Localizing the Latent Structure Canonical Uncertainty: Entropy Profiles for Hidden Markov Models

Jean-Baptiste Durand, Yann Guédon

Localizing the Latent Structure Canonical Uncertainty: Entropy Profiles for Hidden Markov Models

Jean-Baptiste Durand, Yann Guédon

RESEARCH REPORT

N° 7896

February 2012

Project-Teams Mistis and Virtual Plants

Localizing the Latent Structure Canonical Uncertainty: Entropy Profiles for Hidden Markov Models

Jean-Baptiste Durand*, Yann Guédon†

Project-Teams Mistis and Virtual Plants

Research Report n° 7896 — February 2012 — 40 pages

Abstract: This report addresses state inference for hidden Markov models. These models rely on unobserved states, which often have a meaningful interpretation. This makes it necessary to develop diagnostic tools for quantification of state uncertainty. The entropy of the state sequence that explains an observed sequence for a given hidden Markov chain model can be considered as the canonical measure of state sequence uncertainty. This canonical measure of state sequence uncertainty is not reflected by the classic multivariate state profiles computed by the smoothing algorithm, which summarizes the possible state sequences. Here, we introduce a new type of profiles which have the following properties: (i) these profiles of conditional entropies are a decomposition of the canonical measure of state sequence uncertainty along the sequence and makes it possible to localize this uncertainty, (ii) these profiles are univariate and thus remain easily interpretable on tree structures. We show how to extend the smoothing algorithms for hidden Markov chain and tree models to compute these entropy profiles efficiently.

Key-words: Conditional Entropy, Hidden Markov Chain Model, Hidden Markov Tree Model, Plant Structure Analysis

* jean-baptiste.durand@imag.fr, Laboratoire Jean Kuntzmann and INRIA, Mistis, 51 rue des Mathématiques B.P. 53, F-38041 Grenoble cedex 9, France

† guedon@cirad.fr, CIRAD, UMR AGAP and INRIA, Virtual Plants F-34398 Montpellier, France

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Localisation de l'incertitude canonique de structures latentes : profils d'entropie pour des modèles de Markov cachés.

Résumé : Ce rapport concerne l'inférence sur les états de modèles de Markov cachés. Ces modèles se fondent sur des états non observés, qui ont en général une interprétation, dans le contexte d'une application donnée. Ceci rend nécessaire la conception d'outils de diagnostic pour quantifier l'incertitude sur ces états. L'entropie de la séquence d'états associée à une séquence observée, pour un modèle de chaîne de Markov cachée donné, peut être considérée comme la mesure canonique de l'incertitude sur les états. Cette mesure canonique d'incertitude sur la séquence d'états n'est pas reflétée par les profils d'états, multivariés, calculés par l'algorithme de lissage, qui résume les séquences d'états possibles. Nous introduisons ici de nouveaux profils dont les propriétés sont les suivantes : (i) ces profils d'entropie conditionnelle sont une décomposition, le long de cette séquence, de la mesure canonique d'incertitude sur la séquence d'états, ce qui offre la possibilité d'une localisation de cette incertitude, (ii) ces profils sont univariés; ils peuvent donc être facilement utilisés sur des structures arborescentes. Nous montrons comment étendre l'algorithme de lissage sur des chaînes et arbres de Markov cachés afin de calculer ces profils de manière efficace.

Mots-clés : Entropie conditionnelle, modèles de chaînes de Markov cachées, modèles d'arbres de Markov cachés, analyse de l'architecture des plantes

1 Introduction

Hidden Markov chain models have been widely used in signal processing and pattern recognition, for the analysis of sequences with various types of underlying structures – for example succession of homogeneous zones, or noisy patterns (Ephraim & Mehrav, 2002; Zucchini & MacDonald, 2009). This family of models was extended to other kinds of structured data, and particularly to tree graphs (Crouse *et al.*, 1998). Concerning statistical inference for hidden Markov models, we distinguish inference for the unobserved state process from inference for model parameters (Cappé *et al.*, 2005). Our focus here is state inference and more precisely the uncertainty in state sequences.

State inference is particularly relevant in numerous applications where the unobserved states have a meaningful interpretation. In such cases, the state sequence has to be restored. The restored states may be used, typically, in prediction, in segmentation or in denoising. For example Ciriza *et al.* (2011) proposed to optimize the consumption of printers by prediction of the future printing rate from the sequence of printing requests. This rate is related to the parameters of a hidden Markov chain model, and an optimal timeout (time before entering sleep mode) is derived from the restored states. Le Cadre & Tremois (1998) used a vector of restored states in a dynamical system for source tracking in sonar and radar systems. Such use of the state sequence makes assessment of the state uncertainty particularly important.

Not only is state restoration essential for model interpretation, it is generally also used for model diagnostic and validation, for example based on the visualization of functions of the states. The use of restored states in the above-mentioned contexts raises the issue of quantifying the state sequence uncertainty for a given observed sequence, once a hidden Markov model has been estimated. Global quantification of this uncertainty is not sufficient for a precise diagnosis: it is also very important to locate this uncertainty along the sequence, for instance to differentiate zones that are non-ambiguously explained from zones that are ambiguously explained by the estimated model. We have introduced the statistical problem of quantifying state uncertainty in the case of hidden Markov models with discrete state space for sequences, but the same reasoning applies to other families of latent structure models, including hidden semi-Markov models and hidden Markov tree models.

Methods for exploring the state sequences that explain a given observed sequence for a known hidden Markov chain model may be divided into three categories: (i) enumeration of state sequences, (ii) state profiles, which are state sequences summarized in a $J \times T$ array where J is the number of states and T the length of the sequence, (iii) computation of a global measure of state sequence uncertainty. The entropy of the state sequence that explains an observed sequence for a known hidden Markov chain model was proposed as a global measure of the state sequence uncertainty by Hernando *et al.* (2005). We assume here that this conditional entropy is the canonical measure of state sequence uncertainty. Various methods belonging to these three categories have been developed for different families of hidden Markovian models, including hidden Markov chain and hidden semi-Markov chain models; see Guédon (2007) and references therein. We identified some shortcomings of the proposed methods:

- The entropy of the state sequence is not a direct summary of the state profiles based on the smoothed probabilities, due to the marginalization that is intrinsic in the computation of smoothed probabilities. We show that the uncertainty reflected in the classic multivariate state profiles computed by the smoothing algorithm can be summarized as an univariate profile of marginal entropies. Each successive marginal entropy quantifies the uncertainty in the corresponding posterior state distribution for a given time t. The entropy of the state sequence, in contrast, can be decomposed along the sequence as a profile of conditional entropies where the conditioning refers to the preceding states. Using results from

information theory, we show that the profile of conditional entropies is pointwise upper-bounded by the profile of marginal entropies. Hence, the classic state profiles tend to over-represent the state sequence uncertainty and should be interpreted with caution.

- Due to their multidimensional nature, state profiles are difficult to visualize and interpret on trees except in the case of two-state models.

Our objective is to propose efficient algorithms for computing univariate profiles of conditional entropies. These profiles correspond to an additive decomposition of the entropy of the state process along the sequence. As a consequence, each term of the decomposition can be interpreted as a local contribution to entropy. This principle can be extended to more general supporting structures: directed acyclic graphs (DAGs), and in particular, trees. Each contribution is shown to be the conditional entropy of the state at each location, given the past or the future of the state process. This decomposition allows canonical uncertainty to be localized within the structure, which makes the connection between global and local uncertainty easily apprehensible, even for hidden Markov tree models. In this case, we propose to compute in a first stage an univariate profile of conditional entropies that summarizes state uncertainty for each vertex. In a second stage, the usual state profiles computed by the upward-downward algorithm (Durand *et al.*, 2004), or an adaption to trees of the forward-backward Viterbi algorithm of Brushe *et al.* (1998), are visualized on selected paths of interest within the tree. This allows for identification of alternative states at positions with ambiguous state value.

The remainder of this paper is structured as follows. Section 2 focuses on algorithms to compute entropy profiles for state sequences in hidden Markov chain models. These algorithms are based on conditioning on either the past or the future of the process. In Section 3, an additive decomposition of the global state entropy is derived for graphical hidden Markov models indexed by DAGs. Then algorithms to compute entropy profiles conditioned on the parent states and conditioned on the children states are derived in detail in the case of hidden Markov tree models. The use of entropy profiles is illustrated in Section 4 through applications to sequence and tree data. Section 5 consists of concluding remarks.

2 Entropy profiles for hidden Markov chain models

In this section, definitions and notations related to hidden Markov chain (HMC) models are introduced. These are followed by reminders on the classic forward-backward algorithm and the algorithm of Hernando *et al.* (2005) to compute the entropy of the state sequence. These algorithms form the basis of the proposed methodology to compute the state sequence entropy, as the sum of local conditional entropies.

2.1 Definition of a hidden Markov chain model

A J-state HMC model can be viewed as a pair of discrete-time stochastic processes $(\boldsymbol{S}, \boldsymbol{X}) = (S_t, X_t)_{t=0,1,\ldots}$ where $\boldsymbol{S}$ is an unobserved Markov chain with finite state space $\{0, \ldots, J-1\}$ and parameters:

- $\pi_j = P(S_0 = j)$ with $\sum_j \pi_j = 1$ (initial probabilities), and
- $p_{ij} = P(S_t = j | S_{t-1} = i)$ with $\sum_j p_{ij} = 1$ (transition probabilities).

The output process $\boldsymbol{X}$ is related to the state process $\boldsymbol{S}$ by the emission (or observation) probabilities

$$b_j(x) = P(X_t = x | S_t = j) \text{ with } \sum_x b_j(x) = 1.$$

Since the emission distributions $(b_j)_{j=0,\ldots,J-1}$ are such that a given output x may be observed in different states, the state process $\boldsymbol{S}$ cannot be deduced without uncertainty from the outputs, but is observable only indirectly through output process $\boldsymbol{X}$. To simplify the algorithm presentation, we consider a discrete univariate output process. Since this work focuses on conditional distributions of states given the outputs, this assumption is not restrictive.

In the sequel, $X_0^t = x_0^t$ is a shorthand for $X_0 = x_0, \ldots, X_t = x_t$ (this convention transposes to the state sequence $S_0^t = s_0^t$). For a sequence of length T, $X_0^{T-1} = x_0^{T-1}$ is simply noted $\boldsymbol{X} = \boldsymbol{x}$. In the derivation of the algorithms for computing entropy profiles, we will use repeatedly the fact that if $(S_t)_{t=0,1,\ldots}$ is a first-order Markov chain, the time-reversed process is also a first-order Markov chain.

2.2 Reminders: forward-backward algorithm and algorithm of Hernando *et al.* (2005)

The forward-backward algorithm aims at computing the smoothed probabilities $L_t(j) = P(S_t = j|\boldsymbol{X} = \boldsymbol{x})$ and can be stated as follows (Devijver, 1985). The forward recursion is initialized at $t = 0$ and for $j = 0, \ldots, J-1$ as follows:

$$\begin{aligned} F_0(j) &= P(S_0 = j|X_0 = x_0) \\ &= \frac{b_j(x_0)}{N_0}\pi_j. \end{aligned} \tag{1}$$

The recursion is achieved, for $t = 1, \ldots, T-1$ and for $j = 0, \ldots, J-1$, using:

$$\begin{aligned} F_t(j) &= P\left(S_t = j|X_0^t = x_0^t\right) \\ &= \frac{b_j(x_t)}{N_t}\sum_i p_{ij}F_{t-1}(i). \end{aligned} \tag{2}$$

The normalizing factor N_t is obtained directly during the forward recursion as follows

$$\begin{aligned} N_t &= P\left(X_t = x_t|X_0^{t-1} = x_0^{t-1}\right) \\ &= \sum_j P\left(S_t = j, X_t = x_t|X_0^{t-1} = x_0^{t-1}\right), \end{aligned}$$

with

$$P(S_0 = j, X_0 = x_0) = b_j(x_0)\pi_j,$$

and

$$P\left(S_t = j, X_t = x_t|X_0^{t-1} = x_0^{t-1}\right) = b_j(x_t)\sum_i p_{ij}F_{t-1}(i).$$

The backward recursion is initialized at $t = T-1$ and for $j = 0, \ldots, J-1$ as follows:

$$L_{T-1}(j) = P(S_{T-1} = j|\boldsymbol{X} = \boldsymbol{x}) = F_{T-1}(j). \tag{3}$$

The recursion is achieved, for $t = T-2, \ldots, 0$ and for $j = 0, \ldots, J-1$, using:

$$\begin{aligned} L_t(j) &= P(S_t = j|\boldsymbol{X} = \boldsymbol{x}) \\ &= \left\{\sum_k \frac{L_{t+1}(k)}{G_{t+1}(k)}p_{jk}\right\}F_t(j), \end{aligned} \tag{4}$$

where

$$\begin{aligned} G_{t+1}(k) &= P\left(S_{t+1} = k | X_0^t = x_0^t\right) \\ &= \sum_j p_{jk} F_t(j) . \end{aligned}$$

