Context Tree Estimation in Variable Length Hidden Markov Models

Thierry Dumont

Context Tree Estimation in Variable Length Hidden Markov Models

Thierry Dumont
MODAL'X
Université Paris-Ouest, Nanterre, France.
Email: thierry.dumont@u-paris10.fr

***Abstract*—We address the issue of context tree estimation in variable length hidden Markov models. We propose an estimator of the context tree of the hidden Markov process which needs no prior upper bound on the depth of the context tree. We prove that the estimator is strongly consistent. This uses information-theoretic mixture inequalities in the spirit of [1], [2]. We propose an algorithm to efficiently compute the estimator and provide simulation studies to support our result.**

***Index Terms*—Variable length, hidden Markov models, context tree, consistent estimator, mixture inequalities.**

I. Introduction

A variable length hidden Markov model (VLHMM) is a bivariate stochastic process $(X_n, Y_n)_{n\geq 0}$ where $(X_n)_{n\geq 0}$ (the state sequence) is a variable length Markov chain (VLMC) in a state space $\mathbb{X}$ and, conditionally on $(X_n)_{n\geq 0}$, $(Y_n)_{n\geq 0}$ is a sequence of independent variables in an observation space $\mathbb{Y}$ such that the conditional distribution of Y_n given the state sequence (called the emission distribution) depends on X_n only. Such processes fall into the general framework of latent variable processes, and reduce to hidden Markov models (HMM) when the state sequence is a Markov chain. Latent variable processes are used as a flexible tool to model dependent non-Markovian time series, and the statistical problem is to estimate the parameters of the distribution when only $(Y_n)_{n\geq 0}$ is observed. We assume that the hidden process may only take a fixed and known number of values, that is the case where the state space $\mathbb{X}$ is finite with known cardinality k.

The dependence structure of a latent variable process is driven by that of the hidden process $(X_n)_{n\geq 0}$, which is assumed here to be a variable length Markov chain (VLMC). Such processes were first introduced by Rissanen in [3] as a flexible and parsimonious modelization tool for data compression. Recall that a Markov process of order d is such that the conditional distribution of X_n given all past values depends only on the d previous ones $X_{n-1}, \ldots, X_{n-d}$. Nevertheless different past may lead to identical conditional distributions, so that all k^d possible past values are not needed to describe the distribution of the process. A VLMC is such that the probability of the present state depends only on a finite part of the past, and the length of this relevant portion, called context, is a function of the past itself. No context may be a proper suffix of any other context, so that the set of all contexts may be represented as a rooted labelled tree. This set is called the context tree of the VLMC.

Variable length hidden Markov models appear for the first time, to our knowledge, in movement analysis [4], [5]. Human movement analysis is the interpretation of movements as sequences of poses. [5] analyses the movement through 3D rotations of 19 major joints of human body. [4] then use a VLHMM representation where X_n is the pose at time n and Y_n is the body position given by the 3D rotations of the 19 major points. They argue that "VLHMM is superior in its efficiency and accuracy of modeling multivariate time-series data with highly-varied dynamics". [6] studies VLHMM from a theoretical point of view, when the observation process $(Y_n)_{n\geq 0}$ consists in a Bernoulli type perturbation of a binary VLMC $\{X_n\}_{n\geq 0}$ and shows that it is possible to recover the context tree of the chain $\{X_n\}_{n\geq 0}$, using a suitable version of Rissanen's Context algorithm, provided that the Bernoulli noise is small enough.

VLHMM could also be used in WIFI based indoor positioning systems (see [7]). Here X_n is a mobile device position at time n and Y_n is the received signal strength (RSS) vector at time n. Each component of the RSS vector represents the strength of a signal sent by a WIFI access point. In practice, the aim is to estimate the positions of the device $(X_n)_{n\geq 0}$ on the basis of the observations $(Y_n)_{n\geq 0}$. The distribution of Y_n given $X_n = x$ for any location x is beforehand calibrated for a finite number of locations $(L_1, ..., L_k)$. A Markov chain on the finite set $(L_1, ..., L_k)$ is then used to model the sequence of positions $(X_n)_{n\geq 0}$. Again VLHMM model would lead to efficient and accurate estimation of the device position.

The aim of this paper is to provide a statistical analysis of variable length hidden Markov models and, in particular, to propose a consistent estimator of the context tree of the hidden VLMC on the basis of the observations $(Y_n)_{n\geq 0}$ only. We consider a parametrized family of VLHMM, and we use a penalized likelihood method to estimate the context tree of the hidden VLMC. To each possible context tree τ, if Θ_τ is the set of possible parameters, we define

$$\hat{\tau}_n = \underset{\tau}{\operatorname{argmin}} \left\{ - \sup_{\theta \in \Theta_\tau} \log \ell_\theta(Y_{1:n}) + pen(n, \tau) \right\} ,$$

where $\ell_\theta(y_{1:n})$ is the distribution density of the observation $Y_{1:n} = (Y_1, \ldots, Y_n)$ under the parameter θ, with respect to some dominating positive measure. $pen(n, \tau)$ is a penalty that

This work was partially supported by ID Services, 22/24 rue Jean Rostand 91400 Orsay, FRANCE

depends on the number n of observations and on the context tree τ. Our aim is to find penalties for which the estimator is strongly consistent without any prior upper bound on the depth of the context tree, and to provide a practical algorithm to compute the estimator.

Context tree estimation for a VLHMM is similar to order estimation for a HMM in which the order is defined as the unknown cardinality of the state space $\mathbb{X}$. The main difficulty lies in the calibration of the penalty, which requires some understanding of the growth of the likelihood ratios (with respect to orders and to the number of observations). In particular cases, the fluctuations of the likelihood ratios may be understood via empirical process theory, see the recent works [8] for finite state Markov chains and [9] for independent identically distributed observations. Latent variable models are much more complicated, see for instance [10] where it is proved in the HMM situation that the likelihood ratio statistics converges to infinity for overestimated order. We thus use an approach based on information theory tools to understand the behavior of likelihood ratios. Such tools have been successfull for HMM order estimation problems and were used in [2], [1] for discrete observations and in [11] for Poisson emission distributions or Gaussian emission distributions with known variance. Our main result shows that for a penalty of the form $C(\tau)\log n$, $\hat{\tau}_n$ is strongly consistent, that is converges almost surely to the true unknown context tree. Here, $C(\tau)$ has an explicit formulation but is slightly bigger than $(k-1)|\tau|/2$ which gives the popular BIC penalty. We study the important situation of Gaussian emissions with unknown variance, which means that the distributions of Y_0 conditionally on $X_0 = x$, $x \in \mathbb{X}$, are Gaussian with the same unknown variance σ^2 and unknown mean m_x depending on the hidden state x. To our knowledge, this case has not been studied in previous works on HMM order estimation. We prove that our consistency theorem holds in this case.

Computation of the estimator requires computation of the maximum likelihood for all possible context trees. As usual, the EM algorithm may be used to compute the maximum likelihood estimator for the parameters when the context tree is fixed. We propose a suboptimal algorithm to compute the estimator $\widehat{\tau}_n$, which prevents the exploration of a too large number of context trees by pruning an initial maximal context tree of depth $M > 0$, τ_M. In general, the EM algorithm needs to be run several times with different initial values to avoid local extremum traps. In the important situation of Gaussian emissions, we propose a way to choose the initial parameters so that only one run of the EM algorithm is needed. Simulations compare penalized maximum likelihood estimators of the context tree $\tau^\star$ of the hidden VLMC using our penalty and using the BIC penalty. While these simulations highlight the consistency of our estimator, they also indicate that the BIC estimator seems also consistent and reaches the right model for relatively small samples.

The structure of this paper is the following. Section II describes the model and introduces some notations. Section III presents the information theory tools that we use, states the main consistency result and applies it to Poisson emission distributions and Gaussian emission distributions with known variance. Section IV proves the result for Gaussian emission distributions with unknown variance. In section V, we present the algorithm and provide simulation results. The proofs that are not essential at first reading are detailed in the Appendix.

II. Basic setting and notation

Throughout the sequel, the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is assumed to be rich enough to allow the existence of the random processes defined in the paper. Let $\mathbb{X}$ be a finite set whose cardinality is denoted by $|\mathbb{X}| = k$. We identify $\mathbb{X}$ with $\{1, \dots, k\}$. Let $\mathcal{F}_{\mathbb{X}}$ be the finite collection of subsets of $\mathbb{X}$. Let $\mathbb{Y}$ be a Polish space endowed with its Borel sigma-field $\mathcal{F}_{\mathbb{Y}}$. .

A. Context trees and variable length Markov chains

A string $s = x_k x_{k+1} ... x_l \in \mathbb{X}^{l-k+1}$ is denoted by $x_{k:l}$ and its length is then $l(s) = l - k + 1$. The concatenation of the strings u and v is denoted by uv. A string v is a *suffix* of a string s if there exists a string u such that $s = uv$. Note that, in the literature, the term postfix can be used instead. A set τ of strings and possibly semi-infinite sequences is called a *tree* if the following *tree property* holds : no $s \in \tau$ is suffix of any other $s' \in \tau$. If τ is a tree, the elements of τ will be referred to as leaves of the tree τ and, if $s = x_{1:l}$, $l \geq 2$, is a leaf of τ, for any $k = 2, \dots, l$, $x_{k:l}$ will be referred to as (internal) node of τ. The children of a node u of τ are the set of strings xu, $x \in \mathbb{X}$, such that xu is either a node or a leaf of τ. A tree τ is *irreducible* if no element $s \in \tau$ can be replaced by a suffix without violating the tree property. It is *complete* if each node has $|\mathbb{X}|$ children exactly. We denote by $d(\tau)$ the depth of τ: $d(\tau) = \max\{l(s) \mid s \in \tau\}$.

Let $X = (X_n)_{n\in\mathbb{Z}}$ be an ergodic and stationary process on $\mathbb{X}^{\mathbb{Z}}$. For any $m \leq n$ and any $x_{m:n}$ in $\mathbb{X}^{n-m+1}$, write $Q(x_{m:n})$ for $Q(X_{0:n-m} = x_{m:n})$. Throughout the sequel, "Q-almost every semi infinite sequence $x_{-\infty:-1}$" stands for "all semi infinite sequences $x_{-\infty:-1}$ such that for any $M \geq 0$, $Q(x_{-M:-1}) > 0$ "

Definition 1. *Let τ be a tree. τ is called a X-adapted context tree if, for Q-almost every semi infinite sequence $x_{-\infty:-1}$, there exists a unique string s in τ such that s is suffix of $x_{-\infty:-1}$ and:*

$$\forall x_0 \in \mathbb{X},\ \mathbb{P}(X_0 = x_0 \big| X_{-\infty:-1} = x_{-\infty:-1}) = \mathbb{P}(X_0 = x_0 \big| X_{-l(s):-1} = s)\ , \quad (1)$$

Moreover, if for any $s \in \tau$, $Q(s) > 0$ and no proper suffix of s has the property (1), then τ is called the minimal context tree of $(X_n)_{n\in\mathbb{Z}}$ and $(X_n)_{n\in\mathbb{Z}}$ is called a variable length Markov chain (VLMC).

If a tree τ is X-adapted, then for Q-almost every $x_{-\infty:-1}$ there exists a unique string in τ suffix of $x_{-\infty:-1}$. We denote this suffix by $\tau(x_{-\infty:-1})$.
A tree τ is said to be a *subtree* of τ' if for each string s' in τ' there exists a string s in τ suffix of s'. Then if τ is a X-adapted tree, any tree τ' such that τ is a subtree of τ' will be X-adapted.

Definition 2. *Let τ_0 be the minimal context tree of a VLMC $(X_n)_{n\in\mathbb{Z}}$. There exists a unique complete tree $\tau^\star$ such that τ_0 is a subtree of $\tau^\star$ and*

$$|\tau^\star| = \min\left\{|\tau| \ : \ \tau \text{ is a complete tree and } \tau_0 \text{ is a subtree of } \tau\right\} .$$

$\tau^\star$ is called the minimal complete context tree of the distribution Q of the VLMC $(X_n)_{n\in\mathbb{Z}}$.

