint distribution of five consecutive variables $f_{Y_{t+2},Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$.

Remark 2. Both the conditional probability $f_{Y_t|Y_{t-1},X_t}$ and the transition probability $f_{X_t|X_{t-1}}$ can be consistently estimated by directly following the identification procedure. The asymptotic properties of the estimators are similar to those of HMMs.

Remark 3. When $|\mathcal{X}| < |\mathcal{Y}|$, we may still achieve identification by imposing additional restrictions, as we showed in Remark 3 of Theorem 1. We omit the discussion for brevity.

3 Estimation and Monte Carlo Study

The procedure of identification proposed in previous sections is constructive and it can be directly used for estimation. Since the estimation of HMMs and MSMs are similar, we only consider the estimation of HMMs in this section. We also present the asymptotic properties of our estimators and illustrate their performance by a Monte Carlo experiment.

3.1 Consistent Estimation

To estimate HMMs consistently, we first provide an uniformly consistent estimator for the joint distribution of $f_{Y_{t+1},Y_t,Y_{t-1}}$, then employ the constructive identification procedure to estimate the transition probability $f_{X_t|X_{t-1}}$ and the conditional probability $f_{Y_t|X_t}$. More specifically, with the estimated joint distribution $f_{Y_{t+1},Y_t,Y_{t-1}}$, the remaining estimation reduces to diagonalization of matrices which can be estimated directly from the data.

We first present the estimator of $S_{Y_t|X_t}$ based on the following relationship,

$$S_{Y_t|X_t} = \phi(S_{y_{t+1},Y_t,Y_{t-1}} S^{-1}_{Y_t,Y_{t-1}}),$$

where $\phi(\cdot)$ denotes a non-stochastic mapping from a square matrix to its eigenvector matrix described in Eq.(2). To maximize the rate of convergence, we average across all the values of $y_{t+1} \in \mathcal{Y}$ for both sides of Eq.(2), i.e,

$$S_{EY_{t+1},Y_t,Y_{t-1}} S^{-1}_{Y_t,Y_{t-1}} = S_{Y_t|X_t} D_{EY_{t+1}|X_t} S^{-1}_{Y_t|X_t},$$

where the matrices $S_{EY_{t+1},Y_t,Y_{t-1}}$ and $D_{EY_{t+1}|X_t}$ are defined as follows.

$$\begin{aligned} S_{EY_{t+1},Y_t,Y_{t-1}} &\equiv \big(E[Y_{t+1}|Y_t = i, Y_{t-1} = j] S_{Y_t=i,Y_{t-1}=j}\big)_{i,j}, \\ D_{EY_{t+1}|X_t} &\equiv diag\Big(E\left[Y_{t+1}|X_t = 1\right], ..., E\left[Y_{t+1}|X_t = r\right]\Big). \end{aligned}$$

Therefore, the conditional probability matrix $S_{Y_t|X_t}$ is estimated as:

$$\widehat{S}_{Y_t|X_t} = \phi(\widehat{S}_{EY_{t+1},Y_t,Y_{t-1}}\widehat{S}^{-1}_{Y_t,Y_{t-1}}), \tag{9}$$

where the two estimators $\widehat{S}_{EY_{t+1},Y_t,Y_{t-1}}$ and $\widehat{S}^{-1}_{Y_t,Y_{t-1}}$ can both be obtained directly from the observed sample $\{Y_t\}_{t=2,...,T-1}$.

$$\begin{aligned}\widehat{S}_{EY_{t+1},Y_t,Y_{t-1}} &= \left(\frac{1}{T}\sum_{t=2}^{T-1} y_{t+1}\mathbf{1}(Y_t = i, Y_{t-1} = k)\right)_{i,k} \\ \widehat{S}_{Y_t,Y_{t-1}} &= \left(\frac{1}{T}\sum_{t=2}^{T-1} \mathbf{1}(Y_t = k, Y_{t-1} = j)\right)_{k,j}.\end{aligned}$$

Employing the estimator $\widehat{S}_{Y_t|X_t}$, we can estimate the joint probability S_{X_{t+1},X_t} according to the identification equation (6)

$$\widehat{S}_{X_{t+1},X_t} = \widehat{S}^{-1}_{Y_t|X_t}\widehat{S}_{Y_{t+1},Y_t}\left(\widehat{S}^{T}_{Y_t|X_t}\right)^{-1}.$$

Consequently, the transition matrix $S_{X_t|X_{t-1}}$ can be estimated using Eq.(5),

$$(\widehat{S}_{X_t|X_{t-1}})_{i,j} = \frac{(\widehat{S}^{-1}_{Y_t|X_t})_{i,j}}{\sum_j (\widehat{S}_{Y_t|X_{t-1}})_{i,j}}. \tag{10}$$

Asymptotic Properties. Our estimation procedure follows directly identification results, and it only involves sample average and some nonstochastic mappings that do not affect the convergence rate of our estimators. Therefore, we only present a simple summary of the asymptotic properties of our estimators.

From the observed time-series $\{Y_t\}_{t=0}^{\infty}$, the estimator for the joint distribution of Y_{t+1}, Y_t, Y_{t-1}, F_{Z_t} is

$$\widehat{F}_{Z_t}(z_t) = \frac{1}{T}\sum_{t=3}^{T-1} I\left(Y_{t+1} \leq y_{t+1}, Y_t \leq y_t, Y_{t-1} \leq y_{t-1}\right),$$

here $Z_t = (Y_{t+1}, Y_t, Y_{t-1})$ and $z_t = (y_{t+1}, y_t, y_{t-1})$. Under strong mixing conditions which ensure asymptotic independence of the times series, the asymptotic property of an empirical distribution function $\widehat{F}_{Z_t}$ is well-known. (e.g., see Silverman (1983), Liu and Yang (2008)) We state the necessary conditions and the asymptotic results as follows.

Assumption 8. *The series* $\{Y_t\}_{t=0}^{t=\infty}$ *is strong mixing.*[3]

Remark. Strong mixing is commonly imposed in the inference of time-series data, and it is closely related to the assumption of Harris-recurrent Markov chain: Athreya and Pantula (1986) have shown that a Harris-recurrent Markov chain on a general state space is strong mixing, provided there exists a stationary probability distribution for that Markov chain. Oftentimes, Harris-recurrent Markov chain is assumed for the inference of HMMs with general state space, e.g., see Douc, Moulines, Olsson, and Van Handel (2011).

Under Assumption 8, for any z_t

$$\sqrt{TV^{-1}(z_t)}\big(\widehat{F}_{Z_t}(z_t) - F_{Z_t}(z_t)\big) \xrightarrow{d} N(0,1), T \to \infty,$$

where $V(z_t) = \sum_{l=0}^{\infty}\gamma(l), \gamma(l) = E\left[\mathbf{1}\{Z_t \le z_t\}\mathbf{1}\{Z_{t+l} \le z_t\}\right] - F_{Z_t}^2(z_t)$.
This asymptotic result implies the uniform consistency of $\widehat{S}_{Ey_{t+1},Y_t,Y_{t-1}}$ and $\widehat{S}_{Y_t,Y_{t-1}}$, i.e., $\widehat{S}_{EY_{t+1},Y_t,Y_{t-1}} - S_{EY_{t+1},Y_t,Y_{t-1}} = O_p(T^{-1/2}), \widehat{S}_{Y_t,Y_{t-1}} - S_{Y_t,Y_{t-1}} = O_p(T^{-1/2})$. Since $\phi(\cdot)$ is a non-stochastic analytical function, we can obtain $\widehat{S}_{Y_t|X_t} - S_{Y_t|X_t} = O_p(T^{-1/2})$ and consequently $\widehat{S}_{X_{t+1}|X_t} - S_{X_{t+1}|X_t} = O_p(T^{-1/2})$.