These recursions rely on conditional independence properties between hidden and observed variables in HMC models. Several recursions given in Section 2 rely on the following relations, due to the time-reversed process of $(S_t, X_t)_{t=0,1,\ldots}$ being also a hidden first-order Markov chain: for $t = 1, \ldots, T-1$ and for $i, j = 0, \ldots, J-1$,

$$\begin{aligned} P(S_{t-1} = i | S_t = j, \boldsymbol{X} = \boldsymbol{x}) &= P\left(S_{t-1} = i | S_t = j, X_0^t = x_0^t\right) \\ &= P\left(S_{t-1} = i | S_t = j, X_0^{t-1} = x_0^{t-1}\right) , \\ P\left(S_0^{t-1} = s_0^{t-1} | S_t = j, \boldsymbol{X} = \boldsymbol{x}\right) &= P\left(S_0^{t-1} = s_0^{t-1} | S_t = j, X_0^t = x_0^t\right) \\ &= P\left(S_0^{t-1} = s_0^{t-1} | S_t = j, X_0^{t-1} = x_0^{t-1}\right) . \end{aligned}$$

An algorithm was proposed by Hernando *et al.* (2005) for computing the entropy of the state sequence that explains an observed sequence in the case of an HMC model. This algorithm includes the classic forward recursion given by (1) and (2) as a building block. It requires a forward recursion on entropies of partial state sequences S_0^t. (In the sequel, it is understood that the entropy of hidden state variables refers to their conditional entropies given observed values.) This algorithm is initialized at $t = 1$ and for $j = 0, \ldots, J-1$ as follows:

$$\begin{aligned} &H\left(S_0 | S_1 = j, X_0^1 = x_0^1\right) \\ &\quad = -\sum_i P\left(S_0 = i | S_1 = j, X_0^1 = x_0^1\right) \log P\left(S_0 = i | S_1 = j, X_0^1 = x_0^1\right) . \end{aligned} \tag{5}$$

The recursion is achieved, for $t = 2, \ldots, T-1$ and for $j = 0, \ldots, J-1$, using:

$$\begin{aligned} &H\left(S_0^{t-1} | S_t = j, X_0^t = x_0^t\right) \\ &= -\sum_{s_0, \ldots, s_{t-1}} P\left(S_0^{t-1} = s_0^{t-1} | S_t = j, X_0^t = x_0^t\right) \log P\left(S_0^{t-1} = s_0^{t-1} | S_t = j, X_0^t = x_0^t\right) \\ &= -\sum_{s_0, \ldots, s_{t-2}} \sum_i P\left(S_0^{t-2} = s_0^{t-2} | S_{t-1} = i, S_t = j, X_0^t = x_0^t\right) P\left(S_{t-1} = i | S_t = j, X_0^t = x_0^t\right) \\ &\times \left\{\log P\left(S_0^{t-2} = s_0^{t-2} | S_{t-1} = i, S_t = j, X_0^t = x_0^t\right) + \log P\left(S_{t-1} = i | S_t = j, X_0^t = x_0^t\right)\right\} \\ &= -\sum_i P\left(S_{t-1} = i | S_t = j, X_0^{t-1} = x_0^{t-1}\right) \left\{ \sum_{s_0, \ldots, s_{t-2}} P\left(S_0^{t-2} = s_0^{t-2} | S_{t-1} = i, X_0^{t-1} = x_0^{t-1}\right) \right. \\ &\left. \times \log P\left(S_0^{t-2} = s_0^{t-2} | S_{t-1} = i, X_0^{t-1} = x_0^{t-1}\right) + \log P\left(S_{t-1} = i | S_t = j, X_0^t = x_0^t\right) \right\} \\ &= \sum_i P\left(S_{t-1} = i | S_t = j, X_0^{t-1} = x_0^{t-1}\right) \left\{H\left(S_0^{t-2} | S_{t-1} = i, X_0^{t-1} = x_0^{t-1}\right)\right. \\ &\quad \left. - \log P\left(S_{t-1} = i | S_t = j, X_0^{t-1} = x_0^{t-1}\right)\right\} , \end{aligned} \tag{6}$$

with

$$\begin{aligned}
&P\left(S_{t-1}=i|S_t=j,X_0^t=x_0^t\right)\\
&=\frac{P\left(S_t=j,S_{t-1}=i|X_0^{t-1}=x_0^{t-1}\right)}{P\left(S_t=j|X_0^{t-1}=x_0^{t-1}\right)}\\
&=\frac{p_{ij}F_{t-1}\left(i\right)}{G_t\left(j\right)}.
\end{aligned}$$

The forward recursion (6) is a direct consequence of the conditional independence properties within a HMC model and can be interpreted as the chain rule

$$\begin{aligned}
&H\left(S_0^{t-1}|S_t=j,X_0^t=x_0^t\right)\\
&=H\left(S_0^{t-2}|S_{t-1},S_t=j,X_0^t=x_0^t\right)+H\left(S_{t-1}|S_t=j,X_0^t=x_0^t\right)
\end{aligned}\tag{7}$$

with

$$\begin{aligned}
&H\left(S_0^{t-2}|S_{t-1},S_t=j,X_0^t=x_0^t\right)\\
&=-\sum_{s_0,\ldots,s_{t-1}}P\left(S_0^{t-1}=s_0^{t-1}|S_t=j,X_0^t=x_0^t\right)\times\log P\left(S_0^{t-2}=s_0^{t-2}|S_{t-1}=s_{t-1},S_t=j,X_0^t=x_0^t\right)\\
&=-\sum_i P\left(S_{t-1}=i|S_t=j,X_0^t=x_0^t\right)\sum_{s_0,\ldots,s_{t-2}}P\left(S_0^{t-2}=s_0^{t-2}|S_{t-1}=i,X_0^{t-1}=x_0^{t-1}\right)\\
&\quad\times\log P\left(S_0^{t-2}=s_0^{t-2}|S_{t-1}=i,X_0^{t-1}=x_0^{t-1}\right)\\
&=\sum_i P\left(S_{t-1}=i|S_t=j,X_0^{t-1}=x_0^{t-1}\right)H(S_0^{t-2}|S_{t-1}=i,X_0^{t-1}=x_0^{t-1})
\end{aligned}$$

and

$$\begin{aligned}
&H\left(S_{t-1}|S_t=j,X_0^t=x_0^t\right)\\
&=-\sum_i P\left(S_{t-1}=i|S_t=j,X_0^{t-1}=x_0^{t-1}\right)\log P\left(S_{t-1}=i|S_t=j,X_0^{t-1}=x_0^{t-1}\right).
\end{aligned}$$

Using a similar argument as in (6), the termination step is given by

$$\begin{aligned}
&H\left(S_0^{T-1}|\boldsymbol{X}=\boldsymbol{x}\right)\\
&=-\sum_j P\left(S_{T-1}=j|\boldsymbol{X}=\boldsymbol{x}\right)\left\{\sum_{s_0,\ldots,s_{T-2}}P\left(S_0^{T-2}=s_0^{T-2}|S_{T-1}=j,\boldsymbol{X}=\boldsymbol{x}\right)\right.\\
&\left.\times\ \log P\left(S_0^{T-2}=s_0^{T-2}|S_{T-1}=j,\boldsymbol{X}=\boldsymbol{x}\right)+\log P\left(S_{T-1}=j|\boldsymbol{X}=\boldsymbol{x}\right)\right\}\\
&=\sum_j F_{T-1}\left(j\right)\left\{H\left(S_0^{T-2}|S_{T-1}=j,\boldsymbol{X}=\boldsymbol{x}\right)-\log F_{T-1}\left(j\right)\right\}.
\end{aligned}\tag{8}$$

The forward recursion, the backward recursion and the algorithm of Hernando *et al.* (2005) all have complexity in $\mathcal{O}(J^2T)$.

2.3 Entropy profiles for hidden Markov chain models

In what follows, we derive algorithms to compute entropy profiles based on conditional and partial entropies. Firstly, conditioning with respect to past states is considered. Then, conditioning with respect to future states is considered.

The proposed algorithms have a twofold aim, since they focus in computing both

- profiles of partial state sequence entropies $(H(S_0^t|\boldsymbol{X}=\boldsymbol{x}))_{t=0,\ldots,T-1}$
- profiles of conditional entropies $(H(S_t|S_{t-1},\boldsymbol{X}=\boldsymbol{x}))_{t=0,\ldots,T-1}$.

We propose a first solution where the partial state sequence entropies are computed beforehand, and the conditional entropies are deduced from the latter. Then, we propose an alternative solution where the conditional entropies are computed directly, and the partial state sequence entropies are extracted from these.

The profiles of conditional entropies have the noteworthy property that the global state sequence entropy can be decomposed as a sum of entropies conditioned on the past states:

$$H\left(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}\right)=H\left(S_0|\boldsymbol{X}=\boldsymbol{x}\right)+\sum_{t=1}^{T-1}H\left(S_t|S_{t-1},\boldsymbol{X}=\boldsymbol{x}\right). \tag{9}$$

This property comes from the fact that the state sequence $\boldsymbol{S}$ is conditionally a Markov chain given $\boldsymbol{X}=\boldsymbol{x}$.

In this way, the state sequence uncertainty can be localized along the observed sequence, $H(S_t|S_{t-1},\boldsymbol{X}=\boldsymbol{x})$ representing the local contribution at time t to the state sequence entropy. For $t=0,\ldots,T-1$, using conditional independence properties within HMC models, we have

$$\begin{aligned}
&H\left(S_0^t|\boldsymbol{X}=\boldsymbol{x}\right)\\
&=-\sum_{s_0,\ldots,s_t}P\left(S_0^t=s_0^t|\boldsymbol{X}=\boldsymbol{x}\right)\log P\left(S_0^t=s_0^t|\boldsymbol{X}=\boldsymbol{x}\right)\\
&=-\sum_j P\left(S_t=j|\boldsymbol{X}=\boldsymbol{x}\right)\left\{\sum_{s_0,\ldots,s_{t-1}}P\left(S_0^{t-1}=s_0^{t-1}|S_t=j,X_0^t=x_0^t\right)\right.\\
&\quad\left.\times\log P\left(S_0^{t-1}=s_0^{t-1}|S_t=j,X_0^t=x_0^t\right)+\log P\left(S_t=j|\boldsymbol{X}=\boldsymbol{x}\right)\right\}\\
&=\sum_j L_t\left(j\right)\left\{H\left(S_0^{t-1}|S_t=j,X_0^t=x_0^t\right)-\log L_t\left(j\right)\right\} \qquad (10)\\
&=\sum_j L_t\left(j\right)H\left(S_0^{t-1}|S_t=j,X_0^t=x_0^t\right)+H\left(S_t|\boldsymbol{X}=\boldsymbol{x}\right).
\end{aligned}$$

Using a similar argument as in (7), equation (10) can be interpreted as the chain rule

$$H\left(S_0^t|\boldsymbol{X}=\boldsymbol{x}\right)=H\left(S_0^{t-1}|S_t,\boldsymbol{X}=\boldsymbol{x}\right)+H\left(S_t|\boldsymbol{X}=\boldsymbol{x}\right)$$

In this way, the profile of partial state sequence entropies $(H(S_0^t|\boldsymbol{X}=\boldsymbol{x}))_{t=0,\ldots,T-1}$ can be computed as a byproduct of the forward-backward algorithm where the usual forward recursion (2) and the recursion (6) proposed by Hernando *et al.* (2005) are mixed. The conditional entropies are then directly deduced by first-order differencing

$$\begin{aligned}
H\left(S_t|S_{t-1},\boldsymbol{X}=\boldsymbol{x}\right)&=H\left(S_t|S_0^{t-1},\boldsymbol{X}=\boldsymbol{x}\right)\\
&=H\left(S_0^t|\boldsymbol{X}=\boldsymbol{x}\right)-H\left(S_0^{t-1}|\boldsymbol{X}=\boldsymbol{x}\right). \qquad (11)
\end{aligned}$$

As an alternative, the profile of conditional entropies $(H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}))_{t=0,\ldots,T-1}$ could also be computed directly, as

$$\begin{aligned} & H\left(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}\right) \\ & = -\sum_{i,j} P\left(S_t = j, S_{t-1} = i|\boldsymbol{X} = \boldsymbol{x}\right) \log P\left(S_t = j|S_{t-1} = i, \boldsymbol{X} = \boldsymbol{x}\right) \end{aligned} \tag{12}$$

with

$$\left\{ \begin{array}{ll} P\left(S_t = j|S_{t-1} = i, \boldsymbol{X} = \boldsymbol{x}\right) & = L_t\left(j\right) p_{ij} F_{t-1}\left(i\right) / \left\{G_t\left(j\right) L_{t-1}\left(i\right)\right\} \text{ and} \\ P\left(S_t = j, S_{t-1} = i|\boldsymbol{X} = \boldsymbol{x}\right) & = L_t\left(j\right) p_{ij} F_{t-1}\left(i\right) / G_t\left(j\right). \end{array} \right. \tag{13}$$

These latter quantities are directly extracted during the backward recursion (4) of the forward-backward algorithm.