In the sequel we only consider ergodic stationary VLMC, $(X_n)_{n\in\mathbb{Z}}$, whose minimal complete context tree $\tau^\star$ is finite, $(X_n)_{n\in\mathbb{Z}}$ is then a Markov chain of order $d(\tau^\star)$.

Define, for any complete tree τ, the set of transition parameters:

$$\Theta_{t,\tau} = \left\{ (P_{s,i})_{s\in\tau, i\in\mathbb{X}} \ : \ \forall s\in\tau, \forall i\in\mathbb{X}, \ P_{s,i}\geq 0 \text{ and } \sum_{i=1}^{k} P_{s,i} = 1 \right\} .$$

If $(X_n)_{n\in\mathbb{Z}}$ is a VLMC with minimal complete context tree $\tau^\star$ and transition parameters $\theta_t^\star = \left(P^\star_{s,i}\right)_{s\in\tau^\star, i\in\mathbb{X}} \in \Theta_{t,\tau^\star}$, for any complete tree τ such that $\tau^\star$ is a subtree of τ, there exists a unique $\theta_t = (P_{s,i})_{s\in\tau,i\in\mathbb{X}} \in \Theta_{t,\tau}$ that defines the same VLMC transition probabilities, namely: for any $s\in\tau$, there exists a unique $u\in\tau^\star$ which is a suffix of s, and for all $i\in\mathbb{X}$, $P_{s,i} = P^\star_{u,i}$. Notice that a transition parameter θ_t in $\Theta_{t,\tau}$ does not necessarily define a unique VLMC distribution (possibly, θ_t may not define a unique stationary distribution). θ_t defines a unique VLMC distribution if the extended Markov chain $(X_{n-d(\tau)+1:n})_{n\in\mathbb{Z}}$ is irreducible under the parameter θ_t.

B. Variable length hidden Markov models

A variable length hidden Markov model (VLHMM) is a bivariate stochastic process $(X_n, Y_n)_{n\in\mathbb{Z}}$ such that:

- $(X_n)_{n\in\mathbb{Z}}$ (the state sequence) a VLMC with values in $\mathbb{X}$,
- $(Y_n)_{n\in\mathbb{Z}}$ (the observation sequence) is a stochastic process such that the variables $Y_{-n},\dots,Y_n$, $n\geq 0$ are independent conditionally on the state sequence and such that for any $n\in\mathbb{Z}$, the conditional distribution of Y_n given the state sequence (called the emission distribution) depends on X_n only.

We assume that the emission distribution associated with x, namely: the distribution of Y_0 conditionally on $X_0 = x$, $x\in\mathbb{X}$, are absolutely continuous with respect with some positive measure μ defined on a measurable space $(\mathbb{Y},\mathcal{F}_\mathbb{Y})$ and are parametrized by a set $\Theta_e \subset (\mathbb{R}^{d_e})^k \times \mathbb{R}^{m_e}$, so that the set of possible emission densities is $\{(g_{\theta_{e,x},\eta}(.))_{x\in\mathbb{X}}, \ \theta_e = (\theta_{e,1},\dots,\theta_{e,k},\eta)\in\Theta_e\}$. For any complete tree τ, define the parameter set:

$$\Theta_\tau = \Theta_{t,\tau}\times\Theta_e .$$

Define, for $\theta = (\theta_t,\theta_e)\in\Theta_\tau$, $\mathbb{P}_\theta$ the probability of the VLHMM $(X_n,Y_n)_{n\geq 0}$ under the parameter θ: $(X_n)_{n\in\mathbb{Z}}$ is a VLMC with complete context tree τ, transition parameter θ_t, and for any $(u_1,u_2)\in\mathbb{N}^2$, $u_1\leq u_2$, any sets $A_{u_1},\dots,A_{u_2}$ in $\mathcal{F}_Y$, any $x_{u_1:u_2}\in\mathbb{X}^{u_2-u_1+1}$,

$$\mathbb{P}_\theta\left(Y_{u_1}\in A_{u_1},\dots,Y_{u_2}\in A_{u_2}\middle| X_{u_1} = x_{u_1},\dots,X_{u_2} = x_{u_2}\right) = \prod_{u=u_1}^{u_2}\left[\int_{A_u} g_{\theta_{e,x_u},\eta}(y)\mathrm{d}\mu(y)\right] .$$

Throughout the paper we shall assume that the observations $(Y_1,...,Y_n) = Y_{1:n}$ consist in a sample of a VLHMM with true parameter $\theta^\star$ such that $\tau^\star$ is the minimal *complete* context tree associated with the hidden VLMC, and such that $([X_{n-d(\tau^\star)+1},\dots,X_n])_{n\in\mathbb{Z}}$ is a stationary and irreducible Markov chain. In order to define a computable likelihood, we introduce, for any positive integer d, a probability distribution ν_d on $\mathbb{X}^d$ so that, for any complete tree τ and any $\theta = (\theta_t,\theta_e)\in\Theta_\tau$, we set what will be called the likelihood:

$$\forall y_{1:n}\in\mathbb{Y}^n, \ \ell_\theta(y_{1:n}) = \sum_{x_{1:n}\in\mathbb{X}^n}\left[\prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(y_i)\right] q_{\theta_t}(x_{1:n}) ,$$

where, if $\theta_t = (P_{s,x})_{s\in\tau,x\in\mathbb{X}}$,

$$q_{\theta_t}(x_{1:n}) = \sum_{x_{-d(\tau)+1:0}\in\mathbb{X}^{d(\tau)}}\left[\nu_{d(\tau)}(x_{-d(\tau)+1:0}) \cdot\prod_{i=1}^{n} P_{\tau(x_{-d(\tau)+i:i-1}),x_i}\right] .$$

We are concerned with the statistical estimation of the tree $\tau^\star$ using a method that involves no prior upper bound on the depth of $\tau^\star$. Define the following estimator of the minimal complete context tree $\tau^\star$:

$$\hat{\tau}_n = \underset{\tau \text{ complete tree}}{\operatorname{argmin}}\left[-\sup_{\theta\in\Theta_\tau}\log\ell_\theta(Y_{1:n}) + pen(n,\tau)\right] , \quad (2)$$

where $pen(n,\tau)$ is a penalty term depending on the number of observations n and the complete tree τ.
The label switching phenomenon occurs in statistical inference of VLHMM as it occurs in statistical inference of HMM and of population mixtures. That is: applying a label permutation on $\mathbb{X}$ does not change the distribution of $(Y_n)_{n\geq 0}$. Thus, if σ is a permutation of $\{1,...,k\}$ and τ is a complete tree, we define the complete tree $\sigma(\tau)$ by

$$\sigma(\tau) = \left\{\sigma(x_1)...\sigma(x_l)\middle| \ x_{1:l}\in\tau\right\} .$$

Definition 3. *If τ and τ' are two complete trees, we say that τ and τ' are equivalent, and denote it by $\tau\sim\tau'$, if there exists a permutation σ of $\mathbb{X}$ such that $\sigma(\tau) = \tau'$.*

We then choose $pen(n,\tau)$ to be invariant under permutation, that is: for any permutation σ of $\mathbb{X}$, $pen(n,\sigma(\tau)) = pen(n,\tau)$. In this case, for any complete tree τ,

$$-\sup_{\theta\in\Theta_{\hat{\tau}_n}}\log\ell_\theta(Y_{1:n}) + pen(n,\tau) = -\sup_{\theta\in\Theta_{\sigma(\hat{\tau}_n)}}\log\ell_\theta(Y_{1:n}) + pen(n,\sigma(\tau)) ,$$

so that the definition of $\hat{\tau}_n$ requires a choice in the set of minimizers of (2).

Our aim is now to find penalties allowing to prove the strong consistency of $\hat{\tau}_n$, that is such that $\hat{\tau}_n \sim \tau^\star$, $\mathbb{P}_{\theta^\star}$- eventually almost surely as $n \to \infty$.

III. The general strong consistency theorem

In this section, we first recall the tools borrowed from information theory, and set the result that we use in order to find a penalty insuring the strong consistency of $\hat{\tau}_n$. Then we give our general strong consistency theorem, and straightforward applications. Application to Gaussian emissions with unknown variance, which is more involved, is deferred to the next section.

A. An information theoretic inequality

We shall introduce mixture probability distributions on $\mathbb{Y}^n$ and compare them to the maximum likelihood, in the same way as [12] first did; see also [13] and [14] for tutorials and use of such ideas in statistical methods. For any complete tree τ, we define, for all positive integer n, the mixture measure $\mathbb{KT}^n_\tau$ on $\mathbb{Y}^n$ using a prior π^n on Θ_τ:

$$\pi^n(\mathrm{d}\theta) = \pi_t(\mathrm{d}\theta_t) \otimes \pi^n_e(\mathrm{d}\theta_e) \ ,$$

where π^n_e is a prior on Θ_e that may change with n, and π_t the prior on Θ_t such that, if $\theta_t = (P_{s,i})_{s\in\tau, i\in\mathbb{X}}$,

$$\pi_t(\mathrm{d}\theta_t) = \otimes_{s\in\tau}\pi_s(\mathrm{d}(P_{s,i})_{i\in\mathbb{X}}) \ ,$$

where $(\pi_s)_{s\in\tau}$ are Dirichlet $\mathcal{D}(\frac{1}{2}, ..., \frac{1}{2})$ distributions on $[0,1]^{|\mathbb{X}|}$. Then $\mathbb{KT}^n_\tau$ is defined on $\mathbb{Y}^n$ by

$$\mathbb{KT}^n_\tau(y_{1:n}) = \sum_{x_{1:n}\in\mathbb{X}^n} \mathbb{KT}_{\tau,t}(x_{1:n})\mathbb{KT}^n_e(y_{1:n}|x_{1:n}) \ ,$$

where

$$\mathbb{KT}^n_e(y_{1:n}|x_{1:n}) = \int\limits_{\Theta_e} \left[\prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i)\right] \pi^n_e(\mathrm{d}\theta_e) \ ,$$

and

$$\begin{aligned}\mathbb{KT}_{\tau,t}(x_{1:n}) &= \left(\frac{1}{k}\right)^{d(\tau)} \int_{\Theta_t} \mathbb{P}_{\theta_t}\left(x_{d(\tau)+1:n}|x_{1:d(\tau)}\right)\pi_t(\mathrm{d}\theta_t) \\ &= \left(\frac{1}{k}\right)^{d(\tau)} \prod_{s\in\tau} \int\limits_{[0,1]^{|\mathbb{X}|}} \prod_{i=1}^k P_{s,i}^{a^x_s(x_{1:n})} \pi_s(\mathrm{d}(P_{s,i})_{i\in\mathbb{X}}) \ ,\end{aligned}$$

where $a^x_s(x_{1:n})$ is the number of times that x appears in context s, that is $a^x_s(x_{1:n}) = \sum_{i=d(\tau)+1}^n \mathbf{1}_{x_i=x, x_{i-l(s),i-1}=s}$.

The following inequality will be a key tool to control the fluctuations of the likelihood.

