Journal of Mathematical Sciences, Vol. 163, No. 3, 2009

MAXIMUM LIKELIHOOD ESTIMATION FOR GENERAL HIDDEN SEMI-MARKOV PROCESSES WITH BACKWARD RECURRENCE TIME DEPENDENCE

S. Trevezas* **and N. Limnios*** UDC 519

This paper concerns the study of asymptotic properties of the maximum likelihood estimator (MLE) for the general hidden semi-Markov model (HSMM) with backward recurrence time dependence. By transforming the general HSMM into a general hidden Markov model, we prove that under some regularity conditions, the MLE is strongly consistent and asymptotically normal. We also provide useful expressions for asymptotic covariance matrices, involving the MLE of the conditional sojourn times and the embedded Markov chain of the hidden semi-Markov chain. Bibliography: 17 titles.

1. Introduction

Hidden Markov models (HMMs) were first introduced by Baum and Petrie (1966), where the consistency and asymptotic normality of the maximum likelihood estimator (MLE) was proved for this model. In their study, Baum and Petrie consider both the observable and hidden process with a finite state space. The hidden process forms a Markov chain (MC), and the observable process conditioned on the MC forms a sequence of conditionally independent random variables. This class of HMMs is often referred to as probabilistic functions of Markov chains. The conditions for consistency had been weakened in Petrie (1969). Leroux (1992) and Bickel, Ritov, and Ryden (1998) proved the consistency and asymptotic normality of the MLE, respectively, when the observable process has a general state space.

HMMs have a wide range of applications, including speech recognition (see Rabiner (1989) and Rabiner and Juang (1993)), computational biology (see Krogh et al. (1994)), and signal processing (see Elliott and Moore (1995)). The reader is also referred to Ephraim and Merhav (2002) for an overview of statistical and information-theoretic aspects of hidden Markov processes (HMPs). Ferguson (1980) introduced hidden semi-Markov models (HSMMs), where the hidden process actually forms a semi-Markov chain (SMC). This setting allows arbitrary distributions for sojourn times in states of a SMC, rather than geometric distributions in the case of a HMM. Recent papers that concentrate on computational techniques for HSMMs are those of Guédon (2003) and Sansom and Thomson (2001).

To the best of our knowledge, Barbu and Limnios (2006) were the first to study asymptotic properties of the MLE for a HSMM. In this paper, we present a different approach which can be summarized as follows:

(i) We generalize the results for HSMM found therein to the general HSMM, where the state space of the observable process is assumed to be a subset of a Euclidean space. For this purpose, we follow the lines of Leroux (1992) and Bickel et al. (1998);

(ii) we allow the values of the observable process (Y_n), conditioned on a SMC, to depend probabilistically not only on the state Z_n but also on the time for which the system has stayed at this current state (backward recurrence time dependence);

(iii) we use minimal representations for parametric spaces which are involved in our analysis, taking into consideration dependence relations between parameters. We also use for each i and j, general constants $\widetilde{n}_{ij}$ to specify the support for conditional sojourn times, rather than extending the parametric space by identically zero parameters;

(iv) we perform a decomposition of elements of the semi-Markov kernel that is different from that found in Barbu and Limnios (2006).

Taken together, (iii) and (iv) open a way for explicit expressions for asymptotic covariance matrices (as functions of the semi-Markov kernel) which appear in central limit theorems for the MLE of the basic characteristics of the semi-Markov chain.

This paper is organized as follows. In Sec. 2, we introduce the mathematical notation and state the first set of conditions. In Sec. 3, we give a representation of HSMMs as a subclass of HMMs. In Sec. 4, we prove the strong consistency of the MLE of a HSMM, and also of the basic characteristics of a SMC, i.e., conditional sojourn

*Laboratoire de Mathématiques Appliquées de Compiègne Université de Technologie de Compiègne, e-mail: Nikolaos.Limnios@utc.fr, Samis.Trevezas@utc.fr.

Published in *Zapiski Nauchnykh Seminarov POMI*, Vol. 363, 2009, pp. 105–125. Original article submitted December 10, 2008.

 1072-3374/09/1633-0262

times and the embedded Markov chain. In Sec. 5, we prove the asymptotic normality of the MLE of a HSMM and of the previously mentioned characteristics.

2. Preliminaries and assumptions

Let $(Z_n, Y_n)_{n\in\mathbb{N}}$ be a hidden semi-Markov chain defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$, where $\theta \in \Theta$, and Θ is a Euclidean subset which parametrizes our model and will be specified later. We assume that the SMC $(Z_n)_{n\in\mathbb{N}}$ has finite state space $E = \{1, 2, \dots, s\}$ and semi-Markov kernel $(q^\theta_{ij}(k))_{i,j\in E, k\in\mathbb{N}}$. If we denote by $(J_n, S_n)_{n\in\mathbb{N}^*}$ the associated Markov renewal process to Z, then $q^\theta_{ij}(k) = \mathbb{P}_\theta(J_{n+1} = j, S_{n+1} - S_n = k \mid J_n = i)$, $n \ge 1$. The process $(S_n)_{n\in\mathbb{N}^*}$ keeps track of successive time points at which changes of states in $(Z_n)_{n\in\mathbb{N}}$ occur (jump times), and $(J_n)_{n\in\mathbb{N}^*}$ records the visited states at these time points. Under this consideration, $q^\theta_{ii}(k) = 0$ for all $i \in E$, $k \in \mathbb{N}$. We use the notation $\mathbf{Z}^{k_2}_{k_1}$ to denote the vector $(Z_{k_1}, Z_{k_1+1}, \dots, Z_{k_2})$, $k_1 \le k_2$, and $\mathbf{i}_d$ denotes the d-dimensional vector with every component equal to the element $i \in E$. The distribution of $\mathbf{Z}^{S_1}_0$ is selected to be $\mathbb{P}_\theta(\mathbf{Z}^{k-1}_0 = \mathbf{i}_k, Z_k = j, S_1 = k) = p^\theta_{ij}\overline{H}^\theta_i(k-1)/\mu^\theta_{ii}$, where p^θ_{ij} refers to the (i,j) element of the transition matrix of the embedded Markov chain $(J_n)_{n\in\mathbb{N}^*}$, $\overline{H}^\theta_i(\cdot)$ is the survival function in state i, and μ^θ_{ii} is the mean recurrence time in the i-renewal process associated to the semi-Markov chain $(Z_n)_{n\in\mathbb{N}}$. We define later the above quantities as functions of the semi-Markov kernel. The selection of the distribution of $\mathbf{Z}^{S_1}_0$ is naturally justified by the fact that it corresponds to the distribution of the same vector in a semi-Markov system that has worked for an infinite time period and is censored at an arbitrary time point, which can be considered as the beginning of our observation. In order to be well defined, it is enough that $\mu_{ii} < \infty$ for all $i \in E$.

We state the following conditions concerning the subclass of SMCs to be considered:

$(A1)$ There exists a minimal $\widetilde{n} \in \mathbb{N}$ such that $q^\theta_{ij}(k) = 0$ for all $k > \widetilde{n}$, $i, j \in E$, and $\theta \in \Theta$.

$(A2)$ The MC $(J_n)_{n\in\mathbb{N}}$ is irreducible.