In summary, a first possibility is to compute the profile of partial state sequence entropies $(H(S_0^t|\boldsymbol{X} = \boldsymbol{x}))_{t=0,\ldots,T-1}$ using the usual forward and backward recursions combined with (5), (6) and (10), from which the profile of conditional entropies $(H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}))_{t=0,\ldots,T-1}$ is directly deduced by first-order differencing (11). A second possibility is to compute the profile of conditional entropies directly using the usual forward and backward recursions combined with (12) and to deduce the profile of partial state sequence entropies by summation. The time complexity of both algorithms is in $\mathcal{O}(J^2T)$.
The conditional entropy is bounded from above by the marginal entropy (Cover &Thomas, 2006, chap. 2):

$$H\left(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}\right) \leq H\left(S_t|\boldsymbol{X} = \boldsymbol{x}\right),$$

with

$$\begin{aligned} H\left(S_t|\boldsymbol{X} = \boldsymbol{x}\right) & = -\sum_j P\left(S_t = j|\boldsymbol{X} = \boldsymbol{x}\right) \log P\left(S_t = j|\boldsymbol{X} = \boldsymbol{x}\right) \\ & = -\sum_j L_t\left(j\right) \log L_t\left(j\right). \end{aligned}$$

and the difference between the marginal and the conditional entropy is the mutual information between S_t and S_{t-1}, given $\boldsymbol{X} = \boldsymbol{x}$. Thus, the marginal entropy profile $(H(S_t|\boldsymbol{X} = \boldsymbol{x}))_{t=0,\ldots,T-1}$ can be viewed as pointwise upper bounds on the conditional entropy profile $(H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}))_{t=0,\ldots,T-1}$. The profile of marginal entropies can be interpreted as a summary of the classic state profiles given by the smoothed probabilities $(P(S_t = j|\boldsymbol{X} = \boldsymbol{x}))_{t=0,\ldots,T-1;j=0,\ldots,J-1}$. Hence, the difference between the marginal entropy $H(S_t|\boldsymbol{X} = \boldsymbol{x})$ and the conditional entropy $H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x})$ can be seen as a defect of the classic state profiles, which provide a representation of the state sequences such that global uncertainty is overestimated.

Entropy profiles conditioned on the future for hidden Markov chain models The Markov property, which states that the past and the future are independent given the present, essentially treats the past and the future symmetrically. However, there is a lack of symmetry in the parameterization of a Markov chain, with the consequence that only the state process conditioned on the past is often investigated. However, the state uncertainty at time t may be better explained by the values of future states than past states. Consequently, in the present context of state inference, we chose to investigate the state process both forward and backward in time.

Entropy profiles conditioned on the future states rely on the following decomposition of the entropy of the state sequence, as a sum of local entropies where state S_t at time t is conditioned

on the future states:

$$H\left(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}\right)=\sum_{t=0}^{T-2}H\left(S_t|S_{t+1},\boldsymbol{X}=\boldsymbol{x}\right)+H\left(S_{T-1}|\boldsymbol{X}=\boldsymbol{x}\right).$$

This is a consequence of the reverse state process being a Markov chain, given $\boldsymbol{X}=\boldsymbol{x}$.

An algorithm to compute the backward conditional entropies $H(S_{t+1}^{T-1}|S_t=j,X_{t+1}^{T-1}=x_{t+1}^{T-1})$ can be proposed. This algorithm, detailed in Appendix A.1, is similar to that of Hernando *et al.* (2005) but relies on a backward recursion. Using similar arguments as in (10), we have

$$H\left(S_t^{T-1}|\boldsymbol{X}=\boldsymbol{x}\right)=\sum_j L_t\left(j\right)\left\{H\left(S_{t+1}^{T-1}|S_t=j,X_{t+1}^{T-1}=x_{t+1}^{T-1}\right)-\log L_t(j)\right\}. \tag{14}$$

Thus, the profile of partial state sequence entropies $\left(H(S_t^{T-1}|\boldsymbol{X}=\boldsymbol{x})\right)_{t=0,\ldots,T-1}$ can be computed as a byproduct of the forward-backward algorithm, where the usual backward recursion (4) and the backward recursion for conditional entropies (see Appendix A.1) are mixed. The conditional entropies are then directly deduced by first-order differencing

$$\begin{aligned}H\left(S_t|S_{t+1},\boldsymbol{X}=\boldsymbol{x}\right)&=H\left(S_t|S_{t+1}^{T-1},\boldsymbol{X}=\boldsymbol{x}\right)\\&=H\left(S_t^{T-1}|\boldsymbol{X}=\boldsymbol{x}\right)-H\left(S_{t+1}^{T-1}|\boldsymbol{X}=\boldsymbol{x}\right).\end{aligned}$$

The profile of conditional entropies $(H(S_t|S_{t+1},\boldsymbol{X}=\boldsymbol{x}))_{t=0,\ldots,T-1}$ can also be computed directly, as

$$H\left(S_t|S_{t+1},\boldsymbol{X}=\boldsymbol{x}\right)=-\sum_{j,k}P\left(S_t=j,S_{t+1}=k|\boldsymbol{X}=\boldsymbol{x}\right)\log P\left(S_t=j|S_{t+1}=k,\boldsymbol{X}=\boldsymbol{x}\right)$$

with

$$\begin{aligned}P\left(S_t=j|S_{t+1}=k,\boldsymbol{X}=\boldsymbol{x}\right)&=P\left(S_t=j|S_{t+1}=k,X_0^t=x_0^t\right)\\&=p_{jk}F_t\left(j\right)/G_{t+1}\left(k\right)\quad\text{and}\\P\left(S_t=j,S_{t+1}=k|\boldsymbol{X}=\boldsymbol{x}\right)&=L_{t+1}\left(k\right)p_{jk}F_t\left(j\right)/G_{t+1}\left(k\right).\end{aligned}$$

The latter quantities are directly extracted during the forward (2) and backward recursions (4) of the forward-backward algorithm. The conditional entropy is bounded from above by the marginal entropy (Cover & Thomas (2006), chap. 2):

$$H\left(S_t|S_{t+1},\boldsymbol{X}=\boldsymbol{x}\right)\leq H\left(S_t|\boldsymbol{X}=\boldsymbol{x}\right).$$

3 Entropy profiles for hidden Markov tree models

In this section, hidden Markov tree (HMT) models are introduced, as a particular case of graphical hidden Markov (GHM) models. A generic additive decomposition of state entropy in GHM models is proposed, and its implementation is discussed in the case of HMT models.

3.1 Graphical hidden Markov models

Let $\mathcal{G}$ be a directed acyclic graph (DAG) with vertex set $\mathcal{U}$, and $\boldsymbol{S}=(S_u)_{u\in\mathcal{U}}$ be a J-state process indexed by $\mathcal{U}$. Let $\mathcal{G}(\boldsymbol{S})$ be the graph with vertices $\boldsymbol{S}$, isomorphic to $\mathcal{G}$ (so that the set of vertices of $\mathcal{G}(\boldsymbol{S})$ may be assimilated with $\mathcal{U}$). It is assumed that $\boldsymbol{S}$ satisfies the graphical

Markov property with respect to $\mathcal{G}(\boldsymbol{S})$, in the sense defined by Lauritzen (1996). The states S_u are observed indirectly through an output process $\boldsymbol{X} = (X_u)_{u\in\mathcal{U}}$ such that given $\boldsymbol{S}$, the $(X_u)_{u\in\mathcal{U}}$ are independent, and for any u, X_u is independent of $(S_v)_{v\in\mathcal{U};v\neq u}$ given S_u. Then process $\boldsymbol{X}$ is referred to as a GHM model with respect to DAG $\mathcal{G}$.

Let $\mathrm{pa}(u)$ denote the set of parents of $u \in \mathcal{U}$. For any subset E of $\mathcal{U}$, let $\boldsymbol{S}_E$ denote $(S_u)_{u\in E}$. As a consequence from the Markov property of $\boldsymbol{S}$, the following factorization of $P_{\boldsymbol{S}}$ holds for any $\boldsymbol{s}$ (Lauritzen, 1996):

$$P(\boldsymbol{S} = \boldsymbol{s}) = \prod_u P(S_u = s_u|\boldsymbol{S}_{\mathrm{pa}(u)} = \boldsymbol{s}_{\mathrm{pa}(u)}),$$

where $P(S_u = s_u|\boldsymbol{S}_{\mathrm{pa}(u)} = \boldsymbol{s}_{\mathrm{pa}(u)})$ must be understood as $P(S_u = s_u)$ if $\mathrm{pa}(u) = \emptyset$. This factorization property is shown by induction on the vertices in $\mathcal{U}$, starting from the sink vertices (vertices without children), and ending at the source vertices (vertices without parents).

In the particular case where $\mathcal{G}$ is a rooted tree graph, $\boldsymbol{X}$ is called a hidden Markov out-tree with conditionally-independent children states, given their parent state (or more shortly, a hidden Markov tree model). This model was introduced by Crouse *et al.* (1998) in the context of signal and image processing using wavelet trees. The state process $\boldsymbol{S}$ is called a Markov tree.

The following notations will be used for a tree graph $\mathcal{T}$: for any vertex u, $c(u)$ denotes the set of children of u and $\rho(u)$ denotes its parent. Let $\mathcal{T}_u$ denote the complete subtree rooted at vertex u, $\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u$ denote the observed complete subtree rooted at u, $\bar{\boldsymbol{X}}_{c(u)} = \bar{\boldsymbol{x}}_{c(u)}$ denote the collection of observed subtrees rooted at children of vertex u (that is, subtree $\bar{\boldsymbol{x}}_u$ except its root x_u), $\bar{\boldsymbol{X}}_{u\setminus v} = \bar{\boldsymbol{x}}_{u\setminus v}$ the subtree $\bar{\boldsymbol{x}}_u$ except the subtree $\bar{\boldsymbol{x}}_v$ (assuming that $\bar{\boldsymbol{x}}_v$ is a proper subtree of $\bar{\boldsymbol{x}}_u$), and finally $\bar{\boldsymbol{X}}_{b(u)} = \bar{\boldsymbol{x}}_{b(u)}$ the family of brother subtrees $(\bar{\boldsymbol{X}}_v)_{v\in\rho(u);v\neq u}$ of u (assuming that u is not the root vertex). This notation transposes to the state process with for instance $\bar{\boldsymbol{S}}_u = \bar{\boldsymbol{s}}_u$, the state subtree rooted at vertex u. In the sequel, we will use the notation $\mathcal{U} = \{0, \ldots, n-1\}$ to denote the vertex set of a tree with size n, and the root vertex will be $u = 0$. Thus, the entire observed tree can be denoted by $\bar{\boldsymbol{X}}_0 = \bar{\boldsymbol{x}}_0$, although the shorter notation $\boldsymbol{X} = \boldsymbol{x}$ will be used hereafter. These notations are illustrated in Figure 1.

A J-state HMT model $(\boldsymbol{S}, \boldsymbol{X}) = (S_u, X_u)_{u\in\mathcal{U}}$ is defined by the following parameters:

- initial probabilities (for the root vertex) $\pi_j = P(S_0 = j)$ with $\sum_j \pi_j = 1$,
- transition probabilities $p_{jk} = P\left(S_u = k|S_{\rho(u)} = j\right)$ with $\sum_k p_{jk} = 1$,

and by the emission distributions defined as in HMC models by $P(X_u = x|S_u = j) = b_j(x)$.

In GHM models, the state process is conditionally Markovian in the following sense:

Proposition 1 *Let $(\boldsymbol{S}, \boldsymbol{X})$ be a GHM model with respect to DAG $\mathcal{G}$. Then for any $\boldsymbol{x}$, the conditional distribution of $\boldsymbol{S}$ given $\boldsymbol{X} = \boldsymbol{x}$ satisfies the Markov property on $\mathcal{G}$ and for any $\boldsymbol{s}$,*

$$P(\boldsymbol{S} = \boldsymbol{s}|\boldsymbol{X} = \boldsymbol{x}) = \prod_u P(S_u = s_u|\boldsymbol{S}_{pa(u)} = \boldsymbol{s}_{pa(u)}, \boldsymbol{X} = \boldsymbol{x}),$$

where $P(S_u = s_u|\boldsymbol{S}_{pa(u)} = \boldsymbol{s}_{pa(u)}, \boldsymbol{X} = \boldsymbol{x})$ denotes $P(S_u = s_u|\boldsymbol{X} = \boldsymbol{x})$ if $pa(u) = \emptyset$.

Proof To prove this proposition, we consider a potential realization $(\boldsymbol{s}, \boldsymbol{x})$ of process $(\boldsymbol{S}, \boldsymbol{X})$. We introduce the following definitions and notations: for $u \in \mathcal{U}$, $An(u)$ denotes the set of ancestors of u in $\mathcal{G}$; for $A \subset U$, $An(A) = \{An(u)\}_{u\in A}$ and $\bar{An}(A) = An(A) \cup A$. Let $\boldsymbol{S}_A = \boldsymbol{s}_A$ denote the state process indexed by the graph induced by A. By conditional independence of the $(X_u)_{u\in\mathcal{U}}$ given $\boldsymbol{S}$, the process $(\boldsymbol{S}, \boldsymbol{X})$ follows the Markov property on the DAG $\mathcal{G}(\boldsymbol{S}, \boldsymbol{X})$ obtained from $\mathcal{G}(\boldsymbol{S})$ by addition of the set of vertices $\{X_u|u \in \mathcal{U}\}$ and the set of arcs $\{(S_u, X_u)|u \in \mathcal{U}\}$.

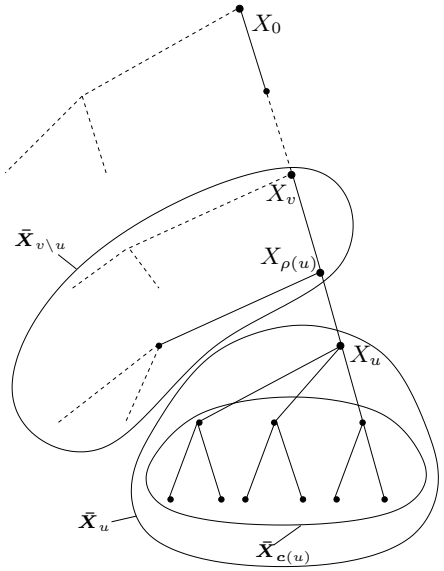


Figure 1: *The notations used for indexing trees*

It is proved by induction on subgraphs A of $\mathcal{G}$ that if $\bar{A}n(A) = A$, then

$$P(\boldsymbol{S}_A = \boldsymbol{s}_A|\boldsymbol{X} = \boldsymbol{x}) = \prod_{v\in A} P(S_v = s_v|\boldsymbol{S}_{\mathrm{pa}(v)} = \boldsymbol{s}_{\mathrm{pa}(v)}, \boldsymbol{X} = \boldsymbol{x}). \tag{15}$$

Since the joint distribution of state vertices in different connected components $(\mathcal{G}_1, \ldots, \mathcal{G}_C)$ of $\mathcal{G}$ can be factorized as $\prod_c P(\boldsymbol{S}_{\mathcal{G}_c} = \boldsymbol{s}_{\mathcal{G}_c}|\boldsymbol{X} = \boldsymbol{x})$, equation (15) is proved separately for each connected component.