Proposition 1. *There exists a finite constant D depending only on k such that for any complete tree τ, and any $y_{1:n} \in \mathbb{Y}^n$:*

$$\begin{aligned}0 \leq \sup_{\theta\in\Theta_\tau} \log \ell_\theta(y_{1:n}) - \log \mathbb{KT}^n_\tau(y_{1:n}) \leq \\ \sup_{x_{1:n}} \left[\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i) - \log \mathbb{KT}^n_e(y_{1:n}|x_{1:n})\right] \\ + \frac{k-1}{2}|\tau|\log n + D \ .\end{aligned}$$

Proof: Let τ be a complete tree. For any $\theta \in \Theta_\tau$,

$$\begin{aligned}\frac{\ell_\theta(y_{1:n})}{\mathbb{KT}^n_\tau(y_{1:n})} &= \frac{\sum\limits_{x_{1:n}} q_{\theta_t}(x_{1:n})\prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i)}{\sum\limits_{x_{1:n}} \mathbb{KT}_\tau(x_{1:n})\mathbb{KT}^n_e(y_{1:n}|x_{1:n})} \ , \\ &\leq \max_{x_{1:n}} \frac{q_{\theta_t}(x_{1:n})\prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i)}{\mathbb{KT}_\tau(x_{1:n})\mathbb{KT}^n_e(y_{1:n}|x_{1:n})} \ .\end{aligned}$$

Thus, using [14],

$$\begin{aligned}\log \frac{\ell_\theta(y_{1:n})}{\mathbb{KT}^n_\tau(y_{1:n})} \leq \sup_{x_{1:n}} \Bigg[\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i) \\ - \log \mathbb{KT}^n_e(y_{1:n}|x_{1:n}) + |\tau|\gamma\left(\frac{n}{|\tau|}\right) + d(\tau)\log k \Bigg] \ ,\end{aligned}$$

where γ is defined, for any positive x, by $\gamma(x) = \frac{k-1}{2}\log x + \log k$ and satisfies that $|\tau|\gamma\left(\frac{n}{|\tau|}\right) + d(\tau)\log k$ is an upper bound of

$$\sup_{\theta_t\in\Theta_{t,\tau}} \sup_{x_{1:n}} \log \frac{q_{\theta_t}(x_{1:n})}{\mathbb{KT}_\tau(x_{1:n})} \ .$$

Then

$$\begin{aligned}\log \frac{\ell_\theta(y_{1:n})}{\mathbb{KT}^n_\tau(y_{1:n})} \leq \sup_{x_{1:n}} \Bigg[\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i) \\ - \log \mathbb{KT}^n_e(y_{1:n}|x_{1:n}) \Bigg] + \frac{k-1}{2}|\tau|\log n + D(\tau) \ ,\end{aligned}$$

where $D(\tau) = -\frac{k-1}{2}|\tau|\log|\tau| + |\tau|\log k + d(\tau)\log k$. Now, since τ is complete, $d(\tau) \leq \frac{|\tau|-k}{k-1}$, so that

$$D(\tau) \leq |\tau|\Big(\log k - \frac{k-1}{2}\log|\tau|\Big) + \frac{|\tau|-k}{k-1}\log k \ .$$

But the upper bound in the inequality tends to $-\infty$ when $|\tau|$ tends to ∞, so that there exists a constant D depending only on k such that for any complete tree τ, $D(\tau) \leq D$. ■

B. Strong consistency theorem

Let $\theta^\star = (\theta^\star_t, \theta^\star_e)$ with $\theta^\star_t = (P^\star_{s,i})_{s\in\tau^\star, i\in\mathbb{X}}$, and $\theta^\star_e = (\theta^\star_{e,1}, ..., \theta^\star_{e,k}, \eta^\star)$ be the true parameters of the VLHMM. Let us now define for any positive α, the penalty:

$$pen_\alpha(n,\tau) = \left[\sum_{t=1}^{|\tau|} \frac{(k-1)t+\alpha}{2}\right]\log n \ . \qquad (3)$$

Notice that the complexity of the model is taken into account through the cardinality of the tree τ.

We need to introduce further assumptions.

- **(A1)**. For any complete tree τ such that $|\tau| \leq |\tau^\star|$ and such that τ and $\tau^\star$ are not equivalent, for any $\theta \in \Theta_\tau$, the random sequence $(\theta_{e,X_n})_{n\in\mathbb{Z}}$ where $(X_n)_{n\in\mathbb{Z}}$ is a VLMC with transition probabilities θ_t, has a different distribution than $(\theta^\star_{e,X_n})_{n\in\mathbb{Z}}$ where $(X_n)_{n\in\mathbb{Z}}$ is a VLMC with transition probabilities $\theta^\star_t$.

- **(A2)**. The family $\{g_{\theta_e}, \theta_e \in \Theta_e,\}$ is such that for any probability distributions $(\alpha_i)_{i=1,...,k}$ and $(\alpha'_i)_{i=1,...,k}$ on $\{1, ..., k\}$, any $(\theta_1, \ldots, \theta_k, \eta) \in \Theta_e$ and $(\theta'_1, \ldots, \theta'_k, \eta') \in \Theta_e$, if

$$\sum_{i=1}^{k} \alpha_i g_{\theta_i,\eta} = \sum_{i=1}^{k} \alpha'_i g_{\theta'_i,\eta'} \ ,$$

then,

$$\sum_{i=1}^{k} \alpha_i \delta_{\theta_i} = \sum_{i=1}^{k} \alpha'_i \delta_{\theta'_i} \text{ and } \eta = \eta' \ .$$

- **(A3)**. For any $y \in \mathbb{Y}$, $\theta_e \longmapsto g_{\theta_e}(y) = (g_{\theta_{e,i},\eta}(y))_{i\in\mathbb{X}}$ is continuous and tends to zero when $||\theta_e||$ tends to infinity.
- **(A4)**. For any $i \in \mathbb{X}$, $E_{\theta^\star}\left[|\log g_{\theta^\star_{e,i},\eta^\star}(Y_1)|\right] < \infty$.
- **(A5)**. For any $\theta_e \in \Theta_e$, there exists $\delta > 0$ such that : $E_{\theta^\star}\left[\sup_{||\theta'_e-\theta_e||<\delta} (\log g_{\theta'_e}(Y_1))^+\right] < \infty$.

Theorem 1. *Assume that* **(A1)** *to* **(A5)** *hold, and that moreover there exists a positive real number b such that*

$$\sup_{\theta_e\in\Theta_e} \sup_{x_{1:n}} \left[\log \prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right] \leq b \log n \ , \quad (4)$$

$\mathbb{P}_{\theta^\star}$ *- eventually almost surely. If one chooses* $\alpha > 2(b+1)$ *in the penalty (3), then* $\hat{\tau}_n \sim \tau^\star$, $\mathbb{P}_{\theta^\star}$ *- eventually almost surely.*

Notice that, to apply this theorem, one has to find a sequence of priors π_e^n on Θ_e such that (4) holds. The remaining of the section will prove that it is possible for situations in which priors may be defined as in previous works about HMM order estimation, while in the next section, we will prove that it is possible to find a prior in the important case of Gaussian emissions with unknown variance.
In the following proof, the assumption (4) insures that $|\hat{\tau}_n| \leq |\tau^\star|$ eventually almost surely, while assumptions **(A1-5)** insure that for any complete tree τ such that $|\tau| < |\tau^\star|$ or $|\tau| = |\tau^\star|$ and $\tau \nsim \tau^\star$, $\hat{\tau}_n \neq \tau^\star$ $\mathbb{P}_{\theta^\star}$ - eventually almost surely. In particular **(A1)** holds whenever $\theta^\star_{e,x} \neq \theta^\star_{e,y}$ if $(x, y) \in \mathbb{X}^2$ and $x \neq y$.

Proof: The proof will be structured as follows : we first prove that $\mathbb{P}_{\theta^\star}$ - eventually almost surely, $|\hat{\tau}_n| \leq |\tau^\star|$. We then prove that for any complete tree τ such that $|\tau| \leq |\tau^\star|$ and $\tau \nsim \tau^\star$, $\hat{\tau}_n \nsim \tau$ $\mathbb{P}_{\theta^\star}$ - eventually almost surely. This will end the proof since there is a finite number of such trees. For any $n \in \mathbb{N}$, we denote by E_n the event

$$E_n \ : \ \left[\sup_{\theta_e\in\Theta_e} \sup_{x_{1:n}} \left(\log \prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right) \leq b \log n \right] .$$

By using (4) and Borel-Cantelli Lemma, to get that $\mathbb{P}_{\theta^\star}$ - eventually almost surely, $|\hat{\tau}_n| \leq |\tau^\star|$, it is enough to show that

$$\sum_{n=1}^{\infty} \mathbb{P}_{\theta^\star}\left\{(|\hat{\tau}_n| > |\tau^\star|) \bigcap E_n\right\} < \infty.$$

Let τ be a complete tree such that $|\tau| > |\tau^\star|$. Using Proposition 1,

$$\begin{aligned}
&\mathbb{P}_{\theta^\star}\left\{(\hat{\tau}_n = \tau) \bigcap E_n\right\} \\
&\leq \mathbb{P}_{\theta^\star}\Bigg\{\Bigg(\sup_{\theta\in\Theta_\tau} \log \ell_\theta(Y_{1:n}) - pen_\alpha(n, \tau) \geq \\
&\qquad \log \ell_{\theta^\star}(Y_{1:n}) - pen_\alpha(n, \tau^\star)\Bigg) \bigcap E_n\Bigg\} , \\
&\leq \mathbb{P}_{\theta^\star}\Bigg\{\Bigg(\log \mathbb{KT}_\tau^n(y_{1:n}) + \sup_{\theta_e\in\Theta_e} \sup_{x_{1:n}} \Bigg[\log \prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) \\
&\qquad - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\Bigg] + \frac{k-1}{2}|\tau| \log n + D \\
&\qquad - \log \ell_{\theta^\star}(Y_{1:n}) + pen_\alpha(n, \tau^\star) - pen_\alpha(n, \tau) \geq 0\Bigg) \\
&\qquad \bigcap E_n\Bigg\} , \\
&\leq \mathbb{P}_{\theta^\star}\left\{\ell_{\theta^\star}(Y_{1:n}) \leq \mathbb{KT}_\tau^n(Y_{1:n})\right\} \exp\left(e_{\tau,n}\right) \ .
\end{aligned}$$

with

$$e_{\tau,n,} = \frac{k-1}{2}|\tau| \log n + b \log n + D + pen_\alpha(n, \tau^\star) - pen_\alpha(n, \tau) \ .$$

But

$$\begin{aligned}
e_{\tau,n} &= \frac{k-1}{2}|\tau| \log n + b \log n + D + \sum_{t=1}^{|\tau^\star|} \frac{(k-1)t + \alpha}{2} \log n \\
&\quad - \sum_{t=1}^{|\tau|} \frac{(k-1)t + \alpha}{2} \log n \ , \\
&= \frac{k-1}{2}|\tau| \log n + b \log n + D \\
&\quad - \sum_{t=|\tau^\star|+1}^{|\tau|} \frac{(k-1)t + \alpha}{2} \log n \ , \\
&\leq -\frac{\alpha}{2}\Big(|\tau| - |\tau^\star|\Big) \log n + b \log n + D \ ,
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{P}_{\theta^\star}\left\{(\hat{\tau}_n = \tau) \bigcap E_n\right\} &\leq e^{-\frac{\alpha}{2}\left(|\tau|-|\tau^\star|\right) \log n + b \log n + D} , \\
&= C.n^{-\frac{\alpha}{2}(|\tau|-|\tau^\star|)+b} ,
\end{aligned}$$

for some constant C. Thus

$$\mathbb{P}_{\theta^\star}\left\{(|\hat{\tau}_n| > |\tau^\star|) \bigcap E_n\right\} \leq C \sum_{t=|\tau^\star|+1}^{\infty} CT(t) n^{-\frac{\alpha}{2}(t-|\tau^\star|)+b} ,$$

where $CT(t)$ is the number of complete trees with t leaves. But using Lemma 2 in [15], $CT(t) \leq 16^t$ so that

$$\mathbb{P}_{\theta^\star}\left\{(|\hat{\tau}_n| > |\tau^\star|)\bigcap E_n\right\} \leq Cn^b 16^{|\tau^\star|}\sum_{t=1}^{\infty}\left[16n^{-\alpha/2}\right]^t , \\ = O(n^{-\alpha/2+b}) ,$$

which is summable if $\alpha > 2(b+1)$.