Remark. In the literature on HMMs, estimation of the parameters has most often been performed using maximum-likelihood estimation. When both X_t and Y_t take values in finite sets as in our analysis, Baum and Petrie (1966) provide results on consistency and asymptotic normality of the maximum-likelihood estimator (MLE). In practice, MLE is often computed using EM (expectation-maximization) algorithm. For HMMs, the EM algorithm was formulated by Baum, Petrie, Soules, and Weiss (1970), and it is known as Baum-Welch (forward-backward) algorithm. Even though the EM of HMMs is efficient, there are two major drawbacks: first, it is well-known that the EM may converge towards a local maximum or even a saddle point whatever optimization algorithm is used. Second, the rate of convergence for the EM algorithm, which is only linear in the vicinity of MLE, can be very slow.[4] In contrast, a prominent advantage

[3]Let $\{Y_t\}_{t=0}^{t=\infty}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$ and $\mathcal{F}_n^m = \sigma\{Y_t : n \le t \le m\}$ be the σ-algebra generated by the random variables $\{Y_n, \cdots, Y_m\}$. Define $\alpha(m) = \sup|P(E\bigcap F)| - P(E)P(F)|$, where the supremum is taken over all $E \in \mathcal{F}_0^n$, $F \in \mathcal{F}_{n+m}^{\infty}$ and n. We say that $\{Y_t\}$ is strong mixing if $\alpha(m)$ tends to zero as m increases to infinity.

[4]Some modifications have been proposed to improve the rate but little is known whether they work well for HMMs (please see Bickel, Ritov, and Ryden (1998) for further discussions.)

of our estimating strategy is that the estimators are global and this overcomes the difficulty of local maximum. Furthermore, our procedure is easy to implement and much faster than EM algorithm. The disadvantage is that our estimator is less efficient than MLE because it involves matrix inversion and diagonalization. Instead of providing a theoretical comparison between MLE and our estimators, we investigate their performance in a Monte Carlo study.

3.2 Monte Carlo Study

To illustrate the performance of our proposed estimators and compare it with that of MLE, we present some monte carlo evidence in this subsection. We consider the following setup of the hidden Markov model: the state space is $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$, $\{X_t\}_{t=1,2,...,T}$ is a stationary and ergodic Markov chain with the matrices of transition probabilities $\boldsymbol{P}$ and the conditional probabilities $\boldsymbol{Q}$:

$$P = \begin{pmatrix} 0.20 & 0.75 & 0.05 \\ 0.70 & 0.20 & 0.05 \\ 0.10 & 0.05 & 0.90 \end{pmatrix}. \quad Q = \begin{pmatrix} 0.80 & 0.20 & 0.05 \\ 0.15 & 0.75 & 0.15 \\ 0.05 & 0.05 & 0.80 \end{pmatrix}.$$

Assuming that the initial state is $X_0 = (1/3, 1/3, 1/3)$, we first generate the (hidden) Markov chain $\{X_t\}_{t=1,2,...,T}$ using the transition matrix $\boldsymbol{P}$. Then $\{Y_t\}_{t=1,2,...,T}$ is generated from $\{X_t\}$ based on the conditional probability matrix $\boldsymbol{Q}$. To utilize the data sufficiently in the estimation, we arrange the data $\{Y_t\}_{t=1}^{T}$ as $\{Y_1, Y_2, Y_3\}$, $\{Y_2, Y_3, Y_4\}$, ... to estimate $S_{EY_{t+1},Y_t,Y_{t-1}}$. As we mentioned previously, the ordering of estimated eigenvalues is achieved based on the property that the matrix $\boldsymbol{Q}$ is strictly diagonally dominant. Our Monte Carlo results are for sample size $T = 2000, 3000, 5000, 8000$ and 200 replications.

Before estimating the parameters of the HMM, we first test the validity of Assumption 2, i.e., whether the matrix S_{Y_{t+1},Y_t} is of full rank. The results show that for all the sample sizes we consider, rank of the matrix S_{Y_{t+1},Y_t} is equal to 3 for each replication.

The estimates of the matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ are presented in Tables 1 and 2, respectively, where the number of iteration for MLE is 2000. The column of "Matrix decomposition" contains results using our method and standard errors are in parentheses. The initial values of $\boldsymbol{P}$ and $\boldsymbol{Q}$ for MLE

are (indicated "initial values #1"):

$$\boldsymbol{P}_0 = \begin{pmatrix} 0.80 & 0.30 & 0.05 \\ 0.05 & 0.40 & 0.35 \\ 0.15 & 0.30 & 0.60 \end{pmatrix}, \ \boldsymbol{Q}_0 = \begin{pmatrix} 0.50 & 0.05 & 0.40 \\ 0.10 & 0.90 & 0.10 \\ 0.40 & 0.05 & 0.50 \end{pmatrix}.$$

The results show that both matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ can be estimated accurately from the observed modest-sized time-series data using our proposed method.[5] Furthermore, the fact that the estimator $\widehat{S}_{Y_t|X_t}$ performs better is consistent with our estimation procedure where $\widehat{S}_{X_t|X_{t-1}}$ is based on $\widehat{S}_{Y_t|X_t}$. As also shown in the tables, the performance of our estimators can be comparable with that of MLE for the same sample size. However, if the initial values are chosen to be close enough to the true values, then MLE outperforms our estimators for all sample sizes. The results for MLE are presented in Table 3, where the initial values of $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the estimates for $T = 2000$ using our method (denoted "initial values #2").

$$\boldsymbol{P}_0 = \begin{pmatrix} 0.40 & 0.91 & 0.07 \\ 0.50 & 0.07 & 0.02 \\ 0.10 & 0.01 & 0.90 \end{pmatrix}, \ \boldsymbol{Q}_0 = \begin{pmatrix} 0.82 & 0.02 & 0.02 \\ 0.17 & 0.86 & 0.17 \\ 0.01 & 0.12 & 0.80 \end{pmatrix}.$$

In practice, it is difficult to obtain initial values that are close enough to the true values for MLE due to the possible existence of local minimum. The results in Table 3 implies that if we employ the estimates of our method as the initial values, the accuracy of MLE will be greatly improved and we even expect a global maximum. Therefore, our method also provides a guidance to choose initial values for MLE, which is an important empirical issue, and the computational convenience of our method further makes such an approach plausible. To illustrate the computational convenience of our method, we provide a comparison of the computing time between our method and MLE in Table 4. The results show that our method is computationally much faster than MLE.[6]

[5] One practical issue of estimation is that the estimated elements of $\boldsymbol{P}$ and $\boldsymbol{Q}$ could be negative or greater than 1. Such results conflict with the fact that each element of the two matrices is a probability and hence between zero and one. In this case, we restrict all elements of $\widehat{S}_{Y_t|X_t}$ and $\widehat{S}_{X_t|X_{t-1}}$ to be between zero and one, then minimize the squared distance between the R.H.S. and L.H.S. of Eq.(4) in estimating $\widehat{S}_{Y_t|X_t}$.