It is easily seen that if u is a source of $\mathcal{G}$, both the right-hand and the left-hand sides of equation (15) are equal to $P(S_u = s_u|\boldsymbol{X} = \boldsymbol{x})$. To prove the induction step, we consider a vertex $u \notin A$ such that $\mathrm{pa}(u) \subset A$. If such vertex does not exist, A is a connected component of $\mathcal{G}$, which terminates the induction.

Otherwise, let A' denote $A \cup \{u\}$. Then $\bar{A}n(A') = A'$ and

$$\begin{aligned} P(\boldsymbol{S}_{A'} = \boldsymbol{s}_{A'}|\boldsymbol{X} = \boldsymbol{x}) =& P(S_u = s_u|\boldsymbol{S}_{\mathrm{pa}(u)} = \boldsymbol{s}_{\mathrm{pa}(u)}, \boldsymbol{S}_{A\setminus\mathrm{pa}(u)} = \boldsymbol{s}_{A\setminus\mathrm{pa}(u)}, \boldsymbol{X} = \boldsymbol{x}) \\ & \times P(\boldsymbol{S}_{\mathrm{pa}(u)} = \boldsymbol{s}_{\mathrm{pa}(u)}, \boldsymbol{S}_{A\setminus\mathrm{pa}(u)} = \boldsymbol{s}_{A\setminus\mathrm{pa}(u)}|\boldsymbol{X} = \boldsymbol{x}) \\ =& P(S_u = s_u|\boldsymbol{S}_{\mathrm{pa}(u)} = \boldsymbol{s}_{\mathrm{pa}(u)}, \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{S}_A = \boldsymbol{s}_A|\boldsymbol{X} = \boldsymbol{x}) \end{aligned}$$

since the Markov property on $\mathcal{G}(\boldsymbol{S}, \boldsymbol{X})$ implies conditional independence of S_u and $\boldsymbol{S}_{A\setminus\mathrm{pa}(u)}$ given $S_{\mathrm{pa}(u)}$ and $\boldsymbol{X}$.

The proof is completed by application of induction equation (15). ■

From application of the chain rule (Cover & Thomas, 2006, chap. 2) to Proposition 1, the following corollary is derived:

Corollary 1 *Let $(\boldsymbol{S}, \boldsymbol{X})$ be a GHM model with respect to DAG $\mathcal{G}$. Then for any $\boldsymbol{x}$,*

$$H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x}) = \sum_u H(S_u|\boldsymbol{S}_{pa(u)}, \boldsymbol{X} = \boldsymbol{x}),$$

where $H(S_u|\boldsymbol{S}_{pa(u)}, \boldsymbol{X} = \boldsymbol{x})$ *denotes* $H(S_u|\boldsymbol{X} = \boldsymbol{x})$ *if* $pa(u) = \emptyset$.

This result extends equation (9) for HMC models to hidden Markov models indexed by DAGs.

It follows from Corollary 1 that the global entropy of the state process can be decomposed as a sum of conditional entropies, where each term is the local contribution of state S_u at vertex u, and corresponds to the conditional entropy of this state given the parent state (or equivalently, given the non-descendant states, from the Markov property on $\mathcal{G}(\boldsymbol{S}, \boldsymbol{X})$).

The remainder of this Section focuses on the derivation of algorithms to compute $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ efficiently in HMT models.

3.2 Reminder: upward-downward algorithm

The upward-downward algorithm aims at computing the smoothed probabilities $\xi_u(j) = P(S_u = j|\boldsymbol{X} = \boldsymbol{x})$ and can be stated as follows (Durand *et al.*, 2004). It consists in three recursions, which all have complexities in $\mathcal{O}(J^2 n)$.

This algorithm requires preliminary computation of the state marginal probabilities $P(S_u = j)$, computed by a downward recursion. This recursion is initialized at the root vertex $u = 0$ and for $j = 0, \ldots, J-1$ as follows:

$$P(S_0 = j) = \pi_j.$$

The recursion is achieved, for vertices $u \neq 0$ taken downwards and for $j = 0, \ldots, J-1$, using:

$$P(S_u = j) = \sum_i p_{ij} P(S_{\rho(u)} = i).$$

The upward recursion is initialized for each leaf as follows. For $j = 0, \ldots, J-1$,

$$\begin{aligned}\beta_u(j) =& P(S_u = j|X_u = x_u)\\ =& \frac{b_j(x_u)P(S_u = j)}{N_u}.\end{aligned}$$

The recursion is achieved, for internal vertices u taken upwards and for $j = 0, \ldots, J-1$, using:

$$\begin{aligned}\beta_{\rho(u),u}(j) =& \frac{P(\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u|S_{\rho(u)} = j)}{P(\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)}\\ =& \sum_k \frac{\beta_u(k)p_{jk}}{P(S_u = k)}\end{aligned}$$

and

$$\begin{aligned}\beta_u(j) =& P(S_u = j|\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)\\ =& \frac{\left\{\prod_{v\in\boldsymbol{c}(u)} \beta_{u,v}(j)\right\} b_j(x_u)P(S_u = j)}{N_u}.\end{aligned}$$

The normalizing factor N_u is obtained directly during the upward recursion by

$$N_u = P(X_u = x_u) = \sum_j b_j(x_u)P(S_u = j)$$

for the leaf vertices, and

$$N_u = \frac{P(\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)}{\prod\limits_{v \in \boldsymbol{c}(u)} P(\bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v)} = \sum_j \left\{ \prod_{v \in \boldsymbol{c}(u)} \beta_{u,v}(j) \right\} b_j(x_u) P(S_u = j)$$

for the internal vertices.
The downward recursion is initialized at the root vertex $u = 0$ and for $j = 0, \ldots, J-1$ as follows:

$$\xi_0(j) = P(S_0 = j | \boldsymbol{X} = \boldsymbol{x}) = \beta_0(j)$$

The recursion is achieved, for vertices $u \neq 0$ taken downwards and for $j = 0, \ldots, J-1$, using:

$$\begin{aligned} \xi_u(j) =& P(S_u = j | \boldsymbol{X} = \boldsymbol{x}) \\ =& \frac{\beta_u(j)}{P(S_u = j)} \sum_i \frac{p_{ij} \xi_{\rho(u)}(i)}{\beta_{\rho(u),u}(i)}. \end{aligned} \tag{16}$$

These recursions rely on conditional independence properties between hidden and observed variables in HMT models. In several recursions given in Section 3, the following relations will be used: for any internal, non-root vertex u and for $j = 1, \ldots, J$,

$$\begin{aligned} P(\bar{\boldsymbol{S}}_{c(u)} = \bar{\boldsymbol{s}}_{c(u)} | S_u = j, \bar{\boldsymbol{S}}_{0\setminus u} = \bar{\boldsymbol{s}}_{0\setminus u}, \boldsymbol{X} = \boldsymbol{x}) & \\ = & P(\bar{\boldsymbol{S}}_{c(u)} = \bar{\boldsymbol{s}}_{c(u)} | S_u = j, S_{\rho(u)} = s_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}) \\ = & P(\bar{\boldsymbol{S}}_{c(u)} = \bar{\boldsymbol{s}}_{c(u)} | S_u = j, \boldsymbol{X} = \boldsymbol{x}) \\ = & \prod_{v \in c(u)} P(\bar{\boldsymbol{S}}_v = \bar{\boldsymbol{s}}_v | S_u = j, \boldsymbol{X} = \boldsymbol{x}) \\ = & \prod_{v \in c(u)} P(\bar{\boldsymbol{S}}_v = \bar{\boldsymbol{s}}_v | S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v), \end{aligned}$$

$$\begin{aligned} P(\bar{\boldsymbol{S}}_u = \bar{\boldsymbol{s}}_u | \bar{\boldsymbol{S}}_{0\setminus u} = \bar{\boldsymbol{s}}_{0\setminus u}, \boldsymbol{X} = \boldsymbol{x}) &= P(\bar{\boldsymbol{S}}_u = \bar{\boldsymbol{s}}_u | S_{\rho(u)} = s_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}) \\ &= P(\bar{\boldsymbol{S}}_u = \bar{\boldsymbol{s}}_u | S_{\rho(u)} = s_{\rho(u)}, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u). \end{aligned}$$

3.3 Algorithms for computing entropy profiles for hidden Markov tree models

In HMT models, the generic decomposition of global state tree entropy yielded by Corollary 1 writes

$$H(\boldsymbol{S} | \boldsymbol{X} = \boldsymbol{x}) = H(S_0 | \boldsymbol{X} = \boldsymbol{x}) + \sum_{u \neq 0} H(S_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}).$$

As in the case of HMC models, such decomposition of $H(\boldsymbol{S} | \boldsymbol{X} = \boldsymbol{x})$ along the tree structure allows the computation of entropy profiles, which rely on conditional and partial state entropies.

In a first approach, conditional entropies $H(S_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ are directly extracted during the downward recursion (16). Then the conditional entropies $H(\bar{\boldsymbol{S}}_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ and the partial state trees entropies $H(\bar{\boldsymbol{S}}_u | \boldsymbol{X} = \boldsymbol{x})$ are computed using an upward algorithm that requires the results of the upward-downward recursion. They are also used in a downward recursion to compute profiles of partial state tree entropies $H(\bar{\boldsymbol{S}}_{0\setminus u} | \boldsymbol{X} = \boldsymbol{x})$.

In a second approach, conditional entropies $H(\bar{\boldsymbol{S}}_{c(u)} | S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)$ are computed directly during the upward recursion given in Section 3.2, without requiring the downward probabilities

$\xi_u(j)$. These conditional entropies are used to compute directly profiles of partial state tree entropies $H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x})$ and $H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x})$.

We also provide an algorithm to compute conditional entropies given the children states $H(S_u|\boldsymbol{S}_{c(u)},\boldsymbol{X}=\boldsymbol{x})$. We show that contrarily to $H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})$, these quantities do not correspond to local contributions to $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})$, but their sum over all vertices u is lower bounded by $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})$.

Computation of partial state tree entropy using conditional entropy of state subtree given parent state Firstly, for every non-root vertex u, the conditional entropy

$$\begin{aligned}&H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\\&=-\sum_{i,j}P(S_u=j,S_{\rho(u)}=i|\boldsymbol{X}=\boldsymbol{x})\log P(S_u=j|S_{\rho(u)}=i,\boldsymbol{X}=\boldsymbol{x}),\end{aligned}\tag{17}$$

is directly extracted during the downward recursion (16), similarly to (13) for HMC models, with

$$\left\{\begin{array}{ll}P(S_u=j|S_{\rho(u)}=i,\boldsymbol{X}=\boldsymbol{x}) & =\beta_u(j)p_{ij}/\{P(S_u=j)\beta_{\rho(u),u}(i)\}\text{ and}\\ P(S_u=j,S_{\rho(u)}=i|\boldsymbol{X}=\boldsymbol{x}) & =\beta_u(j)p_{ij}\xi_{\rho(u)}(i)/\{P(S_u=j)\beta_{\rho(u),u}(i)\}.\end{array}\right.\tag{18}$$

The partial state tree entropy $H(\bar{\boldsymbol{S}}_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})$ is computed using an upward algorithm. Initialization is achieved at the leaf vertices u using equation (17).
The recursion is given, for all non-root vertices u taken upwards, by:

$$\begin{aligned}H(\bar{\boldsymbol{S}}_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})=&H(\bar{\boldsymbol{S}}_{c(u)}|S_u,S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})+H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\\=&\sum_{v\in c(u)}H(\bar{\boldsymbol{S}}_v|S_u,\boldsymbol{X}=\boldsymbol{x})+H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}).\end{aligned}\tag{19}$$

Equation (19) can be interpreted as the chain rule

$$H(\bar{\boldsymbol{S}}_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})=H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})+\sum_{v\in\mathcal{T}_u}H(S_v|S_{\rho(v)},\boldsymbol{X}=\boldsymbol{x}),\tag{20}$$

deduced from factorization

$$\begin{aligned}P(\bar{\boldsymbol{S}}_u=\bar{\boldsymbol{s}}_u|S_{\rho(u)}=s_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})=&P(S_u=s_u|S_{\rho(u)}=s_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\\&\times\prod_{v\in\mathcal{T}_u}P(S_v=s_v|S_{\rho(v)}=s_{\rho(v)},\boldsymbol{X}=\boldsymbol{x}),\end{aligned}$$

which is similar to Proposition 1. An analogous factorization yields

$$\begin{aligned}H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x})=&H(\bar{\boldsymbol{S}}_{c(u)}|S_u,\boldsymbol{X}=\boldsymbol{x})+H(S_u|\boldsymbol{X}=\boldsymbol{x})\\=&\sum_{v\in c(u)}H(\bar{\boldsymbol{S}}_v|S_u,\boldsymbol{X}=\boldsymbol{x})+H(S_u|\boldsymbol{X}=\boldsymbol{x}).\end{aligned}\tag{21}$$

Thus, profiles of partial state tree entropies $\big(H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x})\big)_{u\in\mathcal{U}}$ can be deduced from $\big(H(\bar{\boldsymbol{S}}_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\big)_{u\in\mathcal{U}}$ and the marginal entropies

$$H(S_u|\boldsymbol{X}=\boldsymbol{x})=-\sum_j\xi_u(j)\log\xi_u(j).$$

The global state tree entropy $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})$ is obtained from (21) at root vertex $u=0$.