Let now τ be a tree such that $|\tau| \leq |\tau^\star|$ and $\tau \nsim \tau^\star$. Let τ_M be a complete tree such that τ and $\tau^\star$ are both a subtree of τ_M. Then, by setting for any integer $n \geq d(\tau_M) - 1$, $W_n = [X_{n-d(\tau_M)+1:n}]$, for any $\theta \in \Theta_\tau \cup \Theta_{\tau^\star}$, $(W_n, Y_n)_{n\in\mathbb{Z}}$ is a HMM under $\mathbb{P}_\theta$. Following the proof of Theorem 3 of [16], we obtain that there exists $K > 0$ such that $\mathbb{P}_{\theta^\star}$-eventually a.s.,

$$\frac{1}{n}\log \ell_{\theta^\star}(Y_{1:n}) - \sup_{\theta\in\Theta_\tau}\frac{1}{n}\log \ell_\theta(Y_{1:n}) \geq K ,$$

so that, $\mathbb{P}_{\theta^\star}$-eventually a.s.,

$$\log \ell_{\theta^\star}(Y_{1:n}) - pen(n,\tau^\star) - \sup_{\theta\in\Theta_\tau}\log \ell_\theta(Y_{1:n}) + pen(n,\tau) > 0 ,$$

which concludes the proof of Theorem 1. ■

C. Gaussian emissions with known variance

Here, we do not need the parameter η so we omit it. Then $\Theta_e = \{\theta_e = (m_1, \dots, m_k) \in \mathbb{R}^k\}$. The conditional likelihood is given, for any $\theta_e = (m_x)_{x\in\mathbb{X}}$ by

$$g_{\theta_{e,x}}(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(y-m_x)^2}{2\sigma^2}).$$

Proposition 2. *Assume* **(A1)**. *If one chooses* $\alpha > k+2$ *in the penalty (3),* $\hat{\tau}_n \sim \tau^\star$, $\mathbb{P}_{\theta^\star}$ *- eventually a.s.*

Proof:

The identifiability of the Gaussian model **(A2)** has been proved by Yakowitz and Spragins in [17], it is easy to see that Assumptions **(A3)** to **(A5)** hold. Now, we define the prior measure π_e^n on Θ_e as the probability distribution under which $\theta_e = (m_1, ..., m_k)$ is a vector of k independent random variables with centered Gaussian distribution with variance τ_n^2. Then, using [11], $\mathbb{P}_{\theta^\star}$ -eventually a.s.,

$$\sup_{\theta_e\in\Theta_e}\max_{x_{1:n}\in\mathbb{X}^n}\left[\log\prod_{i=1}^n g_{\theta_{e,x_i}}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right] \\ \leq \frac{k}{2}\log(1+\frac{n\tau_n^2}{k\sigma^2}) + \frac{k}{2\tau_n^2}5\sigma^2\log n .$$

Thus, by choosing $\tau_n^2 = \dfrac{5\sigma^2 k\log(n)}{2}$, we get that for any $\epsilon > 0$,

$$\sup_{\theta_e\in\Theta_e}\max_{x_{1:n}\in\mathbb{X}^n}\left[\log\prod_{i=1}^n g_{\theta_{e,x_i}}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right] \\ \leq \frac{k+\epsilon}{2}\log n ,$$

$\mathbb{P}_{\theta^\star}$ -eventually almost surely, and (4) holds for any $b > \dfrac{k}{2}$. ■

D. Poisson emissions

Now the conditional distribution of Y given $X = x$ is Poisson with mean m_x and

$$\Theta_e = \left\{\theta_e = (m_1, ..., m_k)\middle|\ \forall j \in \mathbb{X},\ m_j > 0\right\} .$$

Proposition 3. *Assume* **(A1)**. *If one chooses* $\alpha > k+2$ *in the penalty (3),* $\hat{\tau}_n \sim \tau^\star$ $\mathbb{P}$ *-eventually a.s.*

Proof:

The identifiability of the Poisson model **(A2)** has been proved by Teicher in [18], it is easy to see that Assumptions **(A3)** to **(A5)** hold. The prior π_e^n on Θ_e is now defined such that $m_1, ..., m_k$ are independent identically distributed with distribution Gamma(t, 1/2). Then, using [11]:

$$\sup_{\theta_e\in\Theta_e}\max_{x_{1:n}\in\mathbb{X}^n}\left\{\log\prod_{i=1}^n g_{\theta_{e,x_i}}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right\} \\ \leq \frac{k}{2}\log\frac{n}{k} + kt\frac{\log n}{\sqrt{\log\log n}} + \frac{k}{2}(1+t\log t) ,$$

$\mathbb{P}_{\theta^\star}$ -eventually a.s.. Then, for any fixed $t > 0$, for any $\epsilon > 0$, eventually almost surely :

$$\sup_{\theta_e\in\Theta_e}\max_{x_{1:n}\in\mathbb{X}^n}\left\{\log g_{\theta_e}(Y_{1:n}|x_{1:n}) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right\} \\ \leq \left(\frac{k}{2}+\epsilon\right)\log n ,$$

$\mathbb{P}_{\theta^\star}$ -eventually almost surely, and (4) holds for any $b > \dfrac{k}{2}$. ■

IV. Gaussian emissions with unknown variance

We consider the situation where the emission distributions are Gaussian with the same, but unknown, variance $\sigma^2 > 0$ and with a mean, $m_x \in \mathbb{R}$, depending on the hidden state x. Let $\eta = -\dfrac{1}{2\sigma^2}$ and $\theta_{e,j} = \dfrac{m_j}{\sigma^2}$ for all $j \in \mathbb{X} = \{1, .., k\}$. Here

$$\Theta_e = \left\{\left(\eta, (\theta_{e,j})_{j=1,\dots,k}\right)\ \middle|\ \theta_{e,j} \in \mathbb{R}, \eta < 0\right\} .$$

If $x_{1:n} \in \mathbb{X}^n$, for any $j \in \mathbb{X}$, we set $I_j = \{i | x_i = j\}$ and $n_j = |I_j|$. For sake of simplicity we omit $x_{1:n}$ in the notation though I_j and n_j depend on $x_{1:n}$. The conditional likelihood is given, for any $x_{1:n}$ in $\mathbb{X}^n$, for any $y_{1:n}$ in $\mathbb{Y}^n$, by

$$\prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i) = \\ \frac{1}{\sqrt{2\pi}^n}\prod_{j=1}^k \exp\left[\eta\sum_{i\in I_j} y_i^2 + \theta_{e,j}\sum_{i\in I_j} y_i - n_j A(\eta, \theta_{e,j})\right] ,$$

where

$$A(\eta, \theta_{e,j}) = -\frac{\theta_{e,j}^2}{4\eta} - \frac{1}{2}\log(-2\eta) .$$

Theorem 2. *Assume* **(A1)**. *If one chooses $\alpha > k+3$ in the penalty (3), then $\hat{\tau}_n \sim \tau^\star$, $\mathbb{P}_{\theta^\star}$ - eventually a.s.*

Proof: We shall prove that Theorem 1 applies. First, it is easy to see that Assumptions **(A3)** to **(A5)** hold and the proof of **(A2)** can be found in [17].
Define now the conjugate exponential prior on Θ_e :

$$\pi_e^n(\mathrm{d}\theta_e) = \exp\Bigg[\alpha_1^n \eta + \sum_{j=1}^k \alpha_{2,j}^n \theta_{e,j} - \sum_{j=1}^k \beta_j^n A\left(\eta, \theta_{e,j}\right)$$
$$- B\left(\alpha_1^n, \alpha_{2,1}^n, \dots, \alpha_{2,k}^n, \beta_1^n, \dots, \beta_k^n\right)\Bigg] \mathrm{d}\eta \mathrm{d}\theta_{e,1} \cdots \mathrm{d}\theta_{e,k} \ ,$$

where the parameters α_1^n, $(\alpha_{2,j}^n)_{j=1,\dots,k}$ and $(\beta_j^n)_{j=1,\dots,k}$ will be chosen later, and the normalizing constant may be computed as

$$\exp\left\{B\left(\alpha_1^n, \alpha_{2,1}^n, \dots, \alpha_{2,k}^n, \beta_1^n, \dots, \beta_k^n\right)\right\} = \frac{2^{k+\frac{\sum_{j=1}^k \beta_j^n}{2}} \pi^{\frac{k}{2}} \Gamma\left(\frac{\sum_{j=1}^k \beta_j^n + k + 2}{2}\right)}{\left(\prod_{j=1}^k \sqrt{\beta_j^n}\right)\left(\alpha_1^n - \sum_{j=1}^k \frac{(\alpha_{2,j}^n)^2}{\beta_j^n}\right)^{\frac{\sum_{j=1}^k \beta_j^n + k + 2}{2}}} \ ,$$

where we recall the Gamma function: $\Gamma(z) = \int_0^{+\infty} u^{z-1} e^{-u} \mathrm{d}u$ for any complex number z. Theorem 2 follows from Theorem 1 and the proposition below. ■

Proposition 4. *The parameters α_1^n, $(\alpha_{2,j}^n)_{j=1,\dots,k}$ and $(\beta_j^n)_{j=1,\dots,k}$ can be chosen such that for any $\epsilon > 0$,*

$$\max_{x_{1:n}}\left\{\sup_{\theta_e \in \Theta_e} \log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}\left(Y_i\right) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right\} \leq \frac{k+1+\epsilon}{2} \log n \ ,$$

$\mathbb{P}_{\theta^\star}$ - eventually a.s.

Proof: For any $x_{1:n} \in \mathbb{X}^n$, the parameters $\left(\widehat{\eta}, (\widehat{\theta}_{e,j})_j\right)$ maximizing the conditional likelihood are given by

$$\widehat{\eta} = -\frac{1}{2\widehat{\sigma}_{x_{1:n}}^2}, \ \widehat{\theta}_{e,j} = \frac{\widehat{m}_{x_{1:n},j}}{\widehat{\sigma}_{x_{1:n}}^2} \ ,$$

with

$$\widehat{m}_{x_{1:n},j} = \frac{\sum_{i \in I_j} Y_i}{n_j}, \ \widehat{\sigma}_{x_{1:n}}^2 = \frac{1}{n} \sum_{j=1}^k \sum_{i \in I_j} \left(Y_i - \widehat{m}_{x_{1:n},j}\right)^2 \ ,$$

so that

$$\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}\left(Y_i\right) \leq -n \log \widehat{\sigma}_{x_{1:n}} - \frac{n}{2} \log 2\pi - \frac{n}{2} \ .$$

Also,

$$\mathbb{KT}_e^n(y_{1:n}|x_{1:n}) = \frac{1}{\sqrt{2\pi}^n} \exp\Bigg[$$
$$B\left(\alpha_1^n + \sum_{i=1}^n Y_i^2, (\alpha_{2,j}^n + \sum_{i \in I_j} Y_i)_{1 \leq j \leq k}, (\beta_j^n + n_j)_{1 \leq j = 1 \leq k}\right)$$
$$- B\left(\alpha_1^n, (\alpha_{2,j}^n)_{1 \leq j \leq k}, (\beta_j^n)_{1 \leq j \leq k}\right)\Bigg] \ .$$