[6] The computer we use has an Intel Core i5 CPU, 3.2GHz and 4GB of RAM, the operating system is Window XP.

Table 1: Estimation results of matrix $\boldsymbol{P}$: initial values #1

Sample size	Matrix decomposition			MLE		
$T = 2000$	0.397(0.186)	0.910(0.231)	0.068(0.050)	0.254(0.096)	0.538(0.294)	0.094(0.070)
	0.500(0.172)	0.074(0.215)	0.024(0.036)	0.629(0.107)	0.166(0.08)	0.267(0.300)
	0.103(0.063)	0.016(0.054)	0.908(0.042)	0.117(0.037)	0.297(0.341)	0.639(0.355)
$T = 3000$	0.241(0.157)	0.904(0.180)	0.046(0.031)	0.245(0.089)	0.563(0.284)	0.090(0.068)
	0.670(0.138)	0.054(0.160)	0.035(0.029)	0.639(0.102)	0.171(0.068)	0.241(0.290)
	0.089(0.038)	0.042(0.029)	0.919(0.021)	0.116(0.034)	0.266(0.328)	0.669(0.347)
$T = 5000$	0.271(0.110)	0.837(0.109)	0.097(0.026)	0.227(0.063)	0.632(0.244)	0.075(0.061)
	0.575(0.095)	0.147(0.101)	0.023(0.025)	0.663(0.080)	0.181(0.053)	0.171(0.245)
	0.154(0.033)	0.016(0.033)	0.880(0.018)	0.110(0.026)	0.187(0.279)	0.754(0.297)
$T = 8000$	0.242(0.102)	0.833(0.093)	0.084(0.026)	0.218(0.051)	0.672(0.204)	0.066(0.051)
	0.622(0.097)	0.145(0.090)	0.029(0.019)	0.677(0.063)	0.189(0.039)	0.131(0.212)
	0.136(0.025)	0.022(0.023)	0.887(0.017)	0.105(0.016)	0.139(0.229)	0.803(0.257)

Table 2: Estimation results of matrix $\boldsymbol{Q}$: initial values #1

Sample size	Matrix decomposition			MLE		
$T = 2000$	0.818(0.137)	0.020(0.185)	0.023(0.032)	0.576(0.320)	0.204(0.031)	0.321(0.379)
	0.178(0.057)	0.861(0.187)	0.173(0.006)	0.165(0.081)	0.762(0.030)	0.151(0.036)
	0.004(0.099)	0.119(0.063)	0.803(0.029)	0.259(0.310)	0.034(0.039)	0.528(0.379)
$T = 3000$	0.776(0.116)	0.174(0.108)	0.063(0.019)	0.598(0.313)	0.204(0.024)	0.289(0.366)
	0.204(0.034)	0.789(0.120)	0.162(0.004)	0.163(0.065)	0.759(0.043)	0.151(0.028)
	0.020(0.104)	0.037(0.053)	0.775(0.016)	0.239(0.297)	0.037(0.056)	0.560(0.369)
$T = 5000$	0.788(0.069)	0.161(0.084)	0.035(0.014)	0.669(0.268)	0.204(0.015)	0.205(0.318)
	0.164(0.028)	0.833(0.099)	0.167(0.003)	0.160(0.047)	0.756(0.020)	0.149(0.020)
	0.048(0.067)	0.006(0.042)	0.798(0.013)	0.171(0.254)	0.040(0.020)	0.646(0.319)
$T = 8000$	0.769(0.045)	0.182(0.067)	0.041(0.014)	0.716(0.218)	0.203(0.011)	0.154(0.270)
	0.167(0.025)	0.805(0.079)	0.157(0.003)	0.154(0.031)	0.754(0.016)	0.150(0.017)
	0.064(0.045)	0.013(0.031)	0.802(0.013)	0.130(0.210)	0.043(0.017)	0.696(0.270)

Table 3: Estimation results of $\boldsymbol{P}$ and $\boldsymbol{Q}$: initial values #2

Sample size	Estimated $\boldsymbol{P}$			Estimated $\boldsymbol{Q}$		
$T = 2000$	0.203(0.027)	0.753(0.053)	0.059(0.045)	0.802(0.032)	0.198(0.019)	0.046(0.029)
	0.696(0.028)	0.197(0.050)	0.052(0.047)	0.148(0.029)	0.753(0.019)	0.149(0.033)
	0.101(0.012)	0.050(0.014)	0.889(0.059)	0.050(0.014)	0.049(0.010)	0.805(0.045)
$T = 3000$	0.203(0.026)	0.753(0.046)	0.056(0.041)	0.801(0.028)	0.199(0.016)	0.047(0.023)
	0.697(0.027)	0.198(0.043)	0.052(0.041)	0.149(0.024)	0.753(0.017)	0.149(0.026)
	0.100(0.009)	0.049(0.011)	0.892(0.049)	0.050(0.010)	0.049(0.008)	0.804(0.033)
$T = 5000$	0.201(0.019)	0.751(0.035)	0.052(0.033)	0.800(0.021)	0.200(0.013)	0.048(0.019)
	0.699(0.019)	0.199(0.033)	0.052(0.032)	0.150(0.018)	0.750(0.013)	0.148(0.021)
	0.010(0.007)	0.050(0.008)	0.896(0.037)	0.050(0.009)	0.005(0.006)	0.804(0.028)
$T = 8000$	0.200(0.015)	0.750(0.028)	0.050(0.026)	0.801(0.016)	0.200(0.010)	0.050(0.015)
	0.700(0.016)	0.200(0.026)	0.050(0.027)	0.150(0.015)	0.750(0.011)	0.049(0.017)
	0.100(0.006)	0.050(0.007)	0.900(0.033)	0.049(0.007)	0.050(0.006)	0.801(0.022)

Table 4: Computing time (seconds)

Sample size	Matrix decomposition	MLE (500 iterations)	MLE (2000 iterations)
$T = 2000$	74	24033	60447
$T = 3000$	79	34088	86908
$T = 5000$	82	55995	124883
$T = 8000$	86	86422	183998

4 Empirical Illustration

In this section, we illustrate the proposed methodology using a dataset on insurance coverage of drugs for patients with chronic diseases. The data set used in our application is compiled from several sources: (1) Catalina Health Resource blinded longitudinal prescription data warehouse (prescription and patient factors), (2) Redbook (price of drugs), First DataBank (brand or generic of drugs), Pharmacy experts (drug performance factors such as side effects, etc.), Verispan PSA (Direct to Consumer, DTC), and the Physician Drug & Diagnosis Audit (drug usage by ICD-9 disease code).