Profiles of partial state tree entropies $\left(H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$ can also be computed using the following downward recursion, initialized at every child u of the root vertex by

$$\begin{aligned} H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x}) =& H(S_0|\boldsymbol{X}=\boldsymbol{x}) + H(\bar{\boldsymbol{S}}_{b(u)}|S_0,\boldsymbol{X}=\boldsymbol{x}) \\ =& H(S_0|\boldsymbol{X}=\boldsymbol{x}) + \sum_{v\in b(u)} H(\bar{\boldsymbol{S}}_v|S_0,\boldsymbol{X}=\boldsymbol{x}). \end{aligned} \tag{22}$$

The downward recursion is given at vertex v with parent $u=\rho(v)$ by

$$\begin{aligned} &H(\bar{\boldsymbol{S}}_{0\setminus v}|\boldsymbol{X}=\boldsymbol{x}) \\ &= H(S_u,\bar{\boldsymbol{S}}_{b(v)}|\bar{\boldsymbol{S}}_{0\setminus u},\boldsymbol{X}=\boldsymbol{x}) + H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x}) \\ &= H(\bar{\boldsymbol{S}}_{b(v)}|S_u,\bar{\boldsymbol{S}}_{0\setminus u},\boldsymbol{X}=\boldsymbol{x}) + H(S_u|\bar{\boldsymbol{S}}_{0\setminus u},\boldsymbol{X}=\boldsymbol{x}) + H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x}) \\ &= \sum_{w\in b(v)} H(\bar{\boldsymbol{S}}_w|S_{\rho(w)},\boldsymbol{X}=\boldsymbol{x}) + H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}) + H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x}), \end{aligned} \tag{23}$$

where for any $w\in b(v)$, $\rho(w)=u$.
Note that equations (20), (22) and (23) can be written under the same form: if $\mathcal{V}$ is a subtree of $\mathcal{T}$, then the entropy of state subtree $\bar{\boldsymbol{S}}_{\mathcal{V}}$ is

$$H(\bar{\boldsymbol{S}}_{\mathcal{V}}|\boldsymbol{X}=\boldsymbol{x}) = \sum_{v\in\mathcal{V}} H(S_v|S_{\rho(v)},\boldsymbol{X}=\boldsymbol{x}),$$

where $H(S_v|S_{\rho(v)},\boldsymbol{X}=\boldsymbol{x})$ refers to $H(S_v|\boldsymbol{X}=\boldsymbol{x})$ if v is the root vertex or if $\rho(v)$ does not belong to $\mathcal{V}$.

Recursion (23) can be terminated at any leaf vertex u using the following equation:

$$\begin{aligned} H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) =& H(S_u|\bar{\boldsymbol{S}}_{0\setminus u},\boldsymbol{X}=\boldsymbol{x}) + H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x}) \\ =& H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}) + H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x}). \end{aligned}$$

In summary, the profile of conditional entropies $\left(H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$ is firstly computed using (17). The conditional entropies are used in (19) to derive the partial state tree entropies $H(\bar{\boldsymbol{S}}_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})$, which are combined with the marginal entropies in (21) to derive profiles of partial state tree entropies $\left(H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$. They are also combined with the conditional entropies in (23) to compute the profiles $\left(H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$. The time complexity of the algorithm is in $\mathcal{O}(J^2n)$.

As in HMC models, the marginal entropy profile $\left(H(S_u|\boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$ can be viewed as pointwise upper bounds on the conditional entropy profile $\left(H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$.

Direct computation of conditional entropy of children state subtrees given each state As an alternative, the entropies $H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u)$ can be computed directly during the upward recursion given in Section 3.2. These are similar to the entropies $H(S_0^{t-1}|S_t=j,X_0^t=x_0^t)$, used in the algorithm of Hernando *et al.* (2005) in HMC models. Therefore, the following algorithm can be seen as a generalization of their approach to HMT models. Its specificity, compared with the approach based on the conditional entropies $H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})$, is that it does not require the results of the downward recursion.

This upward algorithm is initialized at the leaf vertices u by

$$H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u)=0.$$

Since $\bar{\boldsymbol{S}}_{c(u)}$ and $\bar{\boldsymbol{X}}_{0\setminus u}$ are conditionally independent given S_u and $\bar{\boldsymbol{X}}_u$, we have for any state j, $H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) = H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \boldsymbol{X} = \boldsymbol{x})$. Combining this equation with (21) yields

$$\begin{aligned} H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) =& H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_{c(u)} = \bar{\boldsymbol{x}}_{c(u)}) \\ =& \sum_{v\in c(u)} H(\bar{\boldsymbol{S}}_v|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v), \end{aligned}$$

which is similar to the backward recursion (30) in time-reversed HMC models (see Appendix A.1).
Moreover, for any $v \in c(u)$ with $c(v) \neq \emptyset$ and for $j = 0, \ldots, J-1$,

$$\begin{aligned} &H(\bar{\boldsymbol{S}}_v|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &= -\sum_{\bar{\boldsymbol{s}}_{c(v)}, s_v} P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}, S_v = s_v|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &\quad \times \log P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}, S_v = s_v|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &= -\sum_{\boldsymbol{s}_{c(v)}} \sum_k P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}|S_v = k, S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &\quad \times \{\log P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}|S_v = k, S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) + \log P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)\} \\ &= -\sum_k P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \left\{ \sum_{\boldsymbol{s}_{c(v)}} P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}|S_v = k, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right. \\ &\quad \left. \times \log P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}|S_v = k, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) + \log P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right\} \\ &= \sum_k P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \left\{ H(\bar{\boldsymbol{S}}_{c(v)}|S_v = k, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right. \\ &\quad \left. - \log P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right\}. \end{aligned} \tag{24}$$

Thus, the recursion of the upward algorithm is given by

$$\begin{aligned} &H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &= \sum_{v\in c(u)} \left\{ \sum_{s_v} P(S_v = s_v|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \left[H(\bar{\boldsymbol{S}}_{c(v)}|S_v = s_v, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right. \right. \\ &\quad \left. \left. - \log P(S_v = s_v|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right] \right\}, \end{aligned} \tag{25}$$

where $P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) = P(S_v = k|S_u = j, \boldsymbol{X} = \boldsymbol{x})$ is given by equation (18).

The termination step is obtained by similar arguments as equation (21):

$$\begin{aligned} H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x}) =& H(\bar{\boldsymbol{S}}_{c(0)}|S_0, \boldsymbol{X} = \boldsymbol{x}) + H(S_0|\boldsymbol{X} = \boldsymbol{x}) \\ =& \sum_j \beta_0(j) \left\{ H\left(\bar{\boldsymbol{S}}_{c(0)}|S_0 = j, \boldsymbol{X} = \boldsymbol{x}\right) - \log \beta_0(j) \right\}. \end{aligned}$$

If each vertex has a single child, HMT and HMC models coincide, and equation (25) appears as a generalization of (14) for the computation of conditional entropies in time-reversed HMCs.

Using similar arguments as in (24), the partial state tree entropy $H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x})$ can be deduced from the conditional entropies $H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u)$ (with $j=0,\dots,J-1$) as follows:

$$\begin{aligned}H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x}) =& H(\bar{\boldsymbol{S}}_{c(u)}|S_u,\boldsymbol{X}=\boldsymbol{x})+H(S_u|\boldsymbol{X}=\boldsymbol{x})\\ =&\sum_j \xi_u(j)\left\{H\left(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\boldsymbol{X}=\boldsymbol{x}\right)-\log\xi_u(j)\right\}\\ =&\sum_j \xi_u(j)\left\{H\left(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u\right)-\log\xi_u(j)\right\},\end{aligned} \tag{26}$$

where the $(\xi_u(j))_{j=0,\dots,J-1}$ are directly extracted from the downward recursion (16). Moreover, since

$$\begin{aligned}H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x}) =& H(\bar{\boldsymbol{S}}_0|\boldsymbol{X}=\boldsymbol{x})-H(\bar{\boldsymbol{S}}_u|S_{0\setminus u},\boldsymbol{X}=\boldsymbol{x})\\ =& H(\bar{\boldsymbol{S}}_0|\boldsymbol{X}=\boldsymbol{x})-H(\bar{\boldsymbol{S}}_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\end{aligned}$$

and since

$$H(\bar{\boldsymbol{S}}_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})=H(\bar{\boldsymbol{S}}_{c(u)}|S_u,\boldsymbol{X}=\boldsymbol{x})+H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}),$$

the partial state tree entropy $H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x})$ can also be deduced from the conditional entropies $\left(H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u)\right)_{j=0,\dots,J-1}$ using

$$\begin{aligned}&H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x})\\ &=H(\bar{\boldsymbol{S}}_0|\boldsymbol{X}=\boldsymbol{x})-\sum_j \xi_u(j)H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u)-H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}),\end{aligned} \tag{27}$$

but the computation of $H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})$ using (17) is still necessary.

In summary, the profile of partial subtrees entropies $\left(H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u)\right)_{u\in\mathcal{U};\ j=0,\dots,J-1}$ is firstly computed using (25). The profile of partial state tree entropies $\left(H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$ is deduced from these entropies and the smoothed probabilities, using (26). Computation of partial state tree entropies $\left(H(\bar{\boldsymbol{S}}_{0\setminus u}|\boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$ and conditional entropies $\left(H(S_u|S_{\rho(u)},\ \boldsymbol{X}=\boldsymbol{x})\right)_{u\in\mathcal{U}}$ still relies on (23) and (17), essentially, although variant (27) remains possible. The time complexity of the algorithm is in $\mathcal{O}(J^2n)$.

Entropy profiles conditioned on the children states in HMT models Up to this point, the proposed profile of conditional entropies has the property that global state tree entropy is the sum of conditional entropies. This is a consequence of Corollary 1, which translates into HMT models by profiles of state entropy given the parent state.

However, as will be shown in Section 4 (Application), the state uncertainty at vertex u may be better explained by the values of children states than that of the parent state in practical situations. Consequently, profiles based on $H(S_u|\bar{\boldsymbol{S}}_{c(u)},\boldsymbol{X}=\boldsymbol{x})$ have practical importance and are derived below. Since S_u is conditionally independent from $\{S_v\}_{v\in c(u)}$ given $\boldsymbol{S}_{c(u)}$ and $\boldsymbol{X}$, we have $H(S_u|\bar{\boldsymbol{S}}_{c(u)},\boldsymbol{X}=\boldsymbol{x})=H(S_u|\boldsymbol{S}_{c(u)},\boldsymbol{X}=\boldsymbol{x})$. This quantity, bounded from above by the marginal entropy $H(S_u|\boldsymbol{X}=\boldsymbol{x})$, is computed as follows:

$$\begin{aligned}H(S_u|\boldsymbol{S}_{c(u)},\boldsymbol{X}=\boldsymbol{x})=-\sum_j\sum_{\boldsymbol{s}_{c(u)}}& P(S_u=j,\boldsymbol{S}_{c(u)}=\boldsymbol{s}_{c(u)}|\boldsymbol{X}=\boldsymbol{x})\\ &\times\log P(S_u=j|\boldsymbol{S}_{c(u)}=\boldsymbol{s}_{c(u)},\boldsymbol{X}=\boldsymbol{x}),\end{aligned}$$

with

$$P(S_u = j, \boldsymbol{S}_{c(u)} = \boldsymbol{s}_{c(u)} | \boldsymbol{X} = \boldsymbol{x}) = \xi_u(j) \prod_{v \in c(u)} P(S_v = s_v | S_u = j, \boldsymbol{X} = \boldsymbol{x}) \quad (28)$$
$$= \xi_u(j) \prod_{v \in c(u)} \frac{\beta_v(s_v) p_{j s_v}}{P(S_v = s_v) \beta_{u,v}(j)}$$

from equation (18), and where equation (28) comes from conditional independence of $\{S_v\}_{v \in c(u)}$ given S_u. The quantities $\beta_v(k)$, $\beta_{\rho(v),v}(j)$ and $P(S_v = k)$ are directly extracted from the upward recursion in Section 3.2. Consequently,

$$P(S_u = j | \boldsymbol{S}_{c(u)} = \boldsymbol{s}_{c(u)}, \boldsymbol{X} = \boldsymbol{x}) = \frac{\xi_u(j) \prod_{v \in c(u)} [p_{j s_v} / \beta_{u,v}(j)]}{\sum_k \xi_u(k) \prod_{v \in c(u)} [p_{k s_v} / \beta_{u,v}(k)]}.$$

Note that the time complexity of the algorithm for computing entropy profiles conditioned on the children states is in $\mathcal{O}(J^{c+1} n)$ in the case of $c-$ary trees. This makes it the only algorithm among those in this article whose complexity is not in $\mathcal{O}(J^2 n)$.