Recall that for all $z > 0$ (see for instance [19])

$$\sqrt{2\pi} e^{-z} z^{z-\frac{1}{2}} \leq \Gamma(z) \leq \sqrt{2\pi} e^{-z+\frac{1}{12z}} z^{z-\frac{1}{2}} \ ,$$

so that one gets that, for any $x_{1:n} \in \mathbb{X}^n$ and any $\theta_e \in \Theta_e$,

$$\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}\left(Y_i\right) - \log \mathbb{KT}_e^n(y_{1:n}|x_{1:n}) \leq o(\log n)$$
$$-\frac{n}{2} \log \widehat{\sigma}_{x_{1:n}}^2 - \frac{n}{2}\left(1 + \log 2\right) + \frac{k}{2} \log\left(\frac{n + \sum_{j=1}^k \beta_j^n}{k}\right)$$
$$-\Bigg[-\frac{n + \sum_{j=1}^k \beta_j^n + k + 2}{2}$$
$$+\left(\frac{n + \sum_{j=1}^k \beta_j^n + k + 1}{2}\right) \log \frac{n + \sum_{j=1}^k \beta_j^n + k + 2}{2}\Bigg]$$
$$+\frac{n + \sum_{j=1}^k \beta_j^n + k + 2}{2}$$
$$\cdot \log\left(\alpha_1^n + \sum_{i=1}^n Y_i^2 - \sum_{j=1}^k \frac{\left(\alpha_{2,j}^n + \sum_{i \in I_j} Y_i\right)^2}{n_j + \beta_j^n}\right) \ .$$

Choose now,

$$\forall j \in \{1, \dots, k\}, \ \beta_j^n = \alpha_{2,j}^n = \frac{1}{n} \ , \text{ and } \alpha_1^n = k+1 \ . \quad (5)$$

Then one easily gets that for any $x_{1:n} \in \mathbb{X}^n$ and any $\theta_e \in \Theta_e$,

$$\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}\left(Y_i\right) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n}) \leq o(\log n)$$
$$+\frac{n + k/n + k + 2}{2}$$
$$\cdot \log\left(1 + \frac{1}{n\widehat{\sigma}_{x_{1:n}}^2}\Bigg[k + 1 \quad + \sum_{j=1}^k \Bigg\{\widehat{m}_{x_{1:n},j}^2 \left(n_j - \frac{n_j^2}{n_j + 1/n}\right)\right.$$
$$\left. -2\frac{n_j}{n.n_j + 1} \widehat{m}_{x_{1:n},j} - \frac{1}{n^2 n_j + n}\Bigg\}\Bigg]\right)$$
$$+\frac{k+1}{2} \log n + \frac{k/n + k + 2}{2} \log \widehat{\sigma}_{x_{1:n}}^2 \ .$$

Let now $|Y|_{(n)} = \max_{1 \leq i \leq n} |Y_i|$. Then for any $x_{1:n} \in \mathbb{X}^n$,

$$\widehat{\sigma}_{x_{1:n}}^2 \leq |Y|_{(n)}^2 \text{ and } |\widehat{m}_{x_{1:n},j}| \leq |Y|_{(n)}, \ j = 1, \dots, k \ .$$

Also, for any partition $(I_i, \dots, I_k)$ of $\mathbb{R}$ in k intervals, define

:

$$\widehat{\sigma}^2_{I_i,\dots,I_k} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\mathbf{1}_{Y_i\in I_j}\left(Y_i - \frac{\sum\limits_{i'=1}^{n}\mathbf{1}_{Y_{i'}\in I_j}Y_{i'}}{\sum\limits_{i'=1}^{n}\mathbf{1}_{Y_{i'}\in I_j}}\right)^2 ,$$

and

$$\sigma^2_{I_i,\dots,I_k} = \sum_{j=1}^{k}\mathbb{P}_{\theta^\star}(Y_1\in I_k)Var_{\theta^\star}(Y_1|Y_1\in I_k) ,$$

where $Var_{\theta^\star}(Y_1|Y_1 \in I_k)$ is the conditional variance of Y_1 given that $Y_1 \in I_k$.

Lemma 1.

$$\inf_{x_{1:n}\in\mathbb{X}^n}\widehat{\sigma}^2_{x_{1:n}} = \inf_{I_i,\dots,I_k}\widehat{\sigma}^2_{I_i,\dots,I_k} ,$$

where the infimum in the right hand side of the equality is over all partitions of $\mathbb{R}$ *in* k *intervals.*

Proof: Indeed, given a set of observations $Y_{1:n}$ in $\mathbb{R}^n$, there exist optimal centroids, O_j, $j = 1,\dots,k$. The cluster of observations associated with each centroid O_j, $j = 1,\dots,k$, is its Voronoï cell, which is an interval of $\mathbb{R}$. Conversely, given a partition of $\mathbb{R}$ into interval $\{I_j \ : \ j = 1,\dots,k\}$, there exist centroids O_j, $j = 1,\dots,k$ such that $\{I_j \ : \ j = 1,\dots,k\}$ corresponds to the Voronoï partition associated with the centroids $\{O_j \ , \ j = 1,\dots,k\}$. ∎

We now get:

$$\begin{aligned}&\log\prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) - \log\mathbb{KT}^n_e(Y_{1:n}|x_{1:n}) \le o(\log n)\\ &+\frac{n+k/n+k+2}{2}\\ &\cdot\log\left(1+\frac{1}{n\inf\limits_{I_i,\dots,I_k}\widehat{\sigma}^2_{I_i,\dots,I_k}}\left[k+1\right.\right.\\ &\left.\left.+\sum_{j=1}^{k}\left[|Y|^2_{(n)}\left(n_j-\frac{n_j^2}{n_j+1/n}\right)+2\frac{n_j}{n.n_j+1}|Y|(n)\right]\right]\right)\\ &+\frac{k+1}{2}\log n+\frac{k/n+k+2}{2}\log|Y|^2_{(n)} ,\end{aligned}$$

and Proposition 4 follows from the choice (5) and the lemmas below, whose proofs are given in the Appendix. ∎

Lemma 2.

$$\sup_{I_i,\dots,I_k}\left|\widehat{\sigma}^2_{I_i,\dots,I_k} - \sigma^2_{I_i,\dots,I_k}\right|$$

converges to 0 *as* n *tends to infinity* $\mathbb{P}_{\theta^\star}$ *- a.s. (Here the supremum is over all partitions of* $\mathbb{R}$ *in* k *intervals). Also, the infimum* $s_{\inf}$ *of* $\sigma^2_{I_i,\dots,I_k}$ *over all partitions of* $\mathbb{R}$ *in* k *intervals satisfies* $s_{\inf} > 0$.

Lemma 3. $\mathbb{P}_{\theta^\star}$ *- eventually a.s.,* $|Y|^2_{(n)} \le 5\sigma^2_\star\log n$.

V. Algorithm and simulations

In this section we first present our practical algorithm. This algorithm is a suboptimal pruning algorithm, inspired by Rissanen's Context algorithm (see [3]). We then apply it in the case of Gaussian emissions with unknown common variance and compare our estimator with the BIC estimator that is when we choose in (2) the BIC penalty $pen(n,\tau) = \frac{k-1}{2}|\tau|\log n$.

A. Algorithm

We start this section with the definition of the terms used below :

- A maximal node of a complete tree τ is a string u such that, for any x in $\mathbb{X}$, ux belongs to τ. We denote by $N(\tau)$ the set of maximal nodes in the tree τ.
- The score of a complete tree τ on the basis of the observation $(Y_1,\dots,Y_n)$ is the penalized maximum likelihood associated with τ :

$$sc(\tau) = -\sup_{\theta\in\Theta_\tau}\log\ell_\theta(Y_{1:n}) + pen(n,\tau) . \qquad (6)$$

We also require that the emission model belongs to an exponential family such that :

(i) There exists $D \in \mathbb{N}^\star$, a function $s : \mathbb{X}\times\mathbb{Y} \longrightarrow \mathbb{R}^D$ of sufficient statistic and functions $h : \mathbb{X}\times\mathbb{Y} \longrightarrow \mathbb{R}$, $\psi : \Theta_e \longrightarrow \mathbb{R}^D$, and $A : \Theta_e \longrightarrow \mathbb{R}$, such that the emission density can be written as :

$$g_{\theta_e,x,\eta}(y) = h(x,y)\exp\left[\langle\psi(\theta_e), s(x,y)\rangle - A(\theta_e)\right] ,$$

where $\langle .,.\rangle$ denotes the scalar product in $\mathbb{R}^D$.

(ii) For all $S \in \mathbb{R}^D$, the equation :

$$\nabla_{\theta_e}\psi(\theta_e)S - \nabla_{\theta_e}A(\theta_e) = 0 ,$$

where ∇_{θ_e} denotes the gradient, has a unique solution denoted by $\bar{\theta}_e(S)$.

Assumption (ii) states that the function $\bar{\theta}_e : S \in \mathbb{R}^D \to \bar{\theta}_e(S) \in \Theta_e$ that returns the complete data maximum likelihood estimator corresponding to any feasible value of the sufficient statistics is available in closed-form.

The key idea of our algorithm is a "bottom to the top" pruning technique. Starting from the maximal complete tree of depth $M = \lfloor\log n\rfloor$, denoted by τ_M, we change each maximal node into a leaf whenever the resulting tree decreases the score.

Remark 1. *As our algorithm aims at computing* $\widehat{\tau}_n$ *using a pruning procedure starting from a maximal tree* τ_M*, we need to choose the depth* M *of* τ_M*. The choice* $M = \lfloor\log n\rfloor$ *is made so that when* n *is big enough, the true minimal context tree* $\tau^\star$ *is a subtree of* τ_M*, this choice only represents a restriction on the set of candidate trees. As definition* (2) *does not assume any prior bound on the depth of* $\widehat{\tau}_n$*, any other (larger) choice of* M *can be made, for instance* $M = n$*, so that we do not restrict the set of candidate trees (intrinsically,* $M = n$ *is the bound in the context tree maximising algorithm (see [15])). However, as we will see in this section, our algorithm is based on the EM algorithm and thus is prone to the convergence towards local minima of the limiting likelihood (see [20]). Choosing a larger* M*, in addition to increase the computation time of the algorithm,* *will increase the parameter space Θ_t maximal dimension and thus the amount of local minima.*

We start the algorithm by running several iterations of the EM algorithm. During this preliminary step we build estimators of sufficient statistics. These statistics will be used later in the computation of the maximum likelihood estimator $\hat{\theta}_\tau \in \Theta_\tau$ which realizes the supremum in (6) for any complete context tree τ subtree of τ_M.
For any $n \geq 0$, we denote by W_n the vectorial random sequence $W_n = (X_{n-M+1}, \ldots, X_n)$. For n big enough, $M \geq d(\tau^\star)$ and $(W_n)_n$ is a Markov chain. The intermediate quantity (see [20]) needed in the EM algorithm for the HMM (W_n, Y_n) can be written as:
for any (θ, θ') in Θ_{τ_M} :

$$\begin{aligned} Q_{\theta,\theta'} &= E_{\theta'}(\log(p_\theta(W_{1:n}, Y_{1:n}))\big|Y_{1:n}) \ , \\ &= E_{\theta'}(\nu(W_1)\big|Y_{1:n}) + \sum_{i=1}^{n-1} E_{\theta'}(\log P_{\theta_t}(W_i, W_{i+1})\big|Y_{1:n}) \\ &\quad + \sum_{i=1}^{n} E_{\theta'}(\log g_{\theta_{e,W_{i,M}},\eta}(Y_i)\big|Y_{1:n}) \ , \end{aligned}$$

where p_θ denotes the density, parametrized by θ, of $(W_{1:n}, Y_{1:n})$. Notice, for any $\theta \in \Theta_{\tau_M}$, if $(w, w') \in (\mathbb{X}^M)^2$ are such that $w_{2:M} \neq w'_{1:M-1}$, then $P_{\theta_t}(w, w') = 0$.