The patients in the sample are observed to start therapy for their chronic disease for a drug they had not taken before. They are from three cohorts: cohort 1 begins in June, 2002; cohort 2 begins in May, 2002; cohort 3 begins in April, 2002. Each patient is observed for one year. 5920 of the patients in our sample have more than three (re)fills and these patients are used for our estimation. Table 5 presents summary statistics of our sample in analysis.

Empirical specification. Across one year panel, we observe whether a patient is insured or not for her prescribed drug. We attribute the change of observed insurance status $\{Y_t\}_{t=1,2,3} \in \{0,1\}$ to the unobserved binary health status $\{X_t\}_{t=1,2,3}$ (less healthy and healthy) and we are interested in two objectives: how patients' health status evolve and how their health status affect the insurance status. For this purpose, we model $\{X_t, Y_t\}$ as a hidden Markov model and focus on estimating the conditional probability $P(Y_t|X_t)$ and the transition probability of the hidden Markov process $P(X_t|X_{t-1})$, which are both 2×2 matrices.

Justification of assumptions. Before we estimate the model, we provide some discussions on the assumptions 1-4. Assumption 1 and 3 can not be tested directly from the data. However, it is reasonable to impose an assumption of strictly stationary and ergodic since in our application patients have chronic disease and a substantial part (26%) of the sample has taken similar medicine before, the process is stable. Assumption 4 is not testable, but it is reasonable to assume that given all other factors, a more healthy patient is insured with a higher probability than a less healthy patient, i.e., $\Pr(Y_t = 1|X_t = 1) > \Pr(Y_t = 1|X_t = 0)$. Similarly, $\Pr(Y_t = 0|X_t = 0) > \Pr(Y_t = 0|X_t = 1)$ also holds. Hence the matrix $S_{Y_t|X_t}$ is assumed to be strictly diagonally dominant. Assumption 2 is testable and we compute the rank and the condition number of the

Table 5: Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Maximum
# of patients	5920	-	-	-
Age	62.20	13.98	11	90
Gender (M=1)	0.404	0.490	0	1
Insurance (insured=0)	.188	.391	0	1

Table 6: Estimation results

(a) Conditional probability

	Insured	Not insured
Healthy	0.932(0.003)	0.068(0.003)
Less healthy	0.020(0.002)	0.980(0.002)

(b) Transition probability

	Healthy	Less healthy
Healthy	0.996 (0.0007)	0.004(0.0007)
Less healthy	0.000(0.0001)	1.000(0.0001)

matrix S_{Y_2,Y_1}, which describes the joint distribution of Y_2 and Y_1, by bootstrapping 200 times.[7] The resulting rank is 2 ± 0 and the condition number is 4.857 ± 0.161.

Estimation results. The estimated results are presented in Table 6, where the standard errors are estimated using bootstrap (200 times). The results provide patterns of how patients' health status evolves and how it affects their insurance status. Several interesting observations form from the results: (1) a less healthy patient will be not insured with a higher probability than a healthy patient being insured (98% v.s. 93%), which reveals some information about insurer's preference; (2) the transition probabilities show that patients' health status does not change much across two periods, which is not surprising because the patients have chronic disease. The results above are simple since we did not take into account other factors that affect conditional and transition probabilities. Nevertheless, our focus of this empirical example is to illustrate our methodology and also show that our method can be empirically important in analyzing the casual links between variables in economics.

[7]Even though a deterministic relationship between condition number and determinant of a matrix does not exist, a larger condition number is a reasonable indicator of the matrix being closer to singular.

5 Conclusion

We considered the identification of a class of hidden Markov models where both the observed and unobserved components take values in finite spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. We proved that hidden Markov models (Markov switching models) are identified from the joint distribution of three (four) consecutive variables given that the cardinality of $\mathcal{X}$ is equal to that of $\mathcal{Y}$. The approach of identification is constructive and it provides a novel methodology to estimate the hidden Markov models. Comparing with the MLE, our estimators are global, which do not depend on initial values, and they are computationally convenient. Hence our estimators can be used as initial values for MLE. The performance of our methodology is illustrated in a Monte Carlo experiment as well as a simple empirical example. It will be interesting to apply our methodology to the analysis of the casual links between economic variables, e.g., the relationship between health and socioeconomic status as discussed in Adams, Hurdb, McFaddena, Merrillc, and Ribeiroa (2003).

References

Adams, P., M. Hurdb, D. McFaddena, A. Merrillc, and T. Ribeiroa (2003): "Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status," *Journal of Econometrics*, 112, 3–56.

Allman, E., C. Matias, and J. Rhodes (2009): "Identifiability of parameters in latent structure models with many observed variables," *Annals of Statistics*, 37(6A), 3099–3132.

An, Y. (2010): "Nonparametric Identification and Estimation of Level-k Auctions," Manuscript, Johns Hopkins University.

An, Y., Y. Hu, and M. Shum (2010): "Estimating First-Price Auctions with an Unknown Number of Bidders: A Misclassification Approach," *Journal of Econometrics*, 157, 328–341.

Athreya, K., and S. Pantula (1986): "Mixing properties of Harris chains and autoregressive processes," *Journal of applied probability*, 23(1), 880–892.

Baum, L., and T. Petrie (1966): "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics*, 37(6), 1554–1563.

Baum, L., T. Petrie, G. Soules, and N. Weiss (1970): "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The annals of mathematical statistics*, pp. 164–171.

Bickel, P., Y. Ritov, and T. Ryden (1998): "Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models," *The Annals of Statistics*, 26(4), 1614–1635.

Cappé, O., E. Moulines, and T. Rydén (2005): *Inference in hidden Markov models*. Springer Verlag.

Churchill, G. (1992): "Hidden Markov chains and the analysis of genome structure," *Computers & chemistry*, 16(2), 107–115.

Douc, R., and C. Matias (2001): "Asymptotics of the maximum likelihood estimator for general hidden Markov models," *Bernoulli*, 7(3), 381–420.

Douc, R., E. Moulines, J. Olsson, and R. Van Handel (2011): "Consistency of the Maximum Likelihood Estimator for general hidden Markov models," *Annals of Statistics*, 39(1), 474–513.

Douc, R., E. Moulines, and T. Ryden (2004): "Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime," *Annals of Statistics*, 32(5), 2254–2304.

Finesso, L. (1991): "Consistent estimation of the order for Markov and hiddenMarkov chains," Ph.D. thesis, University of Maryland.

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

Hsu, D., S. Kakade, and T. Zhang (2012): "A spectral algorithm for learning hidden markov models," *Journal of Computer and System Sciences*, 78, 1460–1480.

Hu, Y. (2008): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, 144, 27–61.

Hull, J., and A. White (1987): "The pricing of options on assets with stochastic volatilities," *Journal of finance*, 42(2), 281–300.

Jelinek, F. (1997): *Statistical methods for speech recognition*. the MIT Press.

Kruskal, J. (1976): "More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling," *Psychometrika*, 41(3), 281–293.

——— (1977): "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear algebra and its applications*, 18(2), 95–138.

Lauritzen, S. L. (1996): *Graphical models*, vol. 17. Oxford University Press, USA.