In fact, conditions (A1) and (A2) imply that $\mu^\theta_{ii} < \infty$ for all $i \in E$. It can easily be shown that the previously defined distribution of $\mathbf{Z}^{S_1}_0$ implies that the SMC $(Z_n)_{n\in\mathbb{N}}$ is stationary. Because of the stationarity, we can allow $(Z_n)_{n\in\mathbb{N}}$ to be indexed by $n \in \mathbb{Z}$. In this case, we denote $S_0 = -\inf\{k \in \mathbb{N} : Z_{-k-1} \ne Z_{-k}\}$. For the observable process, we assume that $(Y_n)_{n\in\mathbb{N}}$ takes values in a measured space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \nu)$, where usually $\mathcal{Y} \subset \mathbb{R}^q$ for some $q \in \mathbb{N}^*$, $\mathcal{B}(\mathcal{Y})$ denotes the Borel subsets on $\mathcal{Y}$, and ν is a σ-finite measure defined on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. Also, let the conditional probability densities $g_\theta(y \mid i, k)$ denote the densities that correspond to the conditional distribution functions $\mathbb{P}_\theta(Y_n \le y \mid \mathbf{Z}^n_{n-k} = \mathbf{i}_{k+1}, Z_{n-k-1} \ne i)$, $i \in E$, $n, k \in \mathbb{N}$. Under condition (A1), there exist constants $\widetilde{n}_{ij}, \widetilde{n}_i < \infty$, such as $\widetilde{n}_{ij} = \max\{k \in \mathbb{N} : q^\theta_{ij}(k) > 0\}$ and $\widetilde{n}_i = \max_{j\in E}\widetilde{n}_{ij}$. The quantities $\widetilde{n}_{ij}$ express the maximum time period for which the SMC can stay at state i before a direct transition to state j. For practical purposes, these time bounds are assumed to be known from characteristics of the system to which this model can be applied, or they can be imposed by the experimenter as an approximation to a more complicated system. The existence of these time bounds is all what we need for theoretical results which follow. For some $i, j \in E$, $\widetilde{n}_{ij}$ may be equal to zero, and this means that no direct transitions from i to j are allowed. Under condition (A1), possible values of k, referring to the conditional densities $g_\theta(y \mid i, k)$, are those for $0 \le k \le \widetilde{n}_i - 1$. In order to simplify the notation, we denote $D_{ij} = \{1, 2, \dots, \widetilde{n}_{ij}\}$ for $i, j \in E$ such that $\widetilde{n}_{ij} > 0$, and $D_i = \{1, 2, \dots, \widetilde{n}_i\}$.

Let T be a finite index set. Different parametric spaces will be used in the sequel. For the moment, we specify a natural parametric space for the HSMM, i.e.,

$$\Theta := \{q_{ij}(k), \theta_t : k \in D_{ij}, q_{ij}(k) \ge 0, \sum_{j,k} q_{ij}(k) = 1, t \in T\}; \tag{1}$$

in order to distinguish between two different kinds of parameters, we denote

$$\Theta_1 := \{q_{ij}(k) : k \in D_{ij}, q_{ij}(k) \ge 0, \sum_{j,k} q_{ij}(k) = 1\} \tag{2}$$

and

$$\Theta_2 := \{\theta_t : t \in T\}. \tag{3}$$

The space Θ_1 parametrizes elements of the semi-Markov kernel; since $q^\theta_{ij}(k) = pr_{ijk}(\theta) = q_{ij}(k)$ in the natural parametrization, we can suppress the superindex θ from $q^\theta_{ij}(k)$. The space Θ_2 refers to a set of parameters that characterize the conditional densities $g_\theta(y \mid i, k)$. It is possible that they distinguish densities from a specific parametric family, from different parametric families, or represent transition probabilities when $\mathcal{Y}$ is a finite state space. In the most simple case of a single parametric family, we have $g_\theta(y \mid i,k) := g(y \mid \theta(i,k)),\ \theta(i,k) \in A$, where $A \subset \mathbb{R}^m$ for some $m \in \mathbb{N}$. In this case, the index set T which appears in Θ_2 consists of all possible couples (i,k).

From now on, we assume for simplicity that the cardinality of T, denoted d_2, is equal to $\sum_i \widetilde{n}_i$, i.e., one one-dimensional parameter corresponds to each conditional density ($m = 1$). Also, we denote $d_1 = \sum_{i,j} \widetilde{n}_{ij}$ and $d = d_1 + d_2$. Then $\Theta_1 \subset \mathbb{R}^{d_1},\ \Theta_2 \subset \mathbb{R}^{d_2},$ and $\Theta = \Theta_1 \times \Theta_2 \subset \mathbb{R}^d$. Since $\sum_{j,k} q_{ij}(k) = 1$ for all $i \in E$, there are s linear dependence relations between elements of the semi-Markov kernel. In order to have a minimal representation of Θ, we have to express s elements of the kernel as functions of the remaining ones. For this purpose, let $J_i = \{j \in E : \widetilde{n}_{ij} = \widetilde{n}_i\}$. We can choose one element $j_i \in J_i$ for all $i \in E$ and express the s elements as follows:

$$q_{ij_i}(\widetilde{n}_i) = 1 - \sum_{j \in E - \{i, j_i\}} \sum_{1 \le k \le \widetilde{n}_{ij}} q_{ij}(k) - \sum_{1 \le k \le \widetilde{n}_i - 1} q_{ij_i}(k). \tag{4}$$

Now, we are in the position to have a minimal representation by using $\Theta^* := \Theta_1^* \times \Theta_2$ as a parametric space, where Θ_1^* results from Θ_1 after extracting the parameters described above. Then $\Theta_1^* \subset \mathbb{R}^{d_3}$ and $\Theta^* \subset \mathbb{R}^{d_4}$, where $d_3 = d_1 - s$ and $d_4 = d_1 + d_2 - s = d - s$.

3. Representation of the HSMMs as a subclass of HMMs

We claim that the general HSMMs with backward recurrence time dependence can be represented as a subclass of HMMs. For this purpose, it is enough to represent the SMCs that satisfy condition (A1) as a special class of MCs. Let $U = (U_n)_{n \in \mathbb{N}}$ be the sequence of backward recurrence times of the SMC $(Z_n)_{n \in \mathbb{Z}}$ defined as follows:

$$U_n = n - S_{N(n)}, \tag{5}$$

where $N(n) = \max\{k \in \mathbb{N} : S_k \le n\}$.

Let $\overline{H}_i(\cdot)$ be the survival function at state i defined by

$$\overline{H}_i(n) := \mathbb{P}(S_{l+1} - S_l > n \mid J_l = i) = 1 - \sum_{j \in \mathbf{E}} \sum_{k=0}^{n} q_{ij}(k),\ n \in \mathbb{N}, l \in \mathbb{N}^*. \tag{6}$$

It can be shown that the stochastic process $(Z,U) := (Z_n, U_n)_{n \in \mathbb{N}}$ is a Markov chain (see Limnios and Oprisan (2001), Theorem 3.12). In a recent paper, Chryssaphinou et al. (2008) study properties of the process (Z,U). This process plays an important role in understanding of the semi-Markov structure. On one hand, it can be used to study the probabilistic behavior and limit theorems for semi-Markov chains, and on the other hand, it can be used to make statistical inference for semi-Markov chains. This role is extended here in the framework of the HSMMs.

Condition (A1) implies that for all $i \in E$, the maximum time period for which $(Z_n)_{n \in \mathbb{N}}$ can stay at this state is $\widetilde{n}_i$. Therefore, the backward recurrence time $U_n \in \{0, 1, \ldots \widetilde{n}_i - 1\}$, and direct transitions from i to j are restricted to the maximum backward recurrence time $\widetilde{n}_{ij} - 1$. Also, it can easily be verified that conditions (A1) and (A2) and the selection of the distribution of $\mathbf{Z}_0^{S_1}$ as previously mentioned render the process (Z,U) a stationary MC with initial distribution given by $\mathbb{P}_\theta((Z_0, U_0) = (i,k)) = \overline{H}_i(k)/\mu_{ii},\ i \in E,\ 0 \le k \le \widetilde{n}_i - 1$. If we denote by $P = (p_{(i,k_1)(j,k_2)})$ the $d_2 \times d_2$ transition probability matrix of the MC (Z,U), then the following proposition specifies transition probabilities of the above MC as a function of the semi-Markov kernel (see also Barbu and Limnios (to appear)). The proof is easy, and it is omitted here.