The profiles based on $H(S_u | \boldsymbol{S}_{c(u)}, \boldsymbol{X} = \boldsymbol{x})$ satisfy the following property:

Proposition 2

$$H(\boldsymbol{S} | \boldsymbol{X} = \boldsymbol{x}) \leq \sum_u H(S_u | \boldsymbol{S}_{c(u)}, \boldsymbol{X} = \boldsymbol{x})$$

where $H(S_u | \boldsymbol{S}_{c(u)}, \boldsymbol{X} = \boldsymbol{x})$ must be understood as $H(S_u | \boldsymbol{X} = \boldsymbol{x})$ if u is a leaf vertex.

Thus, these entropies cannot be interpreted at the local contribution of vertex u to global state tree entropy $H(\boldsymbol{S} | \boldsymbol{X} = \boldsymbol{x})$, unless equality is obtained in the above equation. (For example, if $\mathcal{T}$ is a linear tree, or in other words a sequence.) To assess the difference between the right-hand and the left-hand parts of the above inequality in practical situations, numerical experiments are performed in Section 4 (Application).

A proof of Proposition 2 is given in Appendix A.2. A consequence of this inequality is that factorization

$$\begin{aligned} &P(S_u = j, S_{c(u)} = s_{c(u)} | \boldsymbol{X} = \boldsymbol{x}) \\ &= P(S_u = j | \boldsymbol{S}_{c(u)} = \boldsymbol{s}_{c(u)}, \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{S}_{c(u)} = \boldsymbol{s}_{c(u)} | \boldsymbol{X} = \boldsymbol{x}) \end{aligned}$$

cannot be pursued through a recursion on the children of u. Essentially, this comes from the fact that any further factorization based on conditional independence between the $(S_v)_{v \in c(u)}$ must involve S_u.

4 Applications of entropy profiles

To illustrate the practical ability of entropy profiles to provide localized information on the state sequence uncertainty, two cases of application are considered. The first case consists of the HMC analysis of the earthquake dataset, published by Zucchini & MacDonald (2009). The second case consists of the HMT analysis of the structure of pine branches, using an original dataset. It is shown in particular that entropy profiles allow regions that are non-ambiguously explained by the estimated model to be differentiated from regions that are ambiguously explained. Their ability to provide accurate interpretation of the model states is also emphasized.

4.1 HMC analysis of earthquakes

The data consists of a single sequence of annual counts of major earthquakes (defined as of magnitude 7 and above) for the years 1900-2000; see Figure 2.

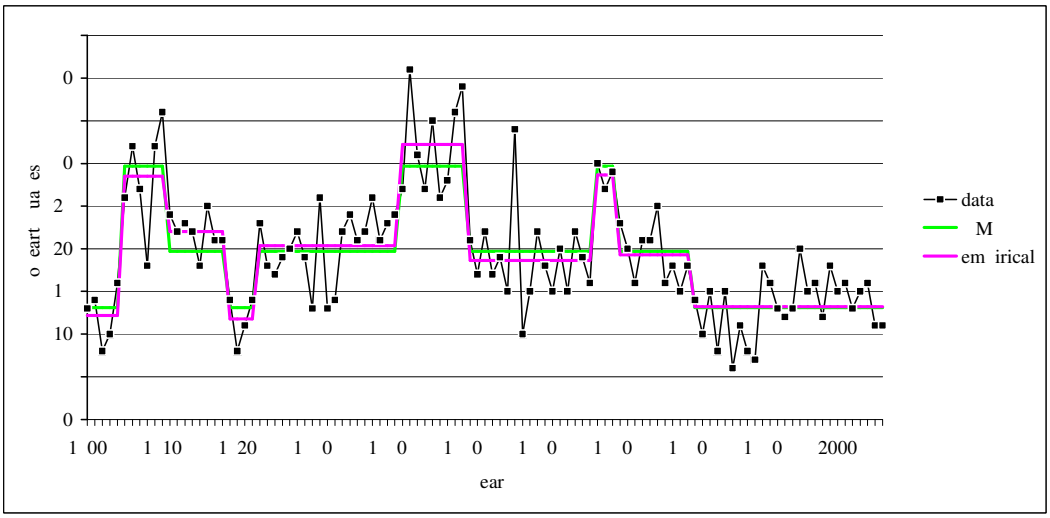


Figure 2: *Earthquake data: Restored state sequence represented as step functions, the level of the segments being either the parameter $\widehat{\lambda}_j$ of the Poisson observation distributions corresponding to the restored state j or the empirical mean estimated for the segment.*

A 3-state stationary HMC model with Poisson observation distributions was estimated on the basis of this earthquake count sequence and the estimated parameters of the Poisson observation distributions were $\widehat{\lambda}_1 = 13.1$, $\widehat{\lambda}_2 = 19.7$ and $\widehat{\lambda}_3 = 29.7$. The restored state sequence is represented in Figure 2 as step functions, the level of the segments being either the parameter $\widehat{\lambda}_j$ of the Poisson observation distributions corresponding to the restored state j or the empirical mean estimated for the segment. The state profiles computed by the forward backward algorithm $\{P\left(S_t = j|\boldsymbol{X} = \boldsymbol{x}\right); j = 0, \ldots, J-1; t = 0, \ldots, T-1\}$ are shown in Figure 3. The entropy of the state sequence that explains the observed sequence for the estimated HMC model is bounded from above by the sum of the marginal entropies

$$\begin{aligned} H\left(S_0^{T-1}|\boldsymbol{X} = \boldsymbol{x}\right) &= \sum_t H\left(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}\right) = 14.9 \\ &< \sum_t H\left(S_t|\boldsymbol{X} = \boldsymbol{x}\right) = 19.9. \end{aligned}$$

For this example, we chose to show the entropies conditional on the past, which are the only meaningful conditional entropies. Since $\log J$ is an upper bound on $H(S_t|\boldsymbol{X} = \boldsymbol{x})$, the scale of these entropy profiles is in theory $[0, \log 3]$. However the scale of the entropy profiles is rather $[0, \log 2]$, since in practice at most two states can explain a given observation equally well; see Figure 4.

Ignoring the dependency structure within the model to assess state uncertainty leads to strong overestimation of this uncertainty. This is highlighted in Figure 4 by the comparison of the profile of entropies conditional on the past and the profile of marginal entropies, and in Figure 5, by

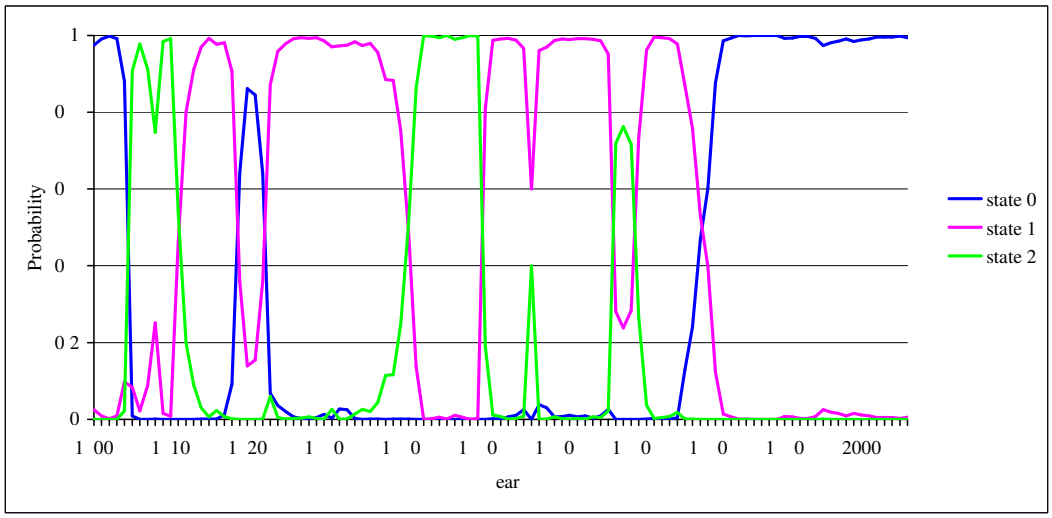


Figure 3: *Earthquake data: state profiles computed by the forward-backward algorithm.*

the comparison of the profile of partial state sequence entropies and the profile of cumulative marginal entropies. It should be recalled that the marginal entropy profile is a direct summary of the uncertainty reflected in the smoothed probability profiles shown in Figure 3. Hence, such profiles should be interpreted with caution.

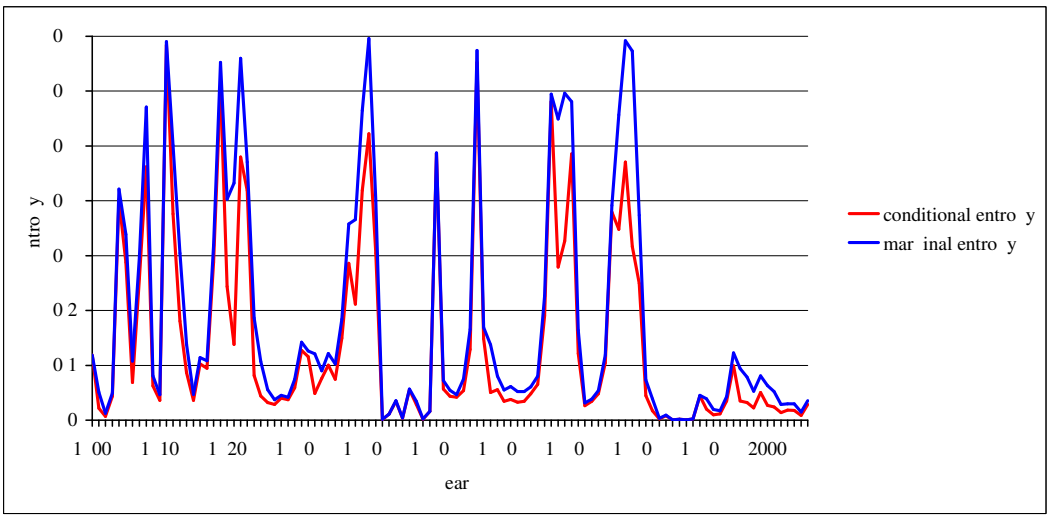


Figure 4: *Earthquake data: Profiles of entropies conditional on the past and of marginal entropies.*

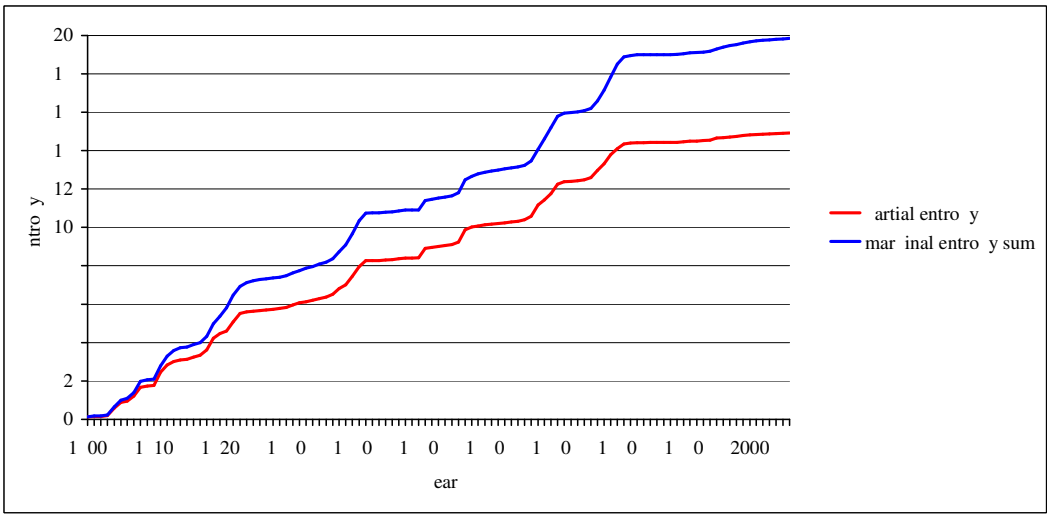


Figure 5: *Earthquake data: Profiles of partial state sequence entropies and of cumulative marginal entropies.*

4.2 Analysis of the structure of Aleppo pines

The aim of this study was to provide a model of the architecture of Aleppo pines. The data set is composed of seven branches of Allepo pines (*Pinus Halepensis* Mill., *Pinaceae*) planted in the south of France (Clapiers, Hérault). The branches come from seven different individuals aged between 35 to 40 years. They were described at the scale of annual shoot, defined as the segment of stem established within a year. Five variables were recorded for each annual shoot: length (in cm), number of branches per tier, number of growth cycles and presence or absence of female cones and of male cones. During a year, the growth of an annual shoot can occur in one to three cycles. An annual shoot with several growth cycles is said to be *polycyclic*. The number of growth cycles beyond the first one corresponds to the third recorded variable. On these seven branches, a total of 836 annual shoots was measured.

4.2.1 Competing models

An HMT model was estimated on basis of the seven branches, to identify classes of annual shoots with comparable values for the variables, and to characterize the succession of the classes within the branches. The branches were considered as mutually independent random realizations of a same HMT model. The emission distributions were multinomial distributions $\mathcal{M}(1; p_1, \ldots, p_V)$ for each variable but the length variable, where V denotes the number of possible values for this variable. The length variable, if included in the model, was assumed to follow a negative binomial distribution, given the state. The five variables were assumed independent given the state. The number of HMT states could not be deduced *a priori* from biological arguments, so it had to be determined using statistical criteria. We resorted to the Bayesian Information Criterion (BIC) to select this number. Although the consistency of BIC was proved for a restricted family of HMC models only (see Boucheron and Gassiat, 2007), its practical ability to provide useful results is established (see *e.g.* Celeux and Durand, 2008). The maximal number of possible states was set to 10. For HMT models where the length variable was discarded, BIC selected a 5-state or

a 6-state model (with respective values of BIC -2,047 and -2,039). The third best model had 4 states, with a BIC value of -2,074. In the case of models including the length variable, a 6-state model was selected (with a BIC value of -10,541) followed by 4-state and 5-state models (with respective values of BIC -10,545 and -10,558). Note that since the estimated HMT models were not ergodic, the theoretical properties of BIC are not established.