Leroux, B. (1992): "Maximum-likelihood estimation for hidden Markov models," *Stochastic processes and their applications*, 40(1), 127–143.

Liu, R., and L. Yang (2008): "Kernel estimation of multivariate cumulative distribution function," *Journal of Nonparametric Statistics*, 20(8), 661–677.

Paz, A. (1971): *Introduction to probabilistic automata*. Academic Press.

Petrie, T. (1969): "Probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics*, 40(1), 97–115.

Rabiner, L., and B. Juang (1993): *Fundamentals of speech recognition*. Prentice hall.

Siddiqi, S., B. Boots, and G. Gordon (2010): "Reduced-rank hidden Markov models," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Silverman, B. (1983): "Convergence of a class of empirical distribution functions of dependent random variables," *Annals of Probability*, 11(3), 745–751.

White, H. (2001): *Asymptotic theory for econometricians*. Academic Press New York.

Appendix

A Proof of Theorem 1

According to the structure of the HMM in our analysis, the joint distribution $f_{Y_{t+1},Y_t,Y_{t-1}}$ which describes the evolution of the observed variable, contains information of the conditional and transition probability of the HMM. To single out the objectives of our identification, we consider the following decomposition of the observed joint distribution of Y_{t+1}, Y_t, Y_{t-1}.

$$\begin{aligned}
f_{Y_{t+1},Y_t,Y_{t-1}} &= \sum_{X_t}\sum_{X_{t-1}} f_{Y_{t+1},Y_t,X_t,Y_{t-1},X_{t-1}} \\
&= \sum_{X_t}\sum_{X_{t-1}} f_{Y_{t+1}|Y_t,X_t} f_{Y_t,X_t|Y_{t-1},X_{t-1}} f_{Y_{t-1},X_{t-1}} \\
&= \sum_{X_t}\sum_{X_{t-1}} f_{Y_{t+1}|X_t} f_{Y_t|X_t} f_{X_t|X_{t-1}} f_{Y_{t-1},X_{t-1}} \\
&= \sum_{X_t} f_{Y_{t+1}|X_t} f_{Y_t|X_t} \left(\sum_{X_{t-1}} f_{X_t|X_{t-1}} f_{Y_{t-1},X_{t-1}} \right) \\
&= \sum_{X_t} f_{Y_{t+1}|X_t} f_{Y_t|X_t} f_{X_t,Y_{t-1}}
\end{aligned}$$

Employing these notations, the equation above for all possible values of $y_{t+1} \in \mathcal{Y}$ can be expressed in matrix form as follows.

$$S_{y_{t+1},Y_t,Y_{t-1}} = S_{Y_t|X_t} D_{y_{t+1}|X_t} S_{X_t,Y_{t-1}} \tag{A.1}$$

Similarly, the observed joint distribution $f_{Y_t,Y_{t-1}}$ can also be expressed as

$$\begin{aligned}
f_{Y_t,Y_{t-1}} &= \sum_{X_t} f_{Y_t|X_t,Y_{t-1}} f_{X_t,Y_{t-1}} \\
&= \sum_{X_t} f_{Y_t|X_t} f_{X_t,Y_{t-1}},
\end{aligned}$$

where the second equality is due to the property of HMM. Again, we rewrite the equation above in matrix form,

$$S_{Y_t,Y_{t-1}} = S_{Y_t|X_t} S_{X_t,Y_{t-1}}. \tag{A.2}$$

In both Eq.(A.1) and Eq.(1), the L.H.S. can be recovered directly from the data while the R.H.S. contain the matrix of interest $S_{Y_t|X_t}$. The intuition of identification is to aggregate the information provided by the two observed joint distribution and achieve identification.

The assumption of full rank or invertibility permits us to take the inverse of Eq.(1)

$$S^{-1}_{Y_t,Y_{t-1}} = S^{-1}_{X_t,Y_{t-1}} S^{-1}_{Y_t|X_t}.$$

It follows from Eq.(A.1) and Eq.(1) that

$$S_{y_{t+1},Y_t,Y_{t-1}} S^{-1}_{Y_t,Y_{t-1}} = S_{Y_t|X_t} D_{y_{t+1}|X_t} S^{-1}_{Y_t|X_t}. \tag{A.3}$$

The R.H.S. of the above equation is an eigen-decomposition of the L.H.S., which can be observed directly from the data $\{Y_t\}$. This relationship implies that $S_{Y_t|X_t}$ and $D_{y_{t+1}|X_t}$ may be identified as the eigenvalue and eigenvector matrices, respectively, and the identification requires normalization of eigenvectors, uniqueness and correct ordering of eigenvalues (eigenvectors).

Since the (i,j)-th element of the matrix $S_{Y_t|X_t}$ is a probability $P(Y_t = i|X_t = j)$, then for any $j = 1, 2, ..., r$ the column sum is one. Hence a natural way to normalize the eigenvector matrix is to divide each column by the corresponding column sum. Moreover, assumptions 3 and 4 help us achieve the uniqueness and correct ordering of eigenvalues (eigenvectors), respectively.

The identification of the transition probability $S_{X_t|X_{t-1}}$ of the Markov chain is based on the identification result of $S_{Y_t|X_t}$. We employ the stationarity again, and focus on the identification of $f_{X_{t+1}|X_t}$. For this purpose, we consider

$$\begin{aligned} f_{Y_{t+1},Y_t} &= \sum_{X_{t+1}} \sum_{X_t} f_{Y_{t+1}|X_{t+1}} f_{X_{t+1}|X_t} f_{Y_t|X_t} f_{X_t} \\ &= \sum_{X_{t+1}} \sum_{X_t} f_{Y_{t+1}|X_{t+1}} f_{X_{t+1},X_t} f_{Y_t|X_t}. \end{aligned}$$

This equation is equivalent to

$$\begin{aligned} S_{Y_{t+1},Y_t} &= S_{Y_{t+1}|X_{t+1}} S_{X_{t+1},X_t} S^T_{Y_t|X_t} \\ &= S_{Y_t|X_t} S_{X_{t+1},X_t} S^T_{Y_t|X_t}, \end{aligned} \tag{A.4}$$

in matrix form, where $S_{Y_{t+1}|X_{t+1}} = S_{Y_t|X_t}$ and $S^T_{Y_t|X_t}$ is the transpose of $S_{Y_t|X_t}$. Considering the invertibility of the conditional probability matrix $S_{Y_t|X_t}$ implied by Assumption 2, we identify the joint probability matrix of the Markov chain S_{X_{t+1},X_t} as

$$\begin{aligned} S_{X_{t+1},X_t} &= S^{-1}_{Y_{t+1}|X_{t+1}} S_{Y_{t+1},Y_t} \left(S^T_{Y_t|X_t}\right)^{-1} \\ &= S^{-1}_{Y_t|X_t} S_{Y_{t+1},Y_t} \left(S^T_{Y_t|X_t}\right)^{-1}. \end{aligned} \tag{A.5}$$