Proposition 1. *Under condition* (A1), *the transition probabilities of the Markov chain* (Z,U) *can be written as:*

$$p_{(i,k_1)(j,k_2)} = \begin{cases} q_{ij}(k_1+1)/\overline{H}_i(k_1) & \textit{if} \quad i \ne j,\ k_2 = 0, \\ & \textit{and} \quad 0 \le k_1 \le \widetilde{n}_{ij} - 1; \\ \overline{H}_i(k_1+1)/\overline{H}_i(k_1) & \textit{if} \quad i = j,\ k_2 - k_1 = 1, \\ & \textit{and} \quad 0 \le k_1 \le \widetilde{n}_i - 2; \\ 0 & \textit{otherwise}, \end{cases} \tag{7}$$

where $\overline{H}_i(\cdot)$ is given by relation (6).

We present here the matrix P in a block form $P = (P_{ij})_{i,j \in E}$, where P_{ij} is an $\widetilde{n}_i \times \widetilde{n}_j$ matrix,

$$P_{ii} = \begin{pmatrix} 0 & p_{(i,0)(i,1)} & 0 & \dots & 0 \\ 0 & 0 & p_{(i,1)(i,2)} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & p_{(i,\widetilde{n}_i-2)(i,\widetilde{n}_i-1)} \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \tag{8}$$

for $i = j$, and

$$P_{ij} = \begin{pmatrix} p_{(i,0)(j,0)} & 0 & \dots & 0 \\ p_{(i,1)(j,0)} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ p_{(i,\widetilde{n}_{ij}-1)(j,0)} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \tag{9}$$

for $i \neq j$.

Remarks.

(1) From relation (7) we conclude that with every semi-Markov kernel that satisfies condition (A1) we can associate a Markov transition matrix with the corresponding transition probabilities.

(2) If we assume additionally (A2), then $p_{(i,k)(i,k+1)} > 0$, $i \in E$, $0 \leq k \leq \widetilde{n}_i - 2$.

(3) If transitions from i to j are not allowed ($\widetilde{n}_{ij} = 0$), then P_{ij} is a null matrix, while if $\widetilde{n}_{ij} = \widetilde{n}_i$, then the first column of P_{ij} has no fixed zero elements.

In Proposition 1, we considered the probabilities $p_{(i,k_1)(j,k_2)}$ as functions of the semi-Markov kernel which is identified with Θ_1 in the natural parametrization. These probabilities will be denoted by $p^{\theta}_{(i,k_1)(j,k_2)}$ whenever we refer to this parametrization. Additionally, we consider a setting where the parametrization fits from the beginning the class of Markov chains described in Proposition 1. Let $\widetilde{\Theta}_1 = \{p_{(i,k_1)(j,k_2)}\} \subset \mathbb{R}^{d_4}$, where all identically zero elements which appear in P have been excluded, and the restrictions imposed on parameters follow from the stochastic nature of the matrix P. Notice that $\widetilde{\Theta}_1$ can be regarded as a natural parametric space of a subclass of d_2-state Markov chains with transition matrices which are given in block form by (8) and (9). The number of parameters that appear in $\widetilde{\Theta}_1$ equals d_4. Since P is a stochastic matrix, there are exactly d_2 linear relations between elements of P. If we exclude one parameter for each row of P, then the remaining number of parameters equals the dimension of Θ_1^*, i.e., d_3.

We denote by $\widetilde{\Theta}_1^* \subset \mathbb{R}^{d_3}$ a minimal representation of $\widetilde{\Theta}_1$. Similarly, we have $\widetilde{\Theta} = \widetilde{\Theta}_1 \times \Theta_2 \subset \mathbb{R}^{d_2+d_4}$ and $\widetilde{\Theta}^* = \widetilde{\Theta}_1^* \times \Theta_2 \subset \mathbb{R}^{d_4}$. Let $P_{\widetilde{\theta}}$ be a generic element of this subclass of $d_2 \times d_2$ stochastic matrices. We prove the existence of the inverse transformation that represents every MC with d_2 states ($d_2 = \sum_{i=1}^{s} \widetilde{n}_i$) and transition matrix $P_{\widetilde{\theta}}$ as an s-state SMC with a kernel that satisfies condition ($A1$).

Proposition 2. *There exists a continuous function Ψ_1 from $\widetilde{\Theta}_1^*$ into Θ_1^* that reparametrizes every d_2-state Markov chain with transition probability matrix given by $P_{\widetilde{\theta}}$ by an s-state semi-Markov chain with a kernel satisfying condition* (A1), *where the states of the* SMC *correspond to the blocks which the decomposition of P indicates from relations* (8) *and* (9).

Proof. By Theorem 6.7 in Barbu and Limnios (to appear), modified by taking into consideration the constants $\widetilde{n}_{ij}$,

$$q_{ij}(k) = \begin{cases} p_{(i,0)(j,0)} & \text{if} \quad k = 1, \\ p_{(i,k-1)(j,0)} \prod_{r=0}^{k-2} p_{(i,r)(i,r+1)} & \text{if} \quad 2 \leq k \leq \widetilde{n}_{ij}, \end{cases} \tag{10}$$

for i, j such that $\widetilde{n}_{ij} > 0$. The proof is completed by letting all the other elements $q_{ij}(k) = 0$ for $\widetilde{n}_{ij} = 0$. For our statistical purposes, we need a specific minimal representation $\widetilde{\Theta}_1^*$ to consider this transformation as a continuous function from the domain $\widetilde{\Theta}_1^*$ to Θ_1^*. For this purpose, we find if convenient to express $p_{(i,k_1)(j_i,0)}$ as a function of the remaining parameters in the same row of P, where j_i is defined before relation (4). Therefore,

$$p_{(i,k_1)(j_i,0)} = \begin{cases} 1 - \sum\limits_{\substack{j:\widetilde{n}_{ij} \geq k_1+1 \\ j \neq j_i}} p_{(i,k_1)(j,0)} - p_{(i,k_1)(i,k_1+1)} & \text{if } \ 0 \leq k_1 \leq \widetilde{n}_i - 2, \\ 1 - \sum\limits_{j \in G_i} p_{(i,k_1)(j,0)} & \text{if } \ k_1 = \widetilde{n}_i - 1, \end{cases} \tag{11}$$

for all $i \in E$, $0 \leq k_1 \leq \widetilde{n}_i - 1$, where $G_i = \{j : j \neq j_i, \ \widetilde{n}_{ij} = \widetilde{n}_i\}$.

We define the desired transformation $\Psi_1 : \widetilde{\Theta}_1^* \to \Theta_1^*$ by

$$\Psi_1(p_{(i,k_1)(j,k_2)}) = (q_{ij}(k)), \tag{12}$$

where the component functions of Ψ_1 are as follows:

$$q_{ij}(1) = \begin{cases} p_{(i,0)(j,0)} & \text{if } \ j \neq j_i, \\ 1 - \sum\limits_{j \in G_i} p_{(i,0),(j,0)} - p_{(i,0),(i,1)} & \text{if } \ j = j_i, \end{cases} \tag{13}$$

and

$$q_{ij}(k) = \begin{cases} p_{(i,k-1)(j,0)} \prod\limits_{r=0}^{k-2} p_{(i,r)(i,r+1)} & \text{if } \ j \neq j_i, 2 \leq k \leq \widetilde{n}_{ij}, \\ \left(1 - \sum\limits_{j \in G_i} p_{(i,k_1)(j,0)} - p_{(i,k_1)(i,k_1+1)}\right) \prod\limits_{r=0}^{k-2} p_{(i,r)(i,r+1)} & \text{if } \ j = j_i, 2 \leq k < \widetilde{n}_i, \end{cases} \tag{14}$$

for $i, j \in E$ such that $\widetilde{n}_{ij} > 0$. By (13) and (14), we conclude that Ψ_1 is continuous.

Remark. (1) The s parameters of Θ_1 that have been excluded in order to obtain Θ_1^* can be written as follows:

$$q_{ij_i}(\widetilde{n}_i) = (1 - \sum_{j \in G} p_{(i,\widetilde{n}_i-1)(j,0)}) \prod_{r=0}^{\widetilde{n}_i-1} p_{(i,r)(i,r+1)}. \tag{15}$$

4. Consistency results

Following the representation of the previous section, the initial HSMM can now be described by that special kind of HMM, $((Z, U), Y)$.