For any $w \in \mathbb{X}^M$ and any $w' \in \mathbb{X}^M$ if we denote by

$$\begin{aligned} &\forall\ i = 1, \ldots, n,\ \Phi^{\theta'}_{i|n}(w) = P_{\theta'}(W_i = w|Y_{1:n}) \ , \\ &\forall\ i = 1, \ldots, n-1, \\ &\qquad \Phi^{\theta'}_{i:i+1|n}(w, w') = P_{\theta'}(W_i = w, W_{i+1} = w'|Y_{1:n}) \ , \end{aligned}$$

and

$$S^{\theta'}_{t,n} = \left(\frac{\left(\sum\limits_{i=1}^{n-1} \Phi^{\theta'}_{i:i+1|n}(w, w') \right)}{n} \right)_{(w,w') \in \mathbb{X}^M} ,$$

$$S^{\theta'}_{e,n} = \frac{1}{n} \sum_{x \in \mathbb{X}} \sum_{i=1}^{n} \left(\sum_{w \in \mathbb{X}^M | w_M = x} \Phi^{\theta'}_{i|n}(w) \right) s(x, Y_i) \ ,$$

then there exists a function C such that :

$$\begin{aligned} \frac{1}{n} Q_{\theta,\theta'} = \frac{1}{n} C(\theta', Y_{1:n}) &+ \left\langle S^{\theta'}_{t,n}, \log P_{\theta_t} \right\rangle \\ &+ \left\langle S^{\theta'}_{e,n}, \psi(\theta_e) \right\rangle - A(\theta_e) \ . \quad (7) \end{aligned}$$

If, for some complete tree τ, we restrict θ_t in $\Theta_{t,\tau}$, then for any s in τ, for any w in $\mathbb{X}^M$ such that s is suffix of w, for any x in $\mathbb{X}$, we have $P_{\theta_t}(w, (w_{2:M} x)) = P_{s,x}(\theta_t)$.

Thus, the vector $P_{s,.}$ maximising this equation is solution of the Lagrangian,

$$\begin{cases} \dfrac{\delta}{\delta P_{s,x}} \left[\dfrac{1}{n} Q_{\theta,\theta'} + \Lambda(\sum\limits_{x' \in \mathbb{X}} P_{s,x'} - 1) \right] = 0 \ ,\ \forall x \in \mathbb{X} \\ \dfrac{\delta}{\delta \Lambda} \quad \left[\dfrac{1}{n} Q_{\theta,\theta'} + \Lambda(\sum\limits_{x' \in \mathbb{X}} P_{s,x'} - 1) \right] = 0 \ , \end{cases}$$

and, finally, the estimator of $\theta_t \in \Theta_{t,\tau}$ maximising the quantity $Q(\theta', .)$ only depends on the sufficient statistic $S^{\theta'}_{t,n}$ and is given by :

$$\bar{P}_{s,x}(S^{\theta'}_{t,n}) = \frac{\sum\limits_{w \in \mathbb{X}^M |\ s \text{ suffix } of\ w} S^{\theta'}_{t,n}(w, (w_{2:M} x))}{\sum\limits_{x' \in \mathbb{X}} \sum\limits_{w \in \mathbb{X}^M |\ s \text{ suffix } of\ w} S^{\theta'}_{t,n}(w, (w_{2:M} x'))} \ . \quad (8)$$

Algorithm 1 Preliminary computation of the sufficient statistics

```
Require: $\theta_0 = (\theta_{t,0}, \theta_{e,0}) \in \Theta_{\tau_M}$ be an initial value for the parameter $\theta$.
Require: Let $t_{EM}$ be a threshold.
$stop = 0$
$i = 0$
while ($stop = 0$) do
    $i = i + 1$
    M step : compute the quantities $S^{\theta_{i-1}}_{t,n}$ and $S^{\theta_{i-1}}_{e,n}$
    E step : set
        $\theta_i = \left( \left( \bar{P}_{w,x}(S^{\theta_{i-1}}_{t,n}) \right)_{w,x} \ , \ \bar{\theta}_e(S^{\theta_{i-1}}_{e,n}) \right)$
    if ($||\theta_i - \theta_{i-1}|| < t_{Em}$) then
        $stop = 1$
    end if
end while
M step : compute the quantities $S^{\theta_i}_{t,n}$ and $S^{\theta_i}_{e,n}$
$S_t = S^{\theta_i}_{t,n}$ and $S_e = S^{\theta_i}_{e,n}$
return $(S_t, S_e)$
```

While Algorithm 1 computes the sufficient statistics S_t and S_e on the basis of the observations $(Y_k)_{k \in \{1,\ldots,n\}}$, Algorithm 2 is our pruning Algorithm. This algorithm starts with the estimation of the exhaustive statistics calling Algorithm 1. As Algorithm 1 is prone to the convergence towards local maxima, in the Gaussian emissions cases, we set our initial parameter value θ_0 after running a preliminary *k-means* algorithm (see [21], [22]): we assign the values $Y_{1:n}$ into k clusters which produces a sequence of "clusters" $\tilde{X}_{1:n}$. A first estimation of the emission parameters is then possible using this clustering, the initial transition parameter $\theta_{0,t} = \left(P^0_{w,i}\right)_{w \in \mathbb{X}^M,\ i \in \mathbb{X}}$ is also computed on the basis of the sequence $\tilde{X}_{1:n}$ using the relation :

$$\forall w \in \mathbb{X}^M,\ \forall x \in \mathbb{X}, P^0_{w,x} = \frac{\sum\limits_{i=1}^{n-M} \mathbf{1}_{\tilde{X}_{i:i+M-1} = w} \mathbf{1}_{\tilde{X}_{i+M} = x}}{\sum\limits_{i=1}^{n-M} \mathbf{1}_{\tilde{X}_{i:i+M-1} = w}} \ .$$

Then, starting with the initialisation $\tau = \tau_M$, we consider, one after the other, the maximal nodes u of τ. We build a new tree τ_{test} by taking out of τ all the contexts s having u as suffix and adding u as a new context: $\tau_{test} = \tau \setminus \left\{ux \middle| ux \in \tau,\ x \in \mathbb{X}\right\} \bigcup \{u\}$. Let $\hat{\theta}_{\text{test}} = \left(((\bar{P}_{s,x}(S_t))_{s \in \tau_{\text{test}}, x \in \mathbb{X}}, \bar{\theta}_e(S_e) \right)$ which, hopefully, becomes an acceptable proxy for $\underset{\theta\in\Theta_{\tau_{test}}}{\operatorname{argmax}} \log \ell_\theta(Y_{1:n})$. Let $-\log \ell_{\hat{\theta}_{\text{test}}}(Y_{1:n}) + pen(n, \tau_{\text{test}})$ be an approximation of the score of the context tree τ_{test} still denoted by $sc(\tau_{\text{test}})$, then, if $sc(\tau_{\text{test}}) < sc(\tau)$, we set $\tau = \tau_{\text{test}}$. In Algorithm 2, the role of τ_2 is to insure that all the branches of τ are tested before shortening again a branch already tested.

Algorithm 2 Bottom to the top pruning algorithm

```
Require: Let $t_{EM}$ a threshold.
Compute $(S_t, S_e)$ with Algorithm 1 with the $t_{EM}$ threshold.
$\hat{\theta} = \left( \left( \bar{P}_{w,x}(S_t) \right)_{w\in\tau_M, x\in\mathbb{X}}, \bar{\theta}_e(S_e) \right)$
Pruning procedure :
$\tau = \tau_2 = \tau_M$
$change = YES$
while ($change = YES$ AND $|\tau| \geq 1$) do
  $change = NO$
  for ($u \in N(\tau)$) do
    if ($u \in N(\tau_2)$) then
      $L_u(\tau_2) = \{s \in \tau_2 | u\ suffix\ of\ s\}$
      $\tau_{\text{test}} = [\tau_2 \setminus L_u(\tau_2)] \bigcup \{u\}$
      $\hat{\theta}_{\text{test}} = \left( \left( \bar{P}_{s,x}(S_t) \right)_{s\in\tau, x\in\mathbb{X}}, \bar{\theta}_e(S_e) \right)$
      if ($sc(\tau_{\text{test}}) < sc(\tau_2)$) then
        $\tau_2 = \tau_{\text{test}}$
        $\hat{\theta} = \hat{\theta}_{\text{test}}$
        $change = YES$
      end if
    end if
  end for
  $\tau = \tau_2$
end while
return $\tau$
```

B. Simulations

We propose to illustrate the a.s convergence of $\hat{\tau}_n$ using Algorithm 2 in the case of Gaussian emission with unknown variance. We set $k = 2$, and use as minimal complete context tree one of the two complete trees represented in Figure 1 and Figure 2. The true transitions probabilities associated with each trees are indicated in boxes under each context.

For each tree $\tau_1^\star$ and $\tau_2^\star$, we will simulate 3 samples of the VLHMM, choosing as true emission parameters $m_0^\star = 0$, $\sigma^{2,\star} = 1$ and $m_1^\star$ varying in $\{2, 3, 4\}$. In the preliminary EM steps, we use as threshold $t_{EM} = 0.001$

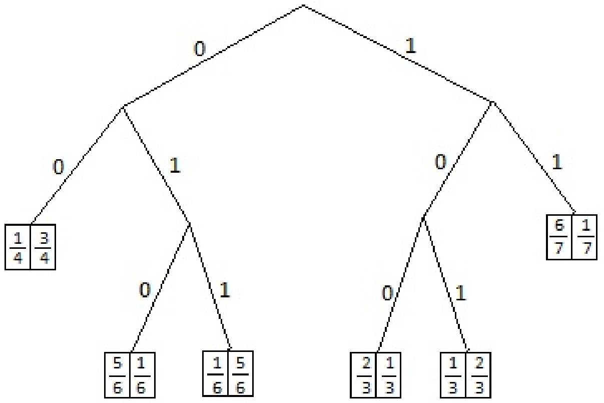

Figure 1: Graphic representation of the complete context tree $\tau_1^\star$ with transition probabilities indicated in the box under each leaf s: $P_{s,0}^\star \mid P_{s,1}^\star$

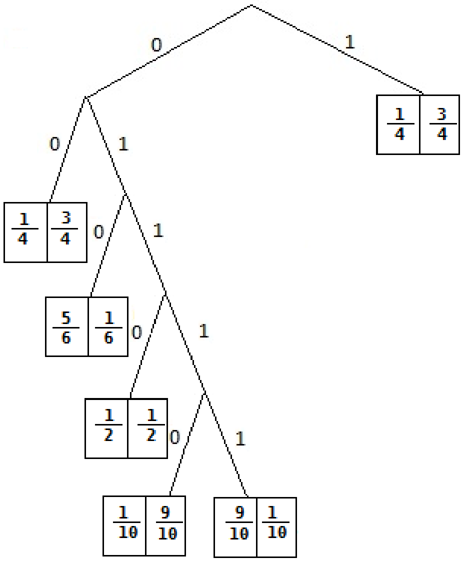

Figure 2: Graphic representation of the complete context tree $\tau_2^\star$ with transition probabilities indicated in the box under each leaf s: $P_{s,0}^\star \mid P_{s,1}^\star$

Algorithm 1 allows to compute the scores, needed in Algorithm 2, only using forward computations of the likelihood. However, the number of trees considered by Algorithm 2 is approximately $2^M - |\widehat{\tau}_n|$ and the computation of the score roughly takes 20 seconds per tested tree (in the case $n = 50000$ and using the R software). Thus, despite the amount of time saved thanks to the preliminary computation of the sufficient statistics, several hours were needed to fulfil the simulations presented in this section. The results of our simulations are summarized in Tables I to IV. The size of the tree $|\widetilde{\tau}_n|$, computed using Algorithm 2, for different values of n and $m_1^\star$ are noticed in Table I when $\tau^\star = \tau_1^\star$ (resp. in the table Figure III when $\tau^\star = \tau_2^\star$) for the two choices of penalties $pen_\alpha(n, \tau) = \sum_{t=1}^{|\tau|} \frac{(k-1)t + \alpha}{2} \log n$ with $\alpha = 5.1$ and $pen(n, \tau) = \frac{k-1}{2}|\tau| \log n$. The first important remark we make regarding Tables I and III is that, on each simulation

$\tau^\star = \tau_1^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (3)			BIC penalty		
$n/m_1^\star$	2	3	4	2	3	4
100	2	2	2	2	3	3
1000	2	2	2	7	**6**	**6**
2000	2	2	4	**6**	**6**	**6**
5000	2	4	4	7	**6**	**6**
10000	4	**6**	**6**	7	**6**	**6**
20000	5	**6**	**6**	**6**	**6**	**6**
30000	5	**6**	**6**	**6**	**6**	**6**
40000	**6**	**6**	**6**	7	**6**	**6**
50000	**6**	**6**	**6**	7	**6**	**6**

Table I: Case $\tau^\star = \tau_1^\star$. Comparison of $|\widetilde{\tau}_n|$ between our estimator and the BIC estimator for different values of n and $m_1^\star$.