Consequently, it is natural to identify the transition matrix $S_{X_{t+1}|X_t}$ as

$$(S_{X_{t+1}|X_t})_{i,j} = \frac{(S_{X_{t+1}|X_t})_{i,j}}{\sum_j (S_{X_{t+1}|X_t})_{i,j}}.$$

B Proof of Theorem 2

We decompose the proof of theorem 2 into several lemmas. Lemma 2 is on the decomposition and representation of the kernel of Markov-switching models. Lemma 3 and 4 construct the identification of the kernel and conditional probability of MSMs, respectively.[8]

Lemma 2. *The kernel of Markov switching models, $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$ can be decomposed into the transition probability of the Markov chain $f_{X_t|X_{t-1}}$ and the conditional probability $f_{Y_t|Y_{t-1},X_t}$, i.e.,*

$$f_{Y_t,X_t|Y_{t-1},X_{t-1}} = f_{Y_t|Y_{t-1},X_t} f_{X_t|X_{t-1}}. \tag{B.1}$$

Under Assumptions 1 and 5, for all $(y_t, y_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$ the kernel (in matrix form) $S_{y_t,X_t|y_{t-1},X_{t-1}}$ can be represented as

$$S_{y_t,X_t|y_{t-1},X_{t-1}} = S^{-1}_{Y_{t+1}|y_t,X_t} S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}} S^{-1}_{Y_t,y_{t-1},Y_{t-2}} S_{Y_t|y_{t-1},X_{t-1}}. \tag{B.2}$$

Proof We derive the decomposition directly as follows.

$$\begin{aligned}
f_{Y_t,X_t|Y_{t-1},X_{t-1}} &= \frac{f_{Y_t,X_t,Y_{t-1},X_{t-1}}}{f_{Y_{t-1},X_{t-1}}} \\
&= \frac{f_{Y_t|X_t,Y_{t-1},X_{t-1}} f_{X_t,Y_{t-1},X_{t-1}}}{f_{Y_{t-1},X_{t-1}}} \\
&= \frac{f_{Y_t|Y_{t-1},X_t} f_{X_t|Y_{t-1},X_{t-1}} f_{Y_{t-1},X_{t-1}}}{f_{Y_{t-1},X_{t-1}}} \\
&= f_{Y_t|Y_{t-1},X_t} f_{X_t|X_{t-1}},
\end{aligned}$$

where the third and the fourth equality are due to the property of MSM: $f_{Y_t|X_t,Y_{t-1},X_{t-1}} = f_{Y_t|Y_{t-1},X_t}$ and $f_{X_t|Y_{t-1},X_{t-1}} = f_{X_t|X_{t-1}}$. For all given $(y_t, y_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$, the relationship above can be expressed in matrix form as

$$S_{y_t,X_t|y_{t-1},X_{t-1}} = S_{y_t|y_{t-1},X_t} S_{X_t|X_{t-1}}.$$

[8] All the derivations involving conditional probabilities can be obtained using the results of graphical models as in Lauritzen (1996).

To show obtain the representation of the kernel $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$, we first investigate the identified joint distribution $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$.

$$\begin{aligned} f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}} &= \sum_{X_t} f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2},X_t} \\ &= \sum_{X_t} f_{Y_{t+1}|Y_t,Y_{t-1},Y_{t-2},X_t} f_{Y_t|Y_{t-1},Y_{t-2},X_t} f_{Y_{t-1},Y_{t-2},X_t} \\ &= \sum_{X_t} f_{Y_{t+1}|Y_t,X_t} f_{Y_t|Y_{t-1},X_t} f_{Y_{t-1},Y_{t-2},X_t}. \end{aligned}$$

where the third equality holds because Y_t does not depend on Y_{t-2} and Y_{t+1} does not depend on Y_{t-1}, Y_{t-2}.

Employing the matrix notation, the equation above for all possible values of $(y_t, y_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$ can be expressed in matrix form as follows.

$$\begin{aligned} S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}} &= S_{Y_{t+1}|y_t,X_t} D_{y_t|y_{t-1},X_t} S_{X_t,y_{t-1},Y_{t-2}} \\ &= S_{Y_{t+1}|y_t,X_t} D_{y_t|y_{t-1},X_t} S_{X_t|y_{t-1},X_{t-1}} S_{X_{t-1},y_{t-1},Y_{t-2}} \\ &= S_{Y_{t+1}|y_t,X_{t+1}} S_{X_{t+1}|y_t,X_t} D_{y_t|y_{t-1},X_t} S_{X_t|y_{t-1},X_{t-1}} S_{X_{t-1}|y_{t-1},Y_{t-2}} D_{y_{t-1},Y_{t-2}} \\ &= S_{Y_{t+1}|y_t,X_{t+1}} S_{X_{t+1}|X_t} D_{y_t|y_{t-1},X_t} S_{X_t|X_{t-1}} S_{X_{t-1}|y_{t-1},Y_{t-2}} D_{y_{t-1},Y_{t-2}}. \end{aligned} \tag{B.3}$$

The simplification of the matrices on the R.H.S. is due to the properties of the MSM.

To construct the representation of the objective $f_{Y_t,X_t|Y_{t-1},X_{t-1}}$, which is described by a matrix $S_{y_t,X_t|y_{t-1},X_{t-1}}$ from the observed joint distribution $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$, we rewrite the matrix $S_{Y_{t+1},y_t,y_{t-1},X_{t-2}}$ as

$$\begin{aligned} S_{Y_{t+1},y_t,y_{t-1},X_{t-2}} &= S_{Y_{t+1}|y_t,X_t} S_{y_t,X_t,y_{t-1},Y_{t-2}} \\ &= S_{Y_{t+1}|y_t,X_t} S_{y_t,X_t|y_{t-1},X_{t-1}} S_{X_{t-1},y_{t-1},Y_{t-2}} \end{aligned} \tag{B.4}$$

Assumption 5 enables us to change the equation above to

$$S_{y_t,X_t|y_{t-1},X_{t-1}} S_{X_{t-1},y_{t-1},Y_{t-2}} = S^{-1}_{Y_{t+1}|y_t,X_t} S_{Y_{t+1},y_t,y_{t-1},X_{t-2}} \tag{B.5}$$

One more step to single out $S_{y_t,X_t|y_{t-1},X_{t-1}}$ is to eliminate $S_{X_{t-1},y_{t-1},Y_{t-2}}$. For this purpose, we use $S_{Y_t,y_{t-1},Y_{t-2}}$ to indicate the joint distribution of Y_t, Y_{t-1}, and Y_{t-2}, then

$$\begin{aligned} S_{Y_t,y_{t-1},Y_{t-2}} &= S_{Y_t|y_{t-1},Y_{t-2},X_{t-1}} S_{y_{t-1},Y_{t-2},X_{t-1}} \\ &= S_{Y_t|y_{t-1},X_{t-1}} S_{y_{t-1},Y_{t-2},X_{t-1}} \end{aligned}$$

Taking into account the full rank (invertibility) of $S_{Y_t|y_{t-1},X_{t-1}}$, we express $S_{y_{t-1},Y_{t-2},X_{t-1}}$ as

$$S_{y_{t-1},Y_{t-2},X_{t-1}} = S^{-1}_{Y_t|y_{t-1},X_{t-1}} S_{Y_t,y_{t-1},Y_{t-2}}$$