The stationarity of (Z, U) implies the stationarity of $((Z, U), Y)$. In the sequel, we assume that the natural parametric space Θ^* is a compact subset of $\mathbb{R}^{d_4}$. Since Θ_1^* is a compact subset of $\mathbb{R}^{d_3}$, it is enough that Θ_2 is compact. If this is not the case, we can use a standard compactification technique (see Leroux (1992) and Kiefer and Wolfowitz (1956)). In the mostly simple case of a single parametric family, we have $g_\theta(y \mid i, k) := g(y \mid \theta(i, k))$, $\theta(i, k) \in A$, where $A \subset \mathbb{R}$. Here $\Theta_2 = A^{d_2}$. The likelihood function for an observation $\{\mathbf{Y}_0^n = \mathbf{y}_0^n\}$ can be written as

$$p_\theta(\mathbf{y}_0^n) = \sum_{(i,k)_0^n} \pi_\theta(i_0, k_0) \prod_{j=0}^{n-1} p^\theta_{(i_j,k_j)(i_{j+1},k_{j+1})} \prod_{j=0}^{n} g(y_j \mid \theta(i_j, k_j)),$$

where $\pi_\theta(i, k)$ is the stationary distribution of P_θ. We denote the real value of parameter by θ_0 and $\widetilde{\theta}_0$ when it refers to Θ^* and $\widetilde{\Theta}^*$, respectively. Since for results on asymptotic normality of some characteristics of the system we obtain asymptotic covariance matrices and calculate derivatives with respect to θ, we keep the minimal representation. The estimation problem is to draw inference about this value from a trajectory of $(Y_n)_{n \in \mathbb{N}}$. The MLE denoted by $\widehat{\theta}_n$ maximizes $p_\theta(\mathbf{y}_0^n)$ over Θ^*. In the "best" case, this is a class which consists of parameters θ which are induced by permutations of a specific value that maximizes the given likelihood. For this reason, we define an equivalence relation $\sim$ in Θ^*, where $\theta_1 \sim \theta_2$ if $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$. Then the results for estimators should be understood in the context of $\Theta^*/\sim$, i.e., in the quotient topology induced by this equivalence (see, e.g., Leroux (1992)).

Now we state some extra conditions in order to deduce that the MLE is consistent. These conditions are found in Leroux (1992), and they are adapted here to our model.

(B1) (*Identifiability condition*) The family of mixtures of at most d_2 elements of $\{g(y \mid \theta), \theta \in A\}$ is identifiable.
(B2) The density function $g(y \mid \cdot)$ is continuous in A for any $y \in \mathbb{R}$.
(B3) $E_{\theta_0}[|\log g(Y_1 \mid \theta_0(i,k))|] < \infty$ for all i, k.
(B4) $E_{\theta_0}[\sup_{|\theta' - \theta| < \delta}(\log g(Y_1 \mid \theta'))^+] < \infty$ for any $\theta \in A$ and some $\delta > 0$, where $x^+ = \max(x, 0)$.

In this setting, the identifiability of our model is guaranteed if (A1), (A2), and (B1) hold, and additionally the $\theta(i,k)$ are distinct (for details, see Leroux (1992)). We are now at the point where the results on consistency for MLE concerning the general HSMMs can be deduced from the corresponding results for the general HMMs. We denote by $(\widehat{q}_{ij}(k,n), \widehat{\theta}_t(n))$ the MLE of $\theta_0 = (q^0_{ij}(k), \theta^0_t)$ over Θ^*.

Theorem 1. *If conditions* (A1)–(A2) and (B1)–(B4) *hold, then the MLE* $\widehat{\theta}_n$ *is a strongly consistent estimator of* θ_0 *in the quotient topology, and, consequently,* $(\widehat{q}_{ij}(k,n))$ *is a strongly consistent estimator of* $(q^0_{ij}(k))$ *in the same sense.*

Proof. By Proposition 1, the general HSMM (Z, Y) parametrized by Θ^* can be viewed as a type of a general HMM $((Z, U), Y)$ with the same parametric space Θ^*. The result would follow from Theorem 3, Sec. 6, in Leroux (1992) if conditions 1–6 of that article hold. Indeed, it is easy to verify that condition 1 of Leroux is deduced from (A1) and (A2). Conditions 2 and 3 are identical to (B1) and (B2). Condition 4 is deduced from the fact that the transition probabilities given in Proposition 1 are continuous functions of the semi-Markov kernel, and Conditions 5 and 6 are identical to (B3) and (B4).

Let matrix (p_{ij}) denote the probability matrix of the embedded Markov chain $(J_n)_{n \in \mathbb{N}}$, and let $(f_{ij}(k))$ be the conditional sojourn times, i.e.,

$$p_{ij} = \begin{cases} \sum_{k=1}^{\widetilde{n}_{ij}} q_{ij}(k) & \text{if } \widetilde{n}_{ij} > 0, \\ 0 & \text{if } \widetilde{n}_{ij} = 0, \end{cases} \tag{16}$$

and

$$f_{ij}(k) = \begin{cases} \frac{q_{ij}(k)}{p_{ij}} & \text{if } \widetilde{n}_{ij} > 0,\ 1 \le k \le \widetilde{n}_{ij}, \\ 0 & \text{if } \widetilde{n}_{ij} = 0, \end{cases} \tag{17}$$

for $i, j \in E$. Since these quantities are expressed as functions of the semi-Markov kernel, we refer to them as p^θ_{ij} and $f^\theta_{ij}(k)$ to show that they are parametrized over Θ^*. Nevertheless, we omit superindex θ for estimators. Therefore, we denote by $(\widehat{p}_{ij}(n))$ and $\left(\widehat{f}_{ij}(k,n)\right)$ the corresponding MLE for the true values $\left(p^0_{ij}\right)$ and $\left(f^0_{ij}(k)\right)$, respectively (regarded as vectors), where we exclude identically zero parameters. Also, let $c_i = \text{card}\{j : \widetilde{n}_{ij} > 0\}$ for all $i \in E$, and let $\widetilde{c} = \sum_i c_i$.

Then the following asymptotic results hold.

Corollary 3. *Under conditions* (A1)–(A2) and (B1)–(B4),

(i) *the MLE of the embedded Markov chain* $(\widehat{p}_{ij}(n))$ *is a strongly consistent estimator of* $\left(p^0_{ij}\right)$;

ii) *the MLE of the conditional sojourn time* $\left(\widehat{f}_{ij}(k,n)\right)$ *is a strongly consistent estimator of* $\left(f^0_{ij}(k)\right)$.

Proof. (i) We define a function $\Phi : \Theta^* \to \mathbb{R}^{\widetilde{c}}$, where $\Phi(\theta) = \Phi(q_{ij}(k), \theta_t) = (\sum_{k=1}^{\widetilde{n}_{ij}} q_{ij}(k)) = (p^\theta_{ij})$ due to relation (16) (for $i, j \in E$ such that $\widetilde{n}_{ij} > 0$). We conclude that $(\widehat{p}_{ij}(n)) = \widehat{\Phi(\theta)}(n) = \Phi(\widehat{\theta}_n) = (\sum_{k=1}^{\widetilde{n}_{ij}} \widehat{q}_{ij}(k,n))$, where the second equality holds by the property of MLE. Consequently, we conclude from the continuous mapping theorem, referring to Theorem 1 together with the continuity of Φ, that

$$(\widehat{p}_{ij}(n)) \xrightarrow[n \to \infty]{a.s.} (p^0_{ij}).$$

(ii) Let $pr_{ijk}(\theta) = q_{ij}(k)$ denote the projection of $\theta \in \Theta^*$ into the corresponding element of the semi-Markov kernel, and let Φ_{ij} be the component function of Φ which corresponds to p^θ_{ij}. Let also $T : \Theta^* \to \mathbb{R}^{d_1}$, where $T(\theta) = (T_{ijk}(\theta)) = (pr_{ijk}(\theta)/\Phi_{ij}(\theta))$. Then

$$(f^\theta_{ij}(k)) = \left(\frac{q_{ij}(k)}{p^\theta_{ij}}\right) = \left(\frac{pr_{ijk}(\theta)}{\Phi_{ij}(\theta)}\right) = T(\theta)$$

for $i, j \in E$ such that $\widetilde{n}_{ij} > 0,\ 1 \le k \le \widetilde{n}_{ij}$. Since T is continuous, the result follows along the line of reasoning of Theorem 1, (i).