$\tau^\star = \tau_1^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (3)			BIC penalty		
$n/m_1^\star$	2	3	4	2	3	4
100	-202	-202	-190	-6	-6	2
1000	-235	-213	-155	4	-2	25
2000	-221	-129	-88	8	-4	4
5000	-144	-36	-20	5	-4	-5
10000	-75	-5	-4	4	-5	-4
20000	-6	-4	-4	10	-4	-4
30000	21	-5	-4	10	-5	-4
40000	12	-4	-3	10	-4	-3
50000	12	-7	-4	10	-4	-4

Table II: Case $\tau^\star = \tau_1^\star$. Score difference $sc(\widetilde{\tau}_n) - sc(\tau^\star)$.

$\tau^\star = \tau_2^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (3)			BIC penalty		
$n/m_1^\star$	2	3	4	2	3	4
100	2	2	2	2	2	2
1000	2	2	2	3	**6**	**6**
2000	2	2	2	**6**	**6**	**6**
5000	2	3	3	**6**	**6**	**6**
10000	3	3	3	**6**	**6**	**6**
20000	3	3	**6**	**6**	**6**	**6**
30000	3	3	**6**	**6**	**6**	**6**
40000	3	**6**	**6**	**6**	**6**	**6**
50000	3	**6**	**6**	**6**	**6**	**6**

Table III: Case $\tau^\star = \tau_2^\star$. Comparison of $|\widetilde{\tau}_n|$ between our estimator and the BIC estimator for different values of n and $m_1^\star$.

$\tau^\star = \tau_2^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (3)			BIC penalty		
n/m_1	2	3	4	2	3	4
100	-201	-202	-195	-10	-6	1
1000	-266	-246	-229	5	-1	-2
2000	-272	-239	67	4	-1	324
5000	-272	-200	-151	2	-2	-5
10000	-242	-128	-52	6	-2	-4
20000	-227	12	-6	6	-6	-6
30000	-191	141	-6	7	-5	-6
40000	-159	-6	-8	8	-6	-8
50000	-136	-6	-9	7	-6	-8

Table IV: Case $\tau^\star = \tau_2^\star$. Score difference $sc(\widetilde{\tau}_n) - sc(\tau^\star)$.

and whatever the penalty we used, when $|\widetilde{\tau}_n| = |\tau^\star|$ we also had $\widetilde{\tau}_n = \tau^\star$, in the same way, each time $|\widetilde{\tau}_n| < |\tau^\star|$ (resp. $|\widetilde{\tau}_n| > |\tau^\star|$), $\widetilde{\tau}_n$ was a subtree of $\tau^\star$ (resp. $\tau^\star$ was a subtree of $\widetilde{\tau}_n$). For any combination of $\tau^\star$ and $m_1^\star$, both estimators seem to converge, except our estimator in the case $\tau^\star = \tau_2^\star$ and $m_1^\star = 2$, where 50 000 measures do not seem sufficient to reach the convergence. However, for small samples, smaller models are systematically chosen with our estimator, while the BIC estimator is reaching the right model for relatively small samples. This behaviour of our estimator shows that our penalty is too heavy.

The score differences $sc(\widetilde{\tau}_n) - sc(\tau^\star)$ Table II when $\tau^\star = \tau_1^\star$ and Table IV when $\tau^\star = \tau_2^\star$ are the differences between the score of $\widetilde{\tau}_n$ computed with the estimated parameter $\hat{\theta}_n$ and the score of $\tau^\star$ computed with the the real parameters. These informations allow us to know when the estimators $\widetilde{\tau}_n, \hat{\theta}_n$ are well estimated by Algorithm 2. Indeed, when $\widetilde{\tau}_n \neq \tau^\star$, if the score of $\tau^\star$ computed with the real transition and emission parameters is smaller than the score of our estimator with estimated parameters (non negative score difference), then the tree, $\widetilde{\tau}_n$, given by Algorithm 2 might not be the expected estimator $\widehat{\tau}_n$ defined by (2). In particular, Table II shows that the over estimation of the BIC estimator in the case $m_1^\star = 2$ (Table II) can be due to a local minima problem: Algorithm 2 selected a tree $\widetilde{\tau}_n$ such that $|\widetilde{\tau}_n| > |\tau^\star|$ whereas $\tau^\star$ had a smaller score. This problem might occur because we use an EM type algorithm which often leads to local minima. Although we try to take an initial value of the parameters in a neighbourhood of the real ones using the preliminary k-means algorithm, this problem persists. Extra EM loops for each tested tree in Algorithm 2 could also provide a better estimation of the parameters and then improve the score estimation for each tested tree, but it would also increase the complexity of the algorithm. An other interpretation of these incorrect estimations is the suboptimality of our pruning procedure. Indeed, Algorithm 2 only explores a restrictive set of trees which does not necessarily contain $\widehat{\tau}_n$.

Finally, we observe that the bigger the quantity $|m_0^\star - m_1^\star|$ is, the quicker the convergence of our estimator or BIC estimator occurs. This phenomenon can be easily understood as very different emission distributions for different states leads to an easier estimation of the underlying state sequence on the basis of the observations and allows us to build a more precise description of the VLMC behaviour.

VI. Conclusion

In this paper, we were interested in the statistical analysis of Variable Length Hidden Markov Models (VLHMM). We have presented such models then we estimated the context tree of the hidden process using penalized maximum likelihood. We have shown how to choose the penalty so that the estimator is strongly consistent without any prior upper bound on the depth or on the size of the context tree of the hidden process. We have proved that our general consistency theorem applies when the emission distributions are Gaussian with unknown means and the same unknown variance. We have proposed a pruning algorithm, probably suboptimal, and have applied it to simulated data sets. This illustrates the consistency of our estimator, but also suggests that smaller penalty could lead to consistent estimation.

Finding the minimal penalty insuring the strong consistency of the estimator with no prior upper bound remains unsolved. A similar problem has been solved by R. van Handel [8] to estimate the order of finite state Markov chains, and by E. Gassiat and R. van Handel [9] to estimate the number of populations in a mixture with i.i.d. observations. The basic idea is that the maximum likelihood behaves as the maximum of approximate chi-square variables, and that the behavior of the maximum likelihood statistic may be investigated using empirical process theory tools to obtain a $\log\log n$ rate of growth. However, it is known for HMM that the maximum likelihood does not behave this way and converges weakly to infinity, see [10]. We bypassed the problem by using information theoretic inequalities, but understanding the path-wise fluctuations of the likelihood in HMM models remains a difficult problem to be solved.

Appendix A
Proof of Lemma 2

For any partition $(I_i,\dots,I_k)$ of $\mathbb{R}$ in k intervals,

$$\sigma^2_{I_i,\dots,I_k} = \sum_{j=1}^{k} \mathbb{P}_{\theta^\star}(Y_1 \in I_k) Var_{\theta^\star}(Y_1 | Y_1 \in I_k) \ ,$$
$$\geq \frac{1}{k} \inf_{I : \mathbb{P}_{\theta^\star}(Y \in I) \geq \frac{1}{k}} Var_{\theta^\star}(Y_1 | Y_1 \in I) \ ,$$

where the infimum is over all intervals I of $\mathbb{R}$. The distribution of Y_1 is the Gaussian mixture with density $g^\star = \sum_{x \in \mathbb{X}} \pi^\star(x) \phi_{m_x^\star, \sigma_\star^2}$, where $\pi^\star$ is the stationary distribution of $(X_n)_{n \geq 0}$ and $\phi_{m_x^\star, \sigma_\star^2}$ is the density of the normal distribution with mean $m_x^\star$ and variance $\sigma_\star^2$. The repartition function $F^\star$ of the distribution of Y_1 is continuous and increasing, with continuous and increasing inverse quantile function. Thus,

$$\inf_{I_i,\dots,I_k} \sigma^2_{I_i,\dots,I_k} \geq \inf_{\substack{-\infty \leq a < b \leq +\infty: \\ F^\star(a) + \frac{1}{k} \leq F^\star(b)}} Var_{\theta^\star}(Y_1 | Y_1 \in]a,b[) \ .$$

But $Var_{\theta^\star}(Y_1 | Y_1 \in]a,b[)$ is a continuous function of (a,b), and the infimum at the righ-hand side of the inequality is attained at some $(\overline{a},\overline{b})$ (eventually infinite) such that $F^\star(\overline{a}) + \frac{1}{k} \leq F^\star(\overline{b})$. Thus $Var_{\theta^\star}(Y_1 | Y_1 \in]\overline{a},\overline{b}[) > 0$, and $s_{inf} > 0$.

For any partition $(I_i,\dots,I_k)$ of $\mathbb{R}$ in k intervals,

$$\hat{\sigma}^2_{I_i,\dots,I_k}(Y_{1:n}) - \sigma^2_{I_i,\dots,I_k} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - E(Y_1^2)$$
$$-\sum_{j=1}^{k} \left(\frac{(\sum_{i=1}^{n} Y_i \mathbf{1}_{I_j}(Y_i))^2}{n^2} \frac{n}{\sum_{i=1}^{n} \mathbf{1}_{I_j}(Y_i)} - \frac{E(Y \mathbf{1}_{I_j}(Y_1))^2}{E(\mathbf{1}_{I_j}(Y_1))} \right) \ ,$$

so that

$$\sup_{I_1,\dots,I_k} \left| \hat{\sigma}^2_{I_i,\dots,I_k}(Y_{1:n}) - \sigma^2_{I_i,\dots,I_k} \right| \leq \frac{1}{n} \left| \sum_{i=1}^{n} Y_i^2 - E(Y_1^2) \right|$$
$$+ k \sup_{I \text{ interval of } \mathbb{R}} \left| \frac{(\sum_{i=1}^{n} Y_i \mathbf{1}_I(Y_i))^2}{n^2} \frac{n}{\sum_{i=1}^{n} \mathbf{1}_I(Y_i)} - \frac{E(Y_1 \mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))} \right| \ .$$

Using [16], $(Y_n)_{n \geq 0}$ is a stationary ergodic process, so that $\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - E(Y_1^2)$ tends to 0 $\mathbb{P}_{\theta^\star}$ a.s. Let $\epsilon > 0$. We now consider separately the intervals I such that $E(\mathbf{1}_I(Y)) \leq \epsilon$ or $E(\mathbf{1}_I(Y)) > \epsilon$.

- Let I be such that $E(\mathbf{1}_I(Y_1)) \leq \epsilon$.