4.2.2 Entropy profiles in the 5-state HMT model without length variable

The estimated transition matrix of the 5-state HMT model is

$$\hat{P} = \begin{bmatrix} 0.18 & 0.47 & 0.33 & 0.02 & 0 \\ 0.01 & 0.51 & 0.45 & 0.00 & 0.03 \\ 0 & 0 & 0.04 & 0.96 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

and the Markov tree is initialized in state 0 with probability 1. It can be seen from $\hat{P}$ that the Markov tree has transient states 0 and 1 and an absorbing class $\{2;3;4\}$, in which the states alternate quasi systematically.

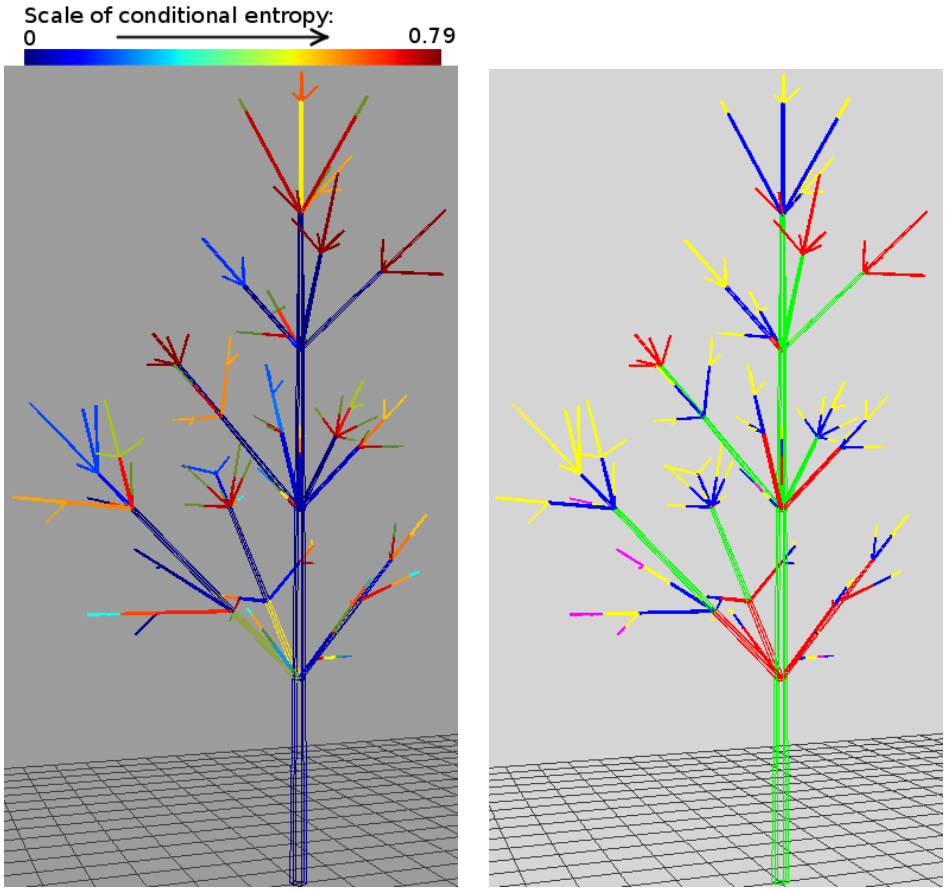
Female cones are potentially present in state 0 only (in state 0, a shoot has female cones with probability 0.14). Male cones are potentially present in state 4 only (a shoot has male cones with probability 0.66). Besides, state 0 is characterized by a high branching intensity (0 to 8 branches) and frequent polycyclism (a shoot is polycyclic with probability 0.95). State 1 is characterized by intermediate branching intensity (0 to 3 branches, unbranched with probability 0.67) and monocyclism. State 2 is characterized by intermediate branching intensity (0 to 4 branches, unbranched with probability 0.81) and rare polycyclism (a shoot is polycyclic with probability 0.06). States 3 and 4 are always monocyclic, and are mostly unbranched (with probability 0.94 and 0.98, respectively). As a consequence, any unbranched, monocyclic, sterile shoot can be in any of the 5 states (respectively with probability 0.002, 0.248, 0.281, 0.346 and 0.123).

From a biological point of view, this model highlights a gradient of vigour, since the states are ordered with decreasing number of growth cycles and branches. This also predicts that class $\{2;3;4\}$ is composed by sterile shoots that have potential polycyclism, alternating with sterile monocyclic shoots, and finally shoots with potential male sexuality.

In the dataset, shoots with male cones (referred to as *male shoots* hereafter) systematically follow sterile shoots. Moreover, they are either located at the tip of a branch, or followed by a unique sterile shoot. This is a consequence of a particular measurement protocol for this dataset, in which individuals were measured just after the occurrence of the first male cones. In contrast, the infinite alternation of two sterile shoots and one male shoot predicted by this model cannot be considered as a general pattern in the pine architecture. A more relevant hypothesis is that after several years of growth, only unbranched monocyclic sterile shoots are produced (or maybe a mixture of both such male and sterile shoots).

To analyze how state ambiguity due to unbranched, monocyclic, sterile shoots affects state restoration, entropy profiles were computed for each branch. Firstly, the annual shoots were represented using a colormap, which is a mapping between colours and the values of conditional entropies $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ (see Figure 6a)). Vertices with lowest conditional entropy are represented in blue, whereas those with highest conditional entropy are in red. In a similar way, the marginal entropy could also be represented using a colormap.

The most likely state tree for each individual was computed using the Viterbi algorithm for HMT models (Durand *et al.*, 2004). This state tree is represented in Figure 6b). This representation shows where the states are located within the tree; for example state 0 is located



a) b)

Figure 6: *Conditional entropy and state tree restoration for a given branch. a) Conditional entropy* $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ *using a colormap. Blue corresponds to lowest entropy and red to highest entropy. b) State tree restoration. The correspondence between states and colours is as follows: state 0 - green ; state 1 - red ; state 2 - blue ; state 3 - yellow ; state 4 - magenta.*

on the main axis (main stem) and at the basis of lateral axis. Moreover, in conjunction with Figure 6a), it highlights some states for which the restoration step is not much ambiguous (in our example, state 0, and to a least extent, state 4). Thus, these states with low entropy correspond to vertices with the highest number of branches, female or male cones. On the contrary, the vertices with highest entropy are mostly unbranched, monocyclic and sterile, and are located at peripheral parts of the plant.

Using the conditional entropy in Figure 6a), peripheral vertices with maximal or minimal conditional entropy can be selected. To further interpret the model with respect to the data, entropy profiles were computed along paths leading to these vertices. These profiles were complemented by so-called *upward-downward Viterbi profiles*. These profiles rely on the following quantities

$$\max_{(\boldsymbol{s}_v)_{v\neq u}} P((S_v = s_v)_{v\neq u}, S_u = j|\boldsymbol{X} = \boldsymbol{x}),$$

for each state j and each vertex u of the tree. Their computation is based on upward and downward dynamic programming recursions, similar to that of Brushe *et al.* (1998), and are not detailed in this paper. Such profiles provide an overview of local alternatives to the state tree restoration given by the Viterbi algorithm. They were used by Guédon (2007) as diagnostic tools for localization of state uncertainty in the context of hidden (semi-)Markov chains. A detailed analysis of the state uncertainty is provided by the entropy profiles.

Female shoots To illustrate how entropy reduction and Viterbi profiles are connected, an example consisting of a path containing a female shoot is considered. This path corresponds to the main axis of the third individual (for which $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) = 52.9$). The path contains 6 vertices, referred to as $\{0,\ldots,5\}$. The female shoot is at vertex 2, and vertex 3 is a bicyclic shoot. Since a female shoot necessarily is in state 0, $H(S_2|\boldsymbol{X}=\boldsymbol{x}) = 0$ (no uncertainty). Since state 0 is quasi systematically preceded by state 0, shoots 0 and 1 are in state 0 with a very high probability and again, $H(S_u|\boldsymbol{X}=\boldsymbol{x}) \approx 0$ for $u = 1, 2$. Shoots 3 is bicyclic, and thus is in state 0 with a very high probability ($H(S_3|\boldsymbol{X}=\boldsymbol{x}) \approx 0$). Shoots 4 and 5, as unbranched, monocyclic, sterile shoots can be in any state. However, due to several impossible transitions in matrix $\hat{P}$, only the following four configurations have non-negligible probabilities for $(S_4, S_5) : (2,3), (1,1), (1,2)$ and $(3,4)$. This is partly highlighted in Figure 7 c) by the Viterbi profile, and results into high mutual information between S_4 and S_5 given $\boldsymbol{X}=\boldsymbol{x}$. For example, $P(S_5 = 3|S_4 = 2, \boldsymbol{X}=\boldsymbol{x})$, $P(S_5 = 4|S_4 = 3, \boldsymbol{X}=\boldsymbol{x})$ and $P(S_5 \in \{1,2\}|S_4 = 1, \boldsymbol{X}=\boldsymbol{x})$ are very close to 1. Thus the downward conditional entropy $H(S_5|\bar{\boldsymbol{S}}_{0\setminus 5}, \bar{\boldsymbol{X}}_0 = \bar{\boldsymbol{x}}_0) = H(S_5|S_4, \bar{\boldsymbol{X}}_0 = \bar{\boldsymbol{x}}_0) = 0.1$, whereas $H(S_5|\bar{\boldsymbol{X}}_0 = \bar{\boldsymbol{x}}_0) = 0.8$. Similarly, the upward conditional entropy $H(S_4|\bar{\boldsymbol{S}}_{c(4)}, \bar{\boldsymbol{X}}_0 = \bar{\boldsymbol{x}}_0) = H(S_4|S_5, \bar{\boldsymbol{X}}_0 = \bar{\boldsymbol{x}}_0)$ is 0.5 whereas $H(S_4|S_5, \bar{\boldsymbol{X}}_0 = \bar{\boldsymbol{x}}_0) = 1.1$ – see both entropy profiles in Figure 7 a) and b). Since there practically is no uncertainty on the value of S_3, the mutual information between S_3 and S_4 given $\boldsymbol{X}=\boldsymbol{x}$ is very low.

Using equation (21), the contribution of the vertices of the considered path $\mathcal{P}$ to the global state tree entropy can be computed as:

$$H(S_0|\boldsymbol{X}=\boldsymbol{x}) + \sum_{\substack{u\in\mathcal{P}\\ u\neq 0}} H(S_u|S_{\rho(u)}, \boldsymbol{X}=\boldsymbol{x}), \tag{29}$$

and is equal to 1.24 in the above example (that is, 0.21 per vertex on average). The global state tree entropy for this individual is 0.37 per vertex, against 0.38 per vertex in the whole dataset.

The contribution of $\mathcal{P}$ to the global state tree entropy corresponds to the sum of the heights of every point of the profile of entropy given parent state in Figure 7b). The mean marginal state entropy for this individual is 0.44 per vertex, which strongly overestimates the mean state tree entropy.

4.2.3 Entropy profiles in the 6-state HMT model without length variable

To assess the ability of the 5-state and the 6-state HMT models to provide state restorations with low uncertainty and relevant interpretation of the results, both models are compared using entropy and Viterbi profiles.