Plug this result into Eq.(B.5),

$$S_{y_t,X_t|y_{t-1},X_{t-1}}S^{-1}_{Y_t|y_{t-1},X_{t-1}}S_{Y_t,y_{t-1},Y_{t-2}} = S^{-1}_{Y_{t+1}|y_t,X_t}S_{Y_{t+1},y_t,y_{t-1},X_{t-2}}$$

The resulting $S_{y_t,X_t|y_{t-1},X_{t-1}}$ is

$$S_{y_t,X_t|y_{t-1},X_{t-1}} = S^{-1}_{Y_{t+1}|y_t,X_t}S_{Y_{t+1},y_t,y_{t-1},X_{t-2}}S^{-1}_{Y_t,y_{t-1},Y_{t-2}}S_{Y_t|y_{t-1},X_{t-1}}.$$

■

Lemma 3. *Suppose a class of MSMs satisfy 1, 5, 6, and 7, then the kernel $f_{Y_{t+1},X_{t+1}|Y_t,X_t}$ is identified from the observed joint probability $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$ for any $t \in \{3,...,T-1\}$.*

Proof According to Eq.(8), identification of $S_{y_t,X_t|y_{t-1},X_{t-1}}$ relies on the identifiability of $S_{Y_{t+1}|y_t,X_t}$ and $S_{Y_t|y_{t-1},X_{t-1}}$. We only show in the following that $f_{Y_{t+1}|Y_t,X_t}$ is identified from the observed joint distribution $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$ since the identifiability of $f_{Y_t|Y_{t-1},X_{t-1}}$ can be achieved similarly from the observed distribution $f_{Y_t,Y_{t-1},Y_{t-2},Y_{t-3}}$ without stationarity. If the process is stationary, then $f_{Y_{t+1}|Y_t,X_t} = f_{Y_t|Y_{t-1},X_{t-1}}$ and both of them can be identified from $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$.

Recall that for all $(y_t, y_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$, we have

$$S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}} = S_{Y_{t+1}|y_t,X_t}D_{y_t|y_{t-1},X_t}S_{X_t,y_{t-1},Y_{t-2}}.$$

This allows us to choose another two points $\widetilde{y}_t$ and $\widetilde{y}_{t-1}$ satisfying $(\widetilde{y}_t, \widetilde{y}_{t-1}) \in \mathcal{Y}\times\mathcal{Y}$. Consequently, $(y_t, \widetilde{y}_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$ and $(\widetilde{y}_t, y_{t-1}) \in \mathcal{Y} \times \mathcal{Y}$. If the equation above is evaluated at the four pairs of points $(y_t, y_{t-1}), (\widetilde{y}_t, \widetilde{y}_{t-1}), (y_t, \widetilde{y}_{t-1})$ and $(\widetilde{y}_t, y_{t-1})$, then we obtain

$$\begin{aligned}
S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}} &= S_{Y_{t+1}|y_t,X_t}D_{y_t|y_{t-1},X_t}S_{X_t,y_{t-1},Y_{t-2}},\\
S_{Y_{t+1},\widetilde{y}_t,y_{t-1},Y_{t-2}} &= S_{Y_{t+1}|\widetilde{y}_t,X_t}D_{\widetilde{y}_t|y_{t-1},X_t}S_{X_t,y_{t-1},Y_{t-2}},\\
S_{Y_{t+1},y_t,\widetilde{y}_{t-1},Y_{t-2}} &= S_{Y_{t+1}|y_t,X_t}D_{y_t|\widetilde{y}_{t-1},X_t}S_{X_t,\widetilde{y}_{t-1},Y_{t-2}},\\
S_{Y_{t+1},\widetilde{y}_t,\widetilde{y}_{t-1},Y_{t-2}} &= S_{Y_{t+1}|\widetilde{y}_t,X_t}D_{\widetilde{y}_t|\widetilde{y}_{t-1},X_t}S_{X_t,\widetilde{y}_{t-1},Y_{t-2}}. \qquad \text{(B.6)}
\end{aligned}$$

Under Assumption 5, both $S_{Y_{t+1}|\widetilde{y}_t,X_t}$ and $D_{\widetilde{y}_t|y_{t-1},X_t}$ have full rank, i.e., are invertible. Then $S_{X_t,y_{t-1},Y_{t-2}}$ can be solved from the second equation above as

$$S_{X_t,y_{t-1},Y_{t-2}} = D^{-1}_{\widetilde{y}_t|y_{t-1},X_t}S^{-1}_{Y_{t+1}|\widetilde{y}_t,X_t}S_{Y_{t+1},\widetilde{y}_t,y_{t-1},Y_{t-2}}.$$

Combining this expression with the first equation in Eq.(B.6) leads to

$$S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}} = S_{Y_{t+1}|y_t,X_t}D_{y_t|y_{t-1},X_t}D^{-1}_{\widetilde{y}_t|y_{t-1},X_t}S^{-1}_{Y_{t+1}|\widetilde{y}_t,X_t}S_{Y_{t+1},\widetilde{y}_t,y_{t-1},Y_{t-2}}.$$

The matrix $S_{Y_{t+1},\widetilde{y}_t,y_{t-1},Y_{t-2}}$ is invertible under the full rank assumption. Hence we can get rid of this matrix from the equation above by post-multiplying its inverse $S^{-1}_{Y_{t+1},\widetilde{y}_t,y_{t-1},Y_{t-2}}$ and denote the resulting matrix as U

$$\begin{aligned}
U &\equiv S_{Y_{t+1},y_t,y_{t-1},Y_{t-2}}S^{-1}_{Y_{t+1},\widetilde{y}_t,y_{t-1},Y_{t-2}}\\
&= S_{Y_{t+1}|y_t,X_t}D_{y_t|y_{t-1},X_t}D^{-1}_{\widetilde{y}_t|y_{t-1},X_t}S^{-1}_{Y_{t+1}|\widetilde{y}_t,X_t}.
\end{aligned}$$

A similar matrix V can be obtained from the third and fourth equality in Eq.(B.6),

$$\begin{aligned} V &\equiv S_{Y_{t+1},\widetilde{y}_t,\widetilde{y}_{t-1},Y_{t-2}} S^{-1}_{Y_{t+1},y_t,\widetilde{y}_{t-1},Y_{t-2}} \\ &= S_{Y_{t+1}|\widetilde{y}_t,X_t} D_{\widetilde{y}_t|\widetilde{y}_{t-1},X_t} D^{-1}_{y_t|\widetilde{y}_{t-1},X_t} S^{-1}_{Y_{t+1}|y_t,X_t}. \end{aligned}$$