Using Cauchy Schwarz inequality,

$$\left(\frac{1}{n} \sum Y_i \mathbf{1}_I(Y_i) \right)^2 \leq \left(\frac{1}{n} \sum Y_i^2 \mathbf{1}_I(Y_i) \right) \times \left(\frac{1}{n} \sum \mathbf{1}_I(Y_i) \right) ,$$
$$E\left(Y_1 \mathbf{1}_I(Y_1)\right)^2 \leq E\left(Y_1^2 \mathbf{1}_I(Y_1)\right) E\left(\mathbf{1}_I(Y_1)\right) \ ,$$

and,

$$E\left(Y_1^2 \mathbf{1}_I(Y_1)\right) \leq \sqrt{E(Y_1^4)} \sqrt{E(\mathbf{1}_I(Y_1))} \leq M\sqrt{\epsilon} \ ,$$

for some fixed positive constant M. Thus,

$$\left| \frac{(\sum_{i=1}^{n} Y_i \mathbf{1}_I(Y_i))^2}{n^2} \frac{n}{\sum_{i=1}^{n} \mathbf{1}_I(Y_i)} - \frac{E(Y_1 \mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))} \right|$$
$$\leq \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \mathbf{1}_I(Y_i) + E(Y_1^2 \mathbf{1}(Y_1)) \ ,$$
$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \mathbf{1}_I(Y_i) - E(Y_1^2 \mathbf{1}(Y_1)) \right| + 2E(Y_1^2 \mathbf{1}(Y_1)) \ ,$$
$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \mathbf{1}_I(Y_i) - E(Y_1^2 \mathbf{1}(Y_1)) \right| + 2M\sqrt{\epsilon} \ .$$

- Let now I be such that $E(\mathbf{1}_I(Y_1)) > \epsilon$.

$$
\begin{aligned}
&\left|\frac{(\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i))^2}{n^2}\frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)} - \frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))}\right| \\
&= \left|\frac{\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i)}{n}\frac{1}{\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}}} - \frac{E(Y_1\mathbf{1}_I(Y_1))}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right| \\
&\times \left|\frac{\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i)}{n}\frac{1}{\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}}} + \frac{E(Y_1\mathbf{1}_I(Y_1))}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right| , \\
&\leq \left[\left|\frac{\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i)}{n}\right|\left|\frac{1}{\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}}} - \frac{1}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right|\right. \\
&+ \left.\left|\frac{\frac{\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i)}{n} - E(Y_1\mathbf{1}_I(Y_1))}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right|\right] \\
&\times \left[\left|\frac{\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i)}{n}\right|\left|\frac{1}{\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}}} + \frac{1}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right|\right. \\
&+ \left.\left|\frac{\frac{\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i)}{n} - E(Y\mathbf{1}_I(Y_1))}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right|\right] , \\
&\leq \left[\left(\frac{\sum_{i=1}^n |Y_i|}{n}\right)\frac{\left|\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}} - \sqrt{E(\mathbf{1}_I(Y_1))}\right|}{\epsilon}\right. \\
&+ \left.\frac{\left|\frac{\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i)}{n} - E(Y_1\mathbf{1}_I(Y_1))\right|}{\sqrt{\epsilon}}\right] \\
&\times \left[\frac{2}{\epsilon}\frac{\sum_{i=1}^n |Y_i|}{n} + \frac{\left|\frac{\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i)}{n} - E(Y_1\mathbf{1}_I(Y_1))\right|}{\sqrt{\epsilon}}\right] .
\end{aligned}
$$

Now, using Lemma 4 below, one gets that, for all positive ϵ,

$$
\begin{aligned}
\limsup_{n\to\infty} \sup_{I \text{ interval of } \mathbb{R}} &\left|\frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))}\right. \\
&\left. - \frac{(\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i))^2}{n^2}\frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)}\right| \leq 2M\sqrt{\epsilon} ,
\end{aligned}
$$

$\mathbb{P}_{\theta^\star}$-a.s. so that

$$
\begin{aligned}
\lim_{n\to\infty} \sup_{I \text{ interval of } \mathbb{R}} &\left|\frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))}\right. \\
&\left. - \frac{(\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i))^2}{n^2}\frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)}\right| = 0 ,
\end{aligned}
$$

$\mathbb{P}_{\theta^\star}$-a.s. and the Lemma follows.

Lemma 4. *The following quantities,*

$$
\begin{aligned}
&\sup_I \left|\frac{1}{n}\sum Y_i^2\mathbf{1}_I(Y_i) - E\left(Y_1^2\mathbf{1}_I(Y)\right)\right| , \\
&\sup_I \left|\frac{1}{n}\sum Y_i\mathbf{1}_I(Y_i) - E\left(Y_1\mathbf{1}_I(Y_1)\right)\right| , \\
&\sup_I \left|\frac{1}{n}\sum \mathbf{1}_I(Y_i) - E\left(\mathbf{1}_I(Y_1)\right)\right| ,
\end{aligned}
$$

where the supremums are over all intervals I in $\mathbb{R}$, tend to 0 *as n tends to infinity, $\mathbb{P}_{\theta^\star}$ a.s.*

Proof: Let us note $\mathcal{F}_a = \{x \to x^a\mathbf{1}_I(x) \ : \ I \text{ interval of } \mathbb{R}\}$ for $a = 0, 1, 2$. Since the sequence of random variables $(Y_n)_{n\geq 0}$ is stationary and ergodic, it is enough to prove that, for $a = 0, 1, 2$, for any positive ϵ, there exists a finite set of functions $\tilde{\mathcal{F}}_a$ such that for any $f \in \mathcal{F}_a$, there exists l, u in $\tilde{\mathcal{F}}_a$ such that $l \leq f \leq u$ and $E(u(Y_1) - l(Y_1)) \leq \epsilon$.
For the cases a=0 or 2 and for any positive ϵ, there exist real numbers : $L^1_{a,\epsilon}$ and $L^2_{a,\epsilon}$ such that $\int_{-\infty}^{L^1_{a,\epsilon}} x^a g^\star(x)dx \leq \epsilon$ and $\int_{L^2_{a,\epsilon}}^{+\infty} x^a g^\star(x)dx \leq \epsilon$, and there exists real numbers $x_{a,1} = L^1_\epsilon < x_{a,2} < .... < x_{a,N_{a,\epsilon}-2} < L^2_{a,\epsilon} = x_{a,N_{a,\epsilon}-1}$ such that $\int_{x_{a,i}}^{x_{a,i+1}} x^a g^\star(x)dx < \epsilon/2$, $i = 1, \ldots, N_{a,\epsilon} - 2$. Then we define

- $I^1_{N_\epsilon} = \mathbb{R}$,
- for any $i = 1, ..., N_{a,\epsilon}$, $I^1_{a,i} = [-\infty \ , \ x_{a,i}]$
- and for any $i = 1, ..., N_{a,\epsilon}$, $I^2_{a,i} = [x_{a,i} \ , \ \infty]$

so that if $\mathcal{I}_a$ is the set $\mathcal{I}_a = \left\{I^j_{a,i} | i = 1, ..., N_{a,\epsilon}, \ j = 1, 2\right\} \bigcup \{[x_{a,i_1}, x_{a,i_2}]\}_{i_1<i_2}$ the set $\tilde{\mathcal{F}}_a = \{x^a\mathbf{1}_I | I \in \mathcal{I}_a\}$ verifies the above conditions.
For the case $a = 1$ the construction of the sequence $x_{a,1} = L^1_\epsilon < x_{a,2} < .... < x_{N_{a,\epsilon}-2} < L^2_{a,\epsilon} = x_{N_{a,\epsilon}-1}$ is such that $\int_{x_i}^{x_{i+1}} |x|g^\star(x)dx < \epsilon/2$ is similar except that we introduce 0 in the sequence : $x_{1:N_{a,\epsilon}}$. ■

Appendix B
Proof of Lemma 3

Let $t_n = 5\sigma_\star^2 \log n$. One has

$$
\begin{aligned}
&\mathbb{P}_{\theta^\star}\left(|Y|^2_{(n)} \geq t_n\right) \\
&\leq \max_{x_{1:n}\in\mathbb{X}^n} \mathbb{P}_{\theta^\star}\left(|Y|^2_{(n)} \geq t_n | X_{1:n} = x_{1:n}\right) , \\
&= \max_{x_{1:n}\in\mathbb{X}^n} \left\{1 - \prod_{i=1}^n \mathbb{P}_{\theta^\star}\left(Y_i^2 \leq t_n | X_i = x_i\right)\right\} , \\
&\leq 1 - \left[\mathbb{P}\left(U^2 \leq \frac{t_n - M}{\sigma_\star}\right)\right]^n ,
\end{aligned}
$$

where $M = \max_{i=1,\ldots,k} m_i^\star$ and U is a Gaussian random variable with distribution $\mathcal{N}(0, 1)$. Then, for large enough n :

$$
\mathbb{P}_{\theta^\star}\left(|Y|^2_{(n)} \geq t_n\right) \leq \frac{1}{n^{3/2}}
$$

and the result follows from Borel Cantelli Lemma.

References

[1] L. Finesso, “Consistent estimation of the order for Markov and hidden Markov chains,” 1990, ph.D. Thesis, Univ. of Maryland.

[2] E. Gassiat and S. Boucheron, “Optimal error exponents in hidden Markov model order estimation,” *IEEE Trans. Info. Theory*, vol. 48, pp. 964–980, 2003.

[3] J. Rissanen, “A universal data compression system,” *IEEE Trans. Inform. Theory*, vol. 29, pp. 656 – 664, 1983.

[4] Y. Wang, L. Zhou, J. Wang, and Z. Liu, “Mining complex time-series by learning markovian models,” in *Proceedings ICDM’06, sixth international conference on data mining*, China, 2005.

[5] Y. Wang, “The variable-length hidden markov model and its applications on sequential data mining,” Departement of computer science, Tech. Rep., 2005.

[6] P. Collet, A. Galves, and F. Leonardi, “Random perturbations of stochastic processes with unbounded variable length memory,” *Electron. J. Probab.*, vol. 13, pp. 1345–1361, 2008.

[7] F. Evennou, “Techniques et technologies de localisation avancées pour terminaux mobiles dans les environnements indoor,” 2007, ph.D. Thesis, Univ. Joseph Fourier, Grenoble, France.

[8] R. van Handel, “On the minimal penalty for Markov order estimation,” *Probability Theory and Related Fields*, vol. 150, pp. 709–738, 2011, 10.1007/s00440-010-0290-y.

[9] E. Gassiat and R. van Handel, “Consistent order estimation and minimal penalties,” *Information Theory, IEEE Transactions on*, vol. 59, no. 2, pp. 1115–1128, Feb 2013.

[10] E. Gassiat and C. Keribin, “The likelihood ratio test for the number of components in a mixture with Markov regime,” *ESAIM: Probability and Statistics*, vol. 4, pp. 25–52, 1 2000.

[11] A. Chambaz, A. Garivier, and E. Gassiat, “A MDL approach to HMM with Poisson and Gaussian emissions. Application to order indentification,” *Journal of Stat. Planning and Inf.*, vol. 139, pp. 962–977, 2009.

[12] R. E. Krichevsky and V. K. Trofimov, “The performance of universal encoding,” *IEEE Trans. Inform. Theory*, vol. 27, pp. 199–207, 1981.

[13] O. Catoni, *Statistical learning theory and stochastic optimization*, ser. Ecole d’Eté de Probabilités de Saint-Flour, XXXI-2001. Berlin: Springer, 2004, vol. 1851.

[14] E. Gassiat, “Codage universel et sélection de modèles emboités,” 2011, lectures notes, M2 Orsay.

[15] A. Garivier, “Consistency of the unlimited bic context tree estimator,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 4630–4635, 2006.

[16] B. Leroux, “Maximum-likelihood estimator for hidden markov models,” *Stoch. Proc. Appl.*, vol. 40, pp. 127–143, 1992.

[17] S. Yakowitz and J. Spragins, “On the identifiability of finite mixtures,” *Ann. Math. Stat.*, vol. 39, no. 1, pp. 209–214, 1968.

[18] H. Teicher, “Identifiability of mixtures,” *Ann. Math. Stat.*, vol. 32, no. 1, pp. 244–248, 1961.

[19] E. Whittaker and G. Watson, *A Course of Modern Analysis*, ser. Cambridge Mathematical Library. Cambridge University Press, 1927.

[20] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*, ser. Springer Series in Statistics. New York: Springer, 2005.

[21] J. McQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the 5th Berkeley Sympos. Math. Statist. and Probability*. Berkeley, California: University of California Press, 1967, pp. 281–297.

[22] M. Inaba, N. Katoh, and H. Imai, “Applications of weighted Voronoi diagrams and randomization to variance based k-clustering,” in *Proceedings of the tenth annual symposium computational geometry*, 1994, pp. 332–339.