5. Asymptotic normality results

Two very useful notions for statistical inference, closely connected with MLE, are the rate of entropy of a stochastic process and the generalized Kullback–Leibler divergence. Because of the stationarity of $((Z,U),Y)$, we may allow $((Z_n,U_n),Y_n)_{n\in\mathbb{N}}$ to be indexed by $n\in\mathbb{Z}$. In this case, the rate of entropy of the stochastic process $((Z,U),Y)$ is defined as

$$-\mathbb{H}(\theta_0) := -\mathbb{E}_{\theta_0}[\log \mathbb{P}_{\theta_0}(Y_0 \mid Y_{-1}, Y_{-2}, \dots)],$$

and the generalized Kullback–Leibler divergence is defined as

$$\mathbb{H}_{\theta_0}(\theta) := \mathbb{E}_{\theta_0}[\log \mathbb{P}_{\theta}(Y_0 \mid Y_{-1}, Y_{-2}, \dots)],\ \theta \in \Theta^*.$$

More details about their use in proofs of consistency can be found in Leroux (1992). We denote by $\sigma(\theta_0)$ the opposite of the Hessian matrix of $\mathbb{H}_{\theta_0}(\theta)$ calculated at θ_0, i.e.,

$$\sigma(\theta_0) = \Big(\sigma_{u,v}(\theta_0)\Big)_{u,v} := -\left(\frac{\partial^2 \mathbb{H}_{\theta_0}(\theta)}{\partial\theta_u \partial\theta_v}\bigg|_{\theta=\theta_0}\right)_{u,v}.$$

The third set of conditions which we impose is based on the paper of Bickel et al. (1998) and ensures asymptotic normality of the MLE. The conditions, adapted to our model, can be stated as follows:

(C1) The MC $(Z_n,U_n)_{n\in\mathbb{N}}$ is aperiodic.

(C2) The conditional densities $g(y \mid \theta(i,k))$ have two continuous derivatives with respect to $\theta \in \Theta^*$ in some neighborhood of θ_0 for all possible values of i,k,y.

(C3) There exists a $\delta > 0$ such that

(i) $E_{\theta_0}\left[\sup_{|\theta-\theta_0(i,k)|<\delta}\left|\frac{d}{d\theta}\log g(Y_1 \mid \theta)\right|^2\right] < \infty$,

(ii) $E_{\theta_0}\left[\sup_{|\theta-\theta_0(i,k)|<\delta}\left|\frac{d^2}{d\theta^2}\log g(Y_1 \mid \theta)\right|\right] < \infty$, and

(iii) $\int \sup_{|\theta-\theta_0(i,k)|<\delta}\left|\frac{d^j}{d\theta^j} g(y \mid \theta)\right| \nu(dy) < \infty$ for $1 \le j \le 2$ and for all i,k.

(C4) For $\theta_0 \in \Theta^*$ there exists a $\delta > 0$ such that if

$$r_{\theta_0}(y) := \sup_{\|\theta-\theta_0\|<\delta} \max_{(i_1,k_1),(i_2,k_2)} \frac{g(y \mid \theta(i_1,k_1))}{g(y \mid \theta(i_2,k_2))},$$

then $\mathbb{P}_{\theta_0}(r_{\theta_0}(Y_1) = \infty \mid (Z_1,U_1) = (i,k)) < 1$ for all i,k.

(C5) The true value θ_0 is an interior point of Θ^*.

(C6) The matrix $\sigma(\theta_0)$ is nonsingular.

Remark. Conditions (C1)–(C3) which involve the densities $g(y \mid \theta(i,k))$ can be replaced by similar conditions for more general conditional densities $g_\theta(y \mid i,k)$, as they appear in Bickel et al. (1998).

Theorem 2. *Under conditions* (A1)–(A2), (B1)–(B4), *and* (C1)–(C6), *the* MLE $\widehat{\theta}_n$ *of* θ_0 *is asymptotically normal, i.e.,*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma(\theta_0)^{-1}).$$

Proof. Since Proposition 1 holds, the result would follow from Theorem 1, Sec. 3 of Bickel et al. (1998) if the conditions for asymptotic normality which are stated there hold. Indeed, conditions (A1), (A2), and (C1) render the process (Z,U) an ergodic Markov chain with finite state space; therefore, condition (A1) of Bickel et al. (1998) is satisfied. Conditions (B1)–(B4), combined with (A1) and (A2) imply condition (A6) of Bickel et al. (1998). The remaining conditions are adapted naturally to our model.

At this point, we connect the two natural parametric spaces Θ^* and $\widetilde{\Theta}^*$ for the general HSMM and the type of the general HMM which we already have considered, respectively, by giving a connection between the two asymptotic covariance matrices of the MLE of the HMM and the MLE of the associated HSMM given by Proposition 1.

As we can see from relation (12), Ψ_1 is differentiable on $\widetilde{\Theta}_1^*$. By extending the domain of Ψ_1 in order to include d_2 parameters for conditional densities but keeping the same range, we define $\Psi : \widetilde{\Theta}^* \to \Theta^*$, where

$\Psi = (\Psi_1, pr_{d_2})$, and pr_{d_2} is the projection function on Θ_2. This function is differentiable at $\widetilde{\theta} \in \widetilde{\Theta}^*$, and we denote by $\Psi^{'}$ the total derivative of Ψ calculated at $\widetilde{\theta}_0$. Let also $\sigma(\widetilde{\theta}_0)^{-1}$ be the asymptotic covariance matrix of the MLE $\widehat{\widetilde{\theta}}_n$ of $\widetilde{\theta}_0$. When necessary, we use the following decomposition of the matrix $\sigma(\widetilde{\theta}_0)^{-1}$:

$$\sigma(\widetilde{\theta}_0)^{-1} = \begin{bmatrix} \overbrace{\sigma(\widetilde{\theta}_0)^{-1}_{11}}^{d_3} & \overbrace{\sigma(\widetilde{\theta}_0)^{-1}_{12}}^{d_2} \\ \sigma(\widetilde{\theta}_0)^{-1}_{21} & \sigma(\widetilde{\theta}_0)^{1}_{22} \end{bmatrix} \begin{matrix} \} d_3 \\ \} d_2 \end{matrix} \quad . \tag{18}$$

The following theorem expresses the asymptotic covariance matrix of the MLE corresponding to the HSMM in terms of the natural parametric space $\widetilde{\Theta}^*$ associated to the HMM.

Theorem 3. *Under conditions* (A1)–(A2), (B1)–(B4), *and* (C1)–(C6), *the MLE $\widehat{\theta}_n$ of θ_0 which corresponds to the natural parametric space of the general HSMM satisfies the following relation:*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Psi^{'}\sigma(\widetilde{\theta}_0)^{-1}(\Psi^{'})^{\top})$$

as $n \to \infty$.