We further investigate the product of U and V

$$\begin{aligned} UV &= S_{Y_{t+1}|y_t,X_t} D_{y_t|y_{t-1},X_t} D^{-1}_{\widetilde{y}_t|y_{t-1},X_t} S^{-1}_{Y_{t+1}|\widetilde{y}_t,X_t} S_{Y_{t+1}|\widetilde{y}_t,X_t} D_{\widetilde{y}_t|\widetilde{y}_{t-1},X_t} D^{-1}_{y_t|\widetilde{y}_{t-1},X_t} S^{-1}_{Y_{t+1}|y_t,X_t} \\ &= S_{Y_{t+1}|y_t,X_t} \left(D_{y_t|y_{t-1},X_t} D^{-1}_{\widetilde{y}_t|y_{t-1},X_t} D_{\widetilde{y}_t|\widetilde{y}_{t-1},X_t} D^{-1}_{y_t|\widetilde{y}_{t-1},X_t} \right) S^{-1}_{Y_{t+1}|y_t,X_t} \\ &= S_{Y_{t+1}|y_t,X_t} \Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t} S^{-1}_{Y_{t+1}|y_t,X_t}, \end{aligned} \tag{B.7}$$

where the matrix $\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t} \equiv D_{y_t|y_{t-1},X_t} D^{-1}_{\widetilde{y}_t|y_{t-1},X_t} D_{\widetilde{y}_t|\widetilde{y}_{t-1},X_t} D^{-1}_{y_t|\widetilde{y}_{t-1},X_t}$ is diagonal and its (k,k)-th element

$$\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}(X_t = k) = \frac{P(y_t|y_{t-1},X_t=k)\Pr(\widetilde{y}_t|\widetilde{y}_{t-1},X_t=k)}{P(\widetilde{y}_t|y_{t-1},X_t=k)P(y_t|\widetilde{y}_{t-1},X_t=k)}. \tag{B.8}$$

Eq.(B.7) implies that the R.H.S. is an eigenvalue-eigenvector decomposition of the observed matrix UV, with $\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}$ and $S_{Y_{t+1}|y_t,X_t}$ being the eigenvalue and eigenvector matrices, respectively. Under Assumption 5, $D_{y_t|y_{t-1},X_t}$ has full rank, which implies $P(y_t|y_{t-1},X_t=k)>0$ for all $k \in \{1,2,...,r\}$. Therefore, all the eigenvalues of the decomposition above are finite. To uniquely determine $S_{Y_{t+1}|y_t,X_t}$ from the decomposition, we need to assure the eigenvalues are distinct and the eigenvectors can be normalized and correctly ordered.

To normalize the eigenvectors, we consider that the (i,j)-th element of $S_{Y_{t+1}|y_t,X_t}$ is $P(Y_{t+1}=i|y_t,X_t=j)$, and we have $\sum_{i=1}^{r} P(Y_{t+1}=i|y_t,X_t=j)=1$ for all $j \in \{1,2,...,r\}$ and every given y_t. This relationship provides a convenient procedure to normalize the eigenvector matrix $S_{Y_{t+1}|X_t}$: divide each column by column sum. The distinct eigenvalues are guaranteed by assumption 6. It remains to order the eigenvalues (eigenvectors) correctly, and this requires some monotonicity of columns of the eigenvector $S_{Y_{t+1}|y_t,X_t}$ matrix or elements of the eigenvalue matrix $\Lambda_{y_t,\widetilde{y}_t,y_{t-1},\widetilde{y}_{t-1},X_t}$. This ordering condition is guaranteed by Assumption 7.

Since the process in analysis is stationary, $S_{Y_{t+1}|y_t,X_t} = S_{Y_t|y_t,X_{t-1}}$ holds. Therefore both $S_{Y_{t+1}|y_t,X_t}$ and $S_{Y_t|y_{t-1},X_{t-1}}$ are identified from the joint probability distribution $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$. Recall in Eq.(8), our identification objective $f_{Y_{t+1},X_{t+1}|Y_t,X_t}$ is determined uniquely by the joint probability distribution $f_{Y_{t+1},Y_t,Y_{t-1},Y_{t-2}}$ if $S_{Y_{t+1}|y_t,X_t}$ and $S_{Y_t|y_{t-1},X_{t-1}}$ are identified. ■

Lemma 4. *Under assumptions 1, 5, 6, and 7, the distribution $f_{Y_t|Y_{t-1},X_t}$ is identified from the observed time-series $\{Y_t\}_{t=1,2,...,T}$ where $T \to \infty$.*

Proof We first express the identification objective $f_{Y_t|Y_{t-1},X_t}$,

$$f_{Y_t,Y_{t-1}} = f_{Y_t|Y_{t-1},X_t} f_{X_t,Y_{t-1}}.$$

For any given $y_t \in \mathcal{Y}$, the equation above has the following matrix form

$$D_{y_t,Y_{t-1}} = S_{y_t|Y_{t-1},X_t} S_{X_t,Y_{t-1}}. \tag{B.9}$$

Hence, the identification of $S_{y_t|Y_{t-1},X_t}$ relies on the identification of $S_{X_t,Y_{t-1}}$. To identify $S_{X_t,Y_{t-1}}$, we consider the joint distribution of Y_{t+1}, Y_t and Y_{t-1} for any given $y_t \in \mathcal{Y}$,

$$S_{Y_{t+1},y_t,Y_{t-1}} = S_{Y_{t+1}|y_t,X_t} S_{X_t,y_t,Y_{t-1}}.$$

In the proof of lemma 3, we show the identification of $S_{Y_{t+1}|y_t,X_t}$. Therefore $S_{X_t,y_t,Y_{t-1}}$ can be identified as

$$S_{X_t,y_t,Y_{t-1}} = S^{-1}_{Y_{t+1}|y_t,X_t} S_{Y_{t+1},y_t,Y_{t-1}}.$$

Consequently, the joint distribution of X_t and Y_{t-1}, $S_{X_t,Y_{t-1}}$ can also be identified,

$$\begin{aligned} S_{X_t,Y_{t-1}} &= \sum_{y_t \in \mathcal{Y}} S_{X_t,y_t,Y_{t-1}} \\ &= \sum_{y_t \in \mathcal{Y}} S^{-1}_{Y_{t+1}|y_t,X_t} S_{Y_{t+1},y_t,Y_{t-1}}. \end{aligned} \tag{B.10}$$

Combining Eq.(B.9) and Eq.(B.10), we obtain the identification of $S_{y_t|Y_{t-1},X_t}$,

$$S_{y_t|Y_{t-1},X_t} = D_{y_t,Y_{t-1}} \left(\sum_{y_t \in \mathcal{Y}} S^{-1}_{Y_{t+1}|y_t,X_t} S_{Y_{t+1},y_t,Y_{t-1}} \right)^{-1}. \tag{B.11}$$

■

Recall the decomposition of the kernel of Markov-switching models in Lemma 2,

$$S_{y_t,X_t|y_{t-1},X_{t-1}} = S_{y_t|y_{t-1},X_t} S_{X_t|X_{t-1}}.$$

This decomposition, together with the identification of $S_{y_t,X_t|y_{t-1},X_{t-1}}$ in Lemma 3, and of $S_{y_t|y_{t-1},X_t}$ in Lemma 4 implies that the transition probability $S_{X_t|X_{t-1}}$ is also identified as

$$S_{X_t|X_{t-1}} = S^{-1}_{y_t|y_{t-1},X_t} S_{y_t,X_t|y_{t-1},X_{t-1}}. \tag{B.12}$$