Consequently,

$$\sqrt{n}(\widehat{q}_{ij}(k,n) - q^0_{ij}(k)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Psi^{'}_1\sigma(\widetilde{\theta}_0)^{-1}_{11}(\Psi^{'}_1)^{\top}),$$

where the matrix $\Psi^{'}$ is given analytically by relations (31)–(35), *and $\Psi^{'}_1$ is the submatrix of $\Psi^{'}$ formed by its first d_3 rows and columns.*

Proof. For any $i \in E$, let $\widetilde{n}_{i\tau_i(1)}, \widetilde{n}_{i\tau_i(2)}, \dots, \widetilde{n}_{i\tau_i(c_i)}$ be the ordered sequence of $\widetilde{n}_{ij}$ for j such that $\widetilde{n}_{ij} > 0$. If some of the elements are equal, the ordering is performed according to the order of indices j as natural numbers. Note that since $\widetilde{n}_{i\tau_i(c_i)} = \widetilde{n}_i$, $\tau_i(c_i) \in J_i$; therefore, we can choose $j_i = \tau_i(c_i)$.

For all $i \in E$, let

$$\underline{q}(i\tau_i(j)) = \begin{cases} \big(q_{i\tau_i(j)}(1), q_{i\tau_i(j)}(2), \dots, q_{i\tau_i(j)}(\widetilde{n}_{i\tau_i(j)})\big) & \text{if } 1 \le j \le c_i - 1, \\ \big(q_{ij_i}(1), q_{ij_i}(2), \dots, q_{ij_i}(\widetilde{n}_{ij_i} - 1)\big) & \text{if } j = c_i, \end{cases} \tag{19}$$

and

$$\underline{q}(i) = \big(\underline{q}(i\tau_i(1)), \underline{q}(i\tau_i(2)), \dots, \underline{q}(ij_i)\big). \tag{20}$$

Then, if we denote by $\underline{\theta}^{(2)}$ the parameters that correspond to Θ_2, the arrangement of parameters of Θ^* can be presented as follows:

$$(q_{ij}(k), \theta_t) = \big(\underline{q}(1), \underline{q}(2), \dots, \underline{q}(s), \underline{\theta}^{(2)}\big). \tag{21}$$

We need the corresponding arrangement of elements of $\widetilde{\Theta}^*$. For this purpose, for all $i \in E$ and $1 \le j \le c_i - 1$, let

$$\underline{p}(ii) = \big(p_{(i,0)(i,1)}, p_{(i,1)(i,2)}, \dots, p_{(i,\widetilde{n}_i-2),(i,\widetilde{n}_i-1)}\big) \tag{22}$$

and

$$\underline{p}(i\tau_i(j)) = \big(p_{(i,0)(\tau_i(j),0)}, p_{(i,1)(\tau_i(j),0)} \dots, p_{(i,\widetilde{n}_{i\tau_i(j)}-1)(\tau_i(j),0)}\big). \tag{23}$$

Then, denoting

$$\underline{p}(i) = \big(\underline{p}(i\tau_i(1)), \underline{p}(i\tau_i(2)), \dots, \underline{p}(i\tau_i(c_i - 1)), \underline{p}(ii)\big), \tag{24}$$

the expression for arrangement of parameters of $\widetilde{\Theta}^*$ is given by

$$(p_{(i,k_1)(j,k_2)}, \theta_t) = \big(\underline{p}(1), \underline{p}(2), \dots, \underline{p}(s), \underline{\theta}^{(2)}\big). \tag{25}$$

Using relations (13), (14), (21), and (25), we get the following block decomposition for Ψ':

$$\Psi' = \begin{pmatrix} M^{(1)} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & M^{(2)} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \\ \mathbf{0} & \mathbf{0} & \dots & M^{(s)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{I}_{d_2} \end{pmatrix}, \tag{26}$$

where $M^{(i)} = \left(\frac{\partial \underline{q}(i)}{\partial \underline{p}(i)}\right)$ for $i \in E$. Using relations (13), (14), (20), and (24), we decompose $M^{(i)}$ into blocks as follows:

$$M^{(i)} = \begin{pmatrix} M_{11}^{(i)} & \mathbf{0} & \dots & \mathbf{0} & M_{1c_i}^{(i)} \\ \mathbf{0} & M_{22}^{(i)} & \dots & \mathbf{0} & M_{2c_i}^{(i)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & M_{c_i-1,c_i-1}^{(i)} & M_{c_i-1,c_i}^{(i)} \\ M_{c_i1}^{(i)} & M_{c_i2}^{(i)} & \dots & M_{c_i,c_i-1}^{(i)} & M_{c_ic_i}^{(i)} \end{pmatrix}, \tag{27}$$

where

$$M_{jj}^{(i)} = \left(\frac{\partial \underline{q}(i\tau_i(j))}{\partial \underline{p}(i\tau_i(j))}\right), \quad M_{jc_i}^{(i)} = \left(\frac{\partial \underline{q}(i\tau_i(j))}{\partial \underline{p}(ii)}\right), \quad \text{and} \quad M_{c_ij}^{(i)} = \left(\frac{\partial \underline{q}(i\tau_i(c_i))}{\partial \underline{p}(i\tau_i(j))}\right),$$

for $1 \le j \le c_i - 1$, and

$$M_{c_ic_i}^{(i)} = \left(\frac{\partial \underline{q}(i\tau_i(c_i))}{\partial \underline{p}(ii)}\right).$$

These four different types of matrices summarize all the information which we want in order to have an explicit matrix form for Ψ', and we study each of them.

For all $i \in E,\ 1 \le k \le \widetilde{n}_i - 1$, let

$$a_i(k) = \prod_{r=0}^{k-1} p_{(i,r)(i,r+1)}, \tag{28}$$

$$a_i(k;l) = \frac{a_i(k)}{p_{(i,l-1)(i,l)}}, \qquad 1 \le l \le k, \tag{29}$$

and

$$b_{iu}^{(j)}(k;l) = p_{(i,u)(\tau_i(j),0)} a_i(k;l), \quad 1 \le l \le k,\ 1 \le u \le \widetilde{n}_{i\tau_i(j)} - 1. \tag{30}$$

Recall that $j_i = \tau_i(c_i)$, and we also use the abbreviations $c_{ij} = \widetilde{n}_{i\tau_i(j)} - 2$ and $c_{ij}^+ = c_{ij} + 1$. Then

$$M_{jj}^{(i)} = \operatorname{diag}\{1, a_i(1), a_i(2), \dots, a_i(c_{ij}^+)\} \tag{31}$$

and

$$M_{c_ij}^{(i)} = \begin{pmatrix} \Delta_{c_ij}^{(i)} \underline{0}^\top & \\ \mathbf{0} & \underline{0}^\top \end{pmatrix}, \tag{32}$$

where

$$\Delta_{c_ij}^{(i)} = -\operatorname{diag}\{1, a_i(1), a_i(2), \dots, a_i(c_{ij})\}, \tag{33}$$

$$M_{jc_i}^{(i)} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ b_{i1}^{(j)}(1;1) & 0 & \dots & 0 & 0 & \dots & 0 \\ b_{i2}^{(j)}(2;1) & b_{i2}^{(j)}(2;2) & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ b_{i,c_{ij}^+}^{(j)}(c_{ij}^+;1) & b_{i,c_{ij}^+}^{(j)}(c_{ij}^+;2) & \dots & b_{i,c_{ij}^+}^{(j)}(c_{ij}^+;c_{ij}^+) & 0 & \dots & 0 \end{pmatrix}, \tag{34}$$

and

$$M_{c_ic_i}^{(i)} = \begin{pmatrix} -1 & 0 & \dots & 0 & 0 \\ b_{i1}^{(j_i)}(1;1) & -a_i(1) & \dots & 0 & 0 \\ b_{i2}^{(j_i)}(2;1) & b_{i2}^{(j_i)}(2;2) & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b_{i,c_{ij_i}}^{(j_i)}(c_{ij_i};1) & b_{i,c_{ij_i}}^{(j_i)}(c_{ij_i};2) & \dots & b_{i,c_{ij_i}}^{(j_i)}(c_{ij_i};c_{ij_i}) & -a_i(c_{ij_i}) \end{pmatrix} \tag{35}$$

Since

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \sqrt{n}(\Psi(\widehat{\widetilde{\theta}}_n) - \Psi(\widetilde{\theta}_0)), \tag{36}$$

where Ψ is differentiable at $\widetilde{\theta}_0$, Theorem 3 follows from Theorem 2 by an application of the delta method.

Remark. In order to find the asymptotic covariance matrix of $\sqrt{n}(\widehat{q}_{ij}(k,n) - q^0_{ij}(k))$ considered at Θ_1 instead of Θ_1^*, we add parameters $q_{ij_i}(\widetilde{n}_i)$ given by (15); using relation (4), we conclude that $\sqrt{n}(\widehat{q}_{ij}(k,n) - q^0_{ij}(k)) \to \mathcal{N}(0, C\Psi_1^{'}\sigma(\widetilde{\theta}_0)^{-1}_{11}(\Psi_1^{'})^\top C^\top)$, where

$$C = \operatorname{diag}\{C_i, i \in E\},\ C_i = \begin{pmatrix} \mathbf{I}_{r_i} \\ -\underline{1} \end{pmatrix}, \quad \text{and} \quad r_i = \sum_{j=1}^{c_i} \widetilde{n}_{i\tau_i(j)} - 1.$$

Let Φ_1 and T_1 be Φ and T, respectively, considered as functions with the domain Θ_1^*, where Φ and T are defined in Corollary 3. The following two propositions state asymptotic normality results for the MLE of characteristics of the semi-Markov system defined by (16) and (17).

Proposition 4. *Under conditions* (A1)–(A2), (B1)–(B4), *and* (C1)–(C6), *the* MLE *of the embedded Markov chain is asymptotically normal, i.e.,*

$$\sqrt{n}((\widehat{p}_{i,j}(n)) - (p^0_{ij})) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Phi_1^{'}\Psi_1^{'}\sigma(\widetilde{\theta}_0)^{-1}(\Phi_1^{'}\Psi_1^{'})^\top),$$

where $\Phi_1^{'}\Psi_1^{'}$ *is given by relations* (41) *and* (42).

Proof. For all $i \in E$, let

$$\underline{p}^e(i) = \big(p_{i\tau_i(1)}, p_{i\tau_i(2)}, \dots, p_{i\tau_i(c_i)}\big). \tag{37}$$

Then the arrangement of parameters (p_{ij}) of the embedded MC can be represented as follows:

$$(p_{ij}) = \big(\underline{p}^e(1), \underline{p}^e(2), \dots, \underline{p}^e(s)\big). \tag{38}$$

Denote $\left(\frac{\partial \underline{p}^e(i_1)}{\partial \underline{q}(i_2)}\right) := \left(\frac{\partial p_{i_1 j_1}}{\partial q_{i_2 j_2}(k)}\right) = \Phi_1{}'$ and $V^{(i)} := \left(\frac{\partial \underline{p}^e(i)}{\partial \underline{q}(i)}\right)$; then

$$\Phi_1{}' = \operatorname{diag}\{V^{(i)}, i \in E\}, \tag{39}$$

where

$$V^{(i)} = \begin{pmatrix} \underline{1}^{(i)}_{11} & \underline{0} & \cdots & \underline{0} & \underline{0} \\ \underline{0} & \underline{1}^{(i)}_{22} & \cdots & \underline{0} & \underline{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \underline{0} & \underline{0} & \cdots & \underline{1}^{(i)}_{c_i-1,c_i-1} & \underline{0} \\ -\underline{1}^{(i)}_{c_i 1} & -\underline{1}^{(i)}_{c_i 2} & \cdots & -\underline{1}^{(i)}_{c_i,c_i-1} & \underline{0} \end{pmatrix}, \tag{40}$$

and $\underline{1}^{(i)}_{jj}$ and $\underline{1}^{(i)}_{c_i j}$, are $\widetilde{n}_{i\tau_i(j)}$-dimensional row vectors with entries 1 for all j such that $1 \le j \le c_i - 1$.

Since $\sqrt{n}\,\big((\widehat{p}_{i,j}(n)) - (p^0_{ij})\big) = \sqrt{n}\,\big(\Phi_1(\widehat{q}_{ij}(k,n)) - \Phi_1(q^0_{ij}(k))\big)$, we use Theorem 3, the differentiability of Φ_1 on Θ_1^*, and apply the delta method to conclude that

$$\sqrt{n}\,\big((\widehat{p}_{i,j}(n)) - (p^0_{ij})\big) \to \mathcal{N}(0, \Phi_1^{'}\Psi_1^{'}\sigma(\widetilde{\theta}_0)^{-1}(\Phi_1^{'}\Psi_1^{'})^\top),$$

where

$$\Phi_1^{'}\Psi_1^{'} = \operatorname{diag}\{V^{(i)}M^{(i)}, i \in E\}, \tag{41}$$

and $V^{(i)}$ and $M^{(i)}$ are given by (40) and (27), respectively.

The explicit form of their product for all $i \in E$ is as follows:

$$V^{(i)}M^{(i)} = \begin{pmatrix} \underline{d}^{(i)}_1 & \underline{0} & \cdots & \underline{0} & \underline{0} \\ \underline{0} & \underline{d}^{(i)}_2 & \cdots & \underline{0} & \underline{0} \\ \vdots & \vdots & \ddots & \vdots & \\ \underline{0} & \underline{0} & \cdots & \underline{d}^{(i)}_{c_i-1} & \underline{0} \\ -\underline{d}^{(i)}_1 & -\underline{d}^{(i)}_2 & \cdots & -\underline{d}^{(i)}_{c_i-1} & \underline{0} \end{pmatrix}, \tag{42}$$

where $\underline{d}^{(i)}_j = \big(1, a_i(1), a_i(2), \dots, a_i(c^+_{ij})\big)$, and the $a_i(k)$ are given by (28).

Proposition 5. *Under conditions* (A1)–(A2), (B1)–(B4), *and* (C1)–(C6), *the* MLE *of conditional sojourn times is asymptotically normal, i.e.,*

$$\sqrt{n}\left((\widehat{f}_{ij}(k,n)) - (f^0_{ij}(k,n))\right) \underset{n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}(0, T_1^{'}\Psi_1^{'}\sigma(\theta_0)^{-1}(T_1^{'}\Psi_1^{'})^t).$$

Proof. For all $i \in E$ and $1 \le j \le c_i$, let

$$\underline{f}(i\tau_i(j)) = \left(f_{i\tau_i(j)}(1), f_{i\tau_i(j)}(2)\dots, f_{i\tau_i(j)}(\widetilde{n}_{i\tau_i(j)})\right); \tag{43}$$

for all $i \in E$, let

$$\underline{f}(i) = \left(\underline{f}(i\tau_i(1)), \underline{f}(i\tau_i(2)), \dots, \underline{f}(i\tau_i(c_i))\right). \tag{44}$$

Then the arrangement of parameters $(f_{ij}(k))$ of conditional sojourn times can be represented as follows:

$$(f_{ij}(k)) = \left(\underline{f}(1), \underline{f}(2), \dots, \underline{f}(s)\right). \tag{45}$$

If we denote $\left(\frac{\partial \underline{f}(i_1)}{\partial \underline{q}(i_2)}\right) := \left(\frac{\partial f_{i_1j_1}(k_1)}{\partial q_{i_2j_2}(k_2)}\right) = T_1^{'}$ and $F^{(i)} := \left(\frac{\partial \underline{f}(i)}{\partial \underline{q}(i)}\right)$,
then

$$T_1^{'} = \operatorname{diag}\{F^{(i)}, i \in E\}, \tag{46}$$

where

$$F^{(i)} = \begin{pmatrix} F^{(i)}_{11} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & F^{(i)}_{22} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & F^{(i)}_{c_i-1,c_i-1} & \mathbf{0} \\ F^{(i)}_{c_i1} & F^{(i)}_{c_i2} & \dots & F^{(i)}_{c_i,c_i-1} & F^{(i)}_{c_ic_i} \end{pmatrix}, \tag{47}$$

and the matrices $F^{(i)}_{j_1j_2} := \left(\frac{\partial \underline{f}(i\tau_i(j_1))}{\partial \underline{q}(i\tau_i(j_2))}\right)$ for different values of j_1 and j_2 which correspond to nonzero matrices in (47) are given by

$$F^{(i)}_{jj} = -\frac{1}{p^2_{i\tau_i(j)}} \begin{pmatrix} -\sum\limits_{k\neq 1} q_{i\tau_i(j)}(k) & q_{i\tau_i(j)}(1) & \dots & q_{i\tau_i(j)}(1) \\ q_{i\tau_i(j)}(2) & -\sum\limits_{k\neq 2} q_{i\tau_i(j)}(k) & \dots & q_{i\tau_i(j)}(2) \\ \vdots & \vdots & \ddots & \vdots \\ q_{i\tau_i(j)}(\widetilde{n}_{i\tau_i(j)}) & q_{i\tau_i(j)}(\widetilde{n}_{i\tau_i(j)}) & \dots & -\sum\limits_{k\neq \widetilde{n}_{i\tau_i(j)}} q_{i\tau_i(j)}(k) \end{pmatrix}, \tag{48}$$

$$F^{(i)}_{c_ij} = \frac{1}{p^2_{ij_i}} \begin{pmatrix} q_{ij_i}(1) & q_{ij_i}(1) & \dots & q_{ij_i}(1) \\ \vdots & \vdots & \ddots & \vdots \\ q_{ij_i}(\widetilde{n}_i - 1) & q_{ij_i}(\widetilde{n}_i - 1) & \dots & q_{ij_i}(\widetilde{n}_i - 1) \\ -\sum\limits_{k\neq \widetilde{n}_i} q_{ij_i}(k) & -\sum\limits_{k\neq \widetilde{n}_i} q_{ij_i}(k) & \dots & -\sum\limits_{k\neq \widetilde{n}_i} q_{ij_i}(k) \end{pmatrix}, \tag{49}$$

and

$$F^{(i)}_{c_ic_i} = \frac{1}{p_{ij_i}} \begin{pmatrix} \mathbf{I}_{s_i} \\ -\underline{1} \end{pmatrix}, \text{ where } \ s_i = c^+_{ij_i}. \tag{50}$$

Since $\sqrt{n}\left((\widehat{f}_{ij}(k)) - (f^0_{ij}(k))\right) = \sqrt{n}\left(T_1(\widehat{q}_{ij}(k,n)) - T_1(q^0_{ij}(k))\right)$, we use Theorem 3, the differentiability of T_1 on Θ_1^*, and apply the delta method to conclude that

$$\sqrt{n}\left((\widehat{f}_{ij}(k)) - (f^0_{ij}(k))\right) \to \mathcal{N}(0, T_1^{'}\Psi_1^{'}\sigma(\theta_0)^{-1}(T_1^{'}\Psi_1^{'})^\top),$$

where

$$T_1^{'}\Psi_1^{'} = \operatorname{diag}\{F^{(i)}M^{(i)}, i \in E\} \tag{51}$$

and $F^{(i)}$ and $M^{(i)}$ are given by (47) and (27), respectively. The explicit form of these matrices for all $i \in E$ is as follows:

$$F^{(i)}M^{(i)} = \begin{pmatrix} D_{11}^{(i)} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & D_{22}^{(i)} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & D_{c_i-1,c_i-1}^{(i)} & \mathbf{0} \\ D_{c_i1}^{(i)} & D_{c_i2}^{(i)} & \dots & D_{c_i,c_i-1}^{(i)} & D_{c_i,c_i}^{(i)} \end{pmatrix}, \tag{52}$$

where

$$D_{jj}^{(i)} = -\frac{1}{p_{i\tau_i(j)}^2} \begin{pmatrix} -\sum\limits_{k\neq 1} q_{i\tau_i(j)}(k) & a_i(1)q_{i\tau_i(j)}(1) & \dots & a_i(c_{ij}^+)q_{i\tau_i(j)}(1) \\ q_{i\tau_i(j)}(2) & -a_i(1)\sum\limits_{k\neq 2} q_{i\tau_i(j)}(k) & \dots & a_i(c_{ij}^+)q_{i\tau_i(j)}(2) \\ \vdots & \vdots & \ddots & \vdots \\ q_{i\tau_i(j)}(\widetilde{n}_{i\tau_i(j)}) & a_i(1)q_{i\tau_i(j)}(\widetilde{n}_{i\tau_i(j)}) & \dots & -a_i(c_{ij}^+)\sum\limits_{k\neq \widetilde{n}_{i\tau_i(j)}} q_{i\tau_i(j)}(k) \end{pmatrix}, \tag{53}$$

$$D_{c_ij}^{(i)} = \frac{1}{p_{ij_i}^2} \begin{pmatrix} -\sum\limits_{k\neq 1} q_{ij_i}(k) & a_i(1)q_{ij_i}(1) & \dots & a_i(c_{ij}^+)q_{ij_i}(1) \\ q_{ij_i}(2) & -a_i(1)\sum_{k\neq 2} q_{ij_i}(k) & \dots & a_i(c_{ij}^+)q_{ij_i}(2) \\ \vdots & \vdots & \ddots & \vdots \\ q_{ij_i}(\widetilde{n}_i) & a_i(1)q_{ij_i}(\widetilde{n}_i) & \dots & -a_i(c_{ij}^+)\sum\limits_{k\neq \widetilde{n}_i} q_{ij_i}(k) \end{pmatrix}, \tag{54}$$

and

$$D_{c_i,c_i}^{(i)} = \sum_{j=1}^{c_i} F_{c_ij}^{(i)} M_{jc_i}^{(i)} \tag{55}$$

for $1 \le j \le c_i - 1$, and $F_{c_ij}^{(i)}$ and $M_{jc_i}^{(i)}$ are given by (34)–(35) and (49)–(50), respectively.

REFERENCES

1. V. Barbu and N. Limnios, "Maximum likelihood estimation for hidden semi-Markov models," *C. R. Acad. Sci. Paris*, **342**, 201–205 (2006).
2. V. Barbu and N. Limnios, *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications*, Springer (2008).
3. L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, **37**, 1554–1563 (1966).
4. P. J. Bickel, Y. Ritov, and T. Ryden, "Asymptotic normality of the maximum-likelihood estimator for general hidden-Markov models," *Ann. Statist.*, **26**, 1614–1635 (1998).
5. O. Chryssaphinou, M. Karaliopoulou, and N. Limnios, "On discrete time semi-Markov chains and applications in words occurrences," *Communications in Statistics – Theory and Methods*, **37**, 1306–1322 (2008).
6. R. A. L. Elliott and J. Moore, *Hidden Markov Models: Estimation and Control*, Springer, New York (1995).
7. Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inf. Theory*, **48**, 1518–1569 (2002).
8. J. Ferguson, "Variable duration models for speech," in: *Proc. Symp. on the Application of Hidden Markov Models to Text and Speech*, Princeton, New Jersey, (1980), 143–179.
9. Y. Guédon, "Estimating hidden semi-Markov chains from discrete sequences," *J. of Computational and Graphical Statistics*, **12** (3), 604–639 (2003).
10. J. Kiefer and J. Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters," *Ann. Math. Stat.*, **27**, 887–906 (1956).
11. A. Krogh, M. Brown, I. S. Mian, K. Sjlander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *J. Molecular Biology*, **235**, 1501–1531 (1994).
12. B. G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Process. Appl.*, **40**, 127–143 (1992).
13. N. Limnios and G. Oprisan, *Semi-Markov Processes and Reliability*, Birkhäuser, Boston (2001).

14. T. Petrie, “Probabilistic functions of finite state Markov chains,” *Ann. Math. Stat.*, **40**, 97–115 (1969).
15. L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, **77**, 257–284 (1989).
16. L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs (1993).
17. J. Sansom and P. Thomson, “Fitting hidden semi-Markov models to break-point rainfall data,” *J. of Applied Probability*, **38A**, 142–157 (2001).