The estimated transition matrix of the 6-state HMT model without the "length" variable is

$$\hat{P} = \begin{bmatrix} 0.16 & 0.56 & 0.16 & 0.12 & 0 & 0 \\ 0 & 0.42 & 0.01 & 0.55 & 0 & 0.02 \\ 0 & 0 & 0.02 & 0.62 & 0.36 & 0 \\ 0 & 0 & 0 & 0.10 & 0.90 & 0 \\ 0 & 0 & 0.01 & 0 & 0 & 0.99 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

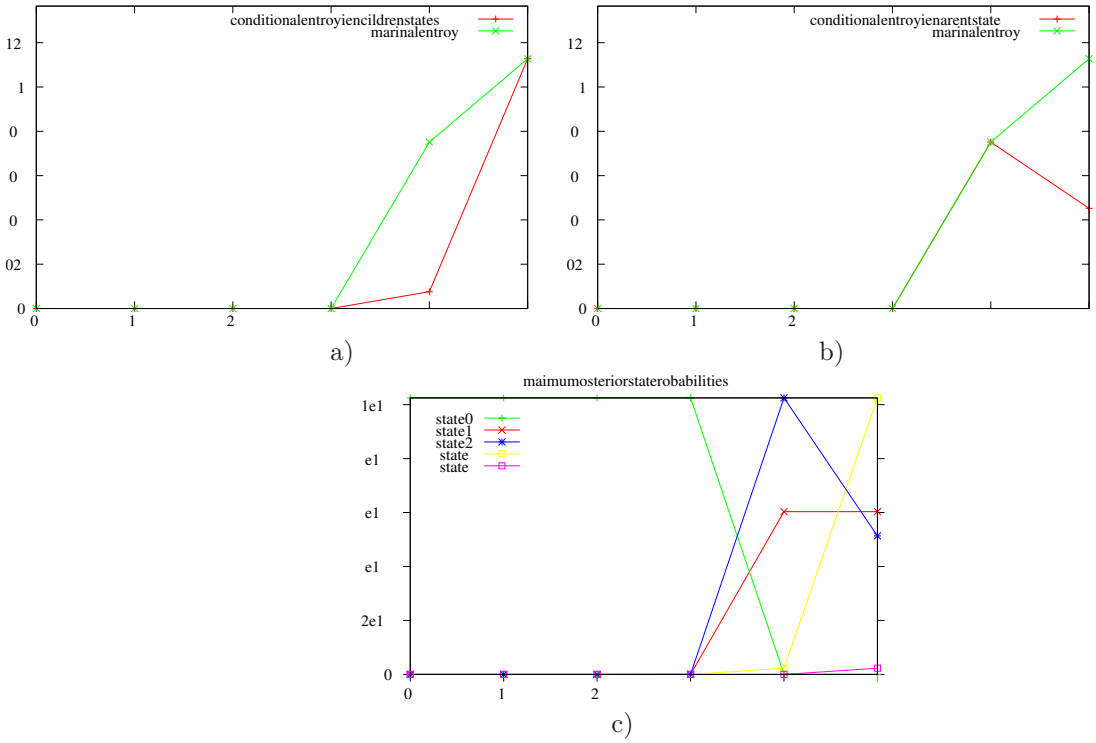


Figure 7: *Entropy profiles along a path containing a female shoot, obtained with a 5-state HMT model without the "length" variable. a) Marginal and conditional entropy given children states. b) Marginal and conditional entropy given parent state. c) State tree restoration with the Viterbi upward-downward algorithm.*

and the Markov tree is initialized in state 0 with probability 1. It can be seen from $\hat{P}$ that the Markov tree has transient states 0 and 1 and an absorbing class $\{2; 3; 4; 5\}$. Any return from state 3, 4 or 5 to state 2 is actually rare, and states 3 to 5 alternate most of the time.

Female cones are potentially present in state 0 only (a shoot has female cones with probability 0.22 in state 0). Male cones are potentially present in state 5 only (a shoot has male cones with probability 0.62). Besides, state 0 is characterized by a high branching intensity (0 to 8 branches) and frequent polycyclism (a shoot is polycyclic with probability 0.92). State 1 is characterized by low branching intensity (0 to 2 branches, unbranched with probability 0.56) and monocyclism. State 2 is characterized by intermediate branching intensity (0 to 6 branches, unbranched with probability 0.21) and bicyclism (a shoot is bicyclic with probability 0.99). States 3 to 5 are always monocyclic, and are mostly unbranched (with probability 0.87, 0.94 and 0.98, respectively). As a consequence, any unbranched, monocyclic, sterile shoot can be in any of the states 0, 1, 3, 4 and 5 (respectively with probability 0.003, 0.205, 0.316, 0.342 and 0.134). States 1, 3 and 4 have rather similar characteristics, although they slightly differ by their branching densities. These states are essentially justified by their particular positions in the plant. The role of state 4 is mainly to represent the state-transition pattern 345, composed by two sterile and one male shoot. In usual Markovian modelling, a binary pattern 001 for the "male cone" variable could be modeled by a second-order markov model, or by a semi-Markov model with Bernoulli sojourn time in value 0. Here, since a first-order Markov tree is considered, state 4 may be thought of as

an additional necessary state to represent this pattern.

From a biological point of view, the approximate reduction of the number of growth cycles and branches along the states is relevant. However, an absorbing class where two sterile shoots and one male shoot tend to indefinitely alternate does not seem justified.

The global state entropy on the whole dataset is 0.36 per vertex on average. This quantity is slightly less than that of the 5-state HMT model. However, the entropy can increase locally on some particular paths.

Female shoots An example consisting in the same branch and path than in Section 4.2.2 is considered (branch with a female shoot). Let us recall that the female shoot is at vertex 2, and vertex 3 is a bicyclic shoot. As in the case of a 5-state model, there is not much uncertainty on the state values at vertices 0 to 3. Only three configurations have non-negligible probabilities for $(S_4, S_5) : (3, 4), (4, 5)$ and $(3, 3)$. The last two configurations are at most 4 times less likely than the most likely configuration. As a consequence, the number and probabilities of the suboptimal state trees is lower for the 6-state model than for the 5-state model (see Figures 7 c) and 8 c)), and the values in the downward entropy profile are also lower (see Figures 7 b) and 8 b)).

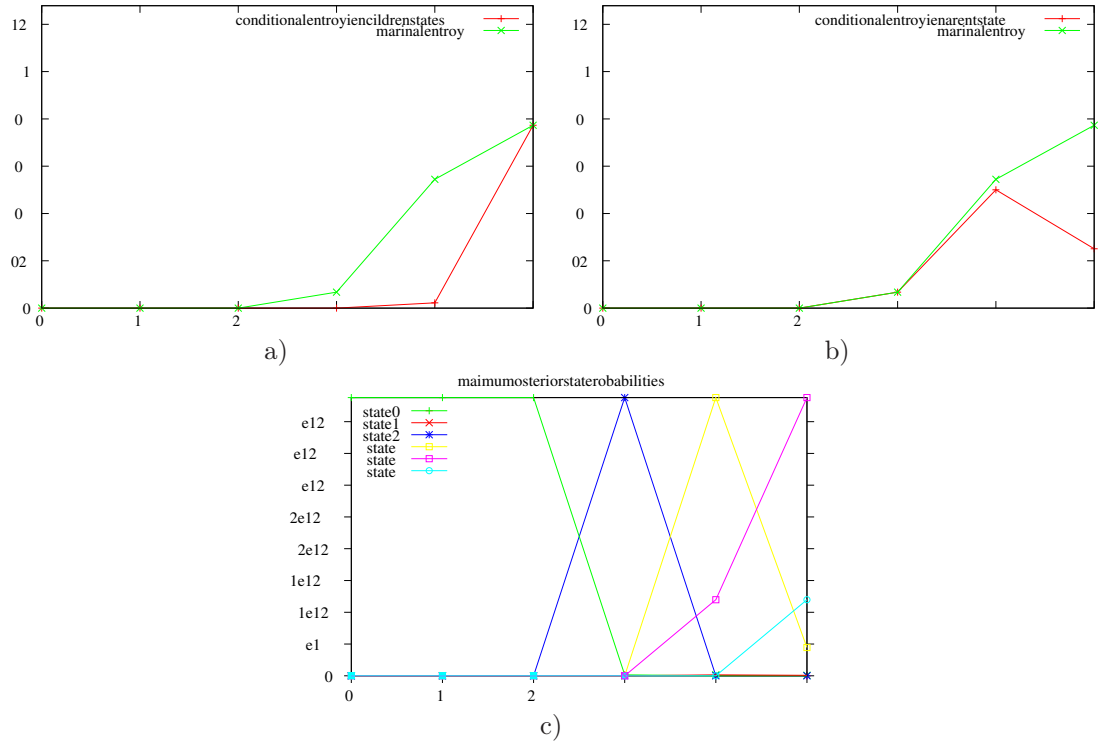


Figure 8: *Entropy profiles along a path containing a female shoot, obtained with a 6-state HMT model without the "length" variable. a) Marginal and conditional entropy given children states. b) Marginal and conditional entropy given parent state. c) State tree restoration with the Viterbi upward-downward algorithm.*

The global state tree entropy for this individual is 0.35 per vertex, and the contribution of the considered path to the global state entropy is 0.17 per vertex, which is lower than for a 5-state

model (*i.e.* 0.21 per vertex).

4.2.4 Entropy profiles in the 6-state HMT model with length variable

The estimated transition matrix of the 6-state HMT model with the "length" variable is

$$\hat{P} = \begin{bmatrix} 0.17 & 0.14 & 0.44 & 0.01 & 0 & 0.24 \\ 0 & 0.18 & 0.18 & 0 & 0 & 0.64 \\ 0 & 0.07 & 0.03 & 0.90 & 0 & 0 \\ 0 & 0.07 & 0.03 & 0 & 0.76 & 0.14 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The hidden states and the state transitions, represented in Figure 9, have the following interpretation. The Markov tree is initialized in state 0 with probability 1. It can be seen from $\hat{P}$ that the Markov tree has transient states 0, transient class $\{1, 2, 3\}$, and two absorbing states 4 and 5. The only possible transitions to a previously-visited state are $2 \to 1$, $3 \to 1$ and $3 \to 2$.

The states are ordered by decreasing length, except for state 5, which has slightly longer shoots than state 4. Female cones are potentially present in state 0 only (a shoot has female cones with probability 0.13 in state 0). Male cones are potentially present in state 4 essentially, and any shoot in state 4 has male cones with probability 1. Male cones may also be present in states 0 and 5 (with probability 0.02 and 0.03, respectively). Besides, state 0 is characterized by a high branching intensity (0 to 8 branches) and frequent polycyclism (a shoot is polycyclic with probability 0.89). State 1 is characterized by intermediate branching intensity (1 to 3 branches, never unbranched) and monocyclism with rare bicyclism (a shoot is monocyclic with probability 0.96). State 2 is characterized by low branching intensity (0 to 3 branches, unbranched with probability 0.74) and monocyclism with rare bicyclism (a shoot is monocyclic with probability 0.9). States 3 to 5 are always monocyclic, and are mostly unbranched (with probability 0.94 and 0.98, respectively). As a consequence, any unbranched, monocyclic, sterile shoot can be in any of the states 0, 2, 3 and 5 (respectively with probability 0.001, 0.261, 0.367 and 0.371). This characteristic of the model will be shown to be the source of state uncertainty for such shoots. States 3 and 5 differ mostly by their shoot length distributions.

From a biological point of view, this model highlights a gradient of vigor, since the states are ordered by decreasing length, and also roughly by decreasing number of growth cycles and branches. The existence of an absorbing class corresponding to unbranched, monocyclic, shoots of short length (either male or sterile) predicted by this estimated HMT model is more consistent with biological *a priori* knowledge on Aleppo pine architecture, than the models in Sections 4.2.2 and 4.2.3.

A detailed analysis of state uncertainty has been performed on three paths (extracted from two distinct individuals), chosen for the contrasted situations they yield:

Case 1) Female shoots Firstly, the same path containing a female shoot as in Sections 4.2.2 and 4.2.3 is considered. Let us recall that the female shoot is at vertex 2, and vertex 3 is a bicyclic shoot. Since a female shoot necessarily is in state 0, $H(S_2|\boldsymbol{X} = \boldsymbol{x}) = 0$ (no uncertainty). Since state 0 is systematically preceded by state 0, shoots 0 and 1 are in state 0 with probability one and again, $H(S_u|\boldsymbol{X} = \boldsymbol{x}) = 0$ for $u = 0, 1$. Shoot 3 is bicyclic, and thus is in state 0 with a very high probability ($H(S_3|\boldsymbol{X} = \boldsymbol{x}) \approx 0$). Shoots 4 and 5, as unbranched, monocyclic, sterile shoots can be in any state, except states 1 and 4. However, due to several impossible transitions in matrix $\hat{P}$, and given the lengths of these shoots, only the following three configurations have non-negligible probabilities for (S_4, S_5) : $(5, 5), (2, 3)$ and $(3, 5)$. This is partly highlighted in

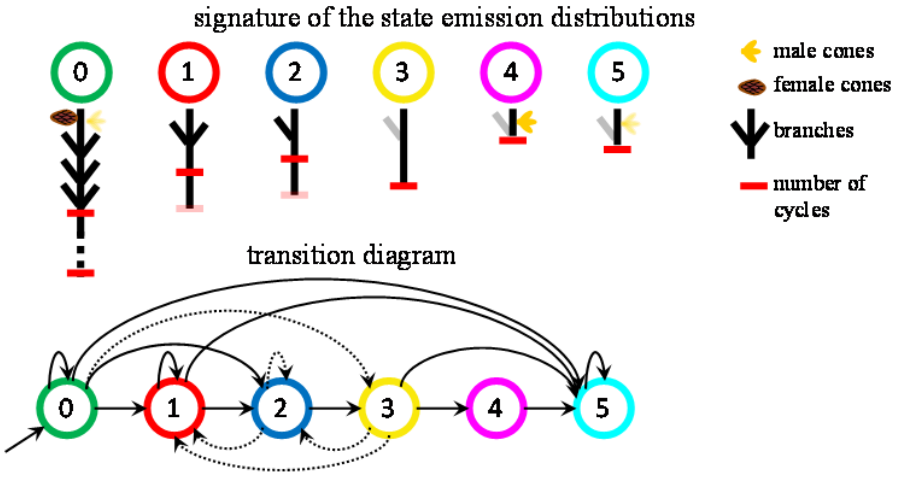


Figure 9: *6-state HMT model: transition diagram and symbolic representation of the state signatures (conditional mean values of the variables given the states, depicted by typical shoots). Dotted arrows correspond to transitions with associated probability* < 0.1. *Mean shoot lengths given each state are proportional to segment lengths, except for state 0 (which mean length is slightly more than twice the mean length for state 1).*

Figure 10c) by the Viterbi profile. As a consequence, S_5 can be deduced from S_4, which results into high mutual information between S_4 and S_5 given $\boldsymbol{X} = \boldsymbol{x}$. Thus the conditional entropy $H(S_5|\bar{\boldsymbol{S}}_{0\setminus 5}, \boldsymbol{X} = \boldsymbol{x}) = H(S_5|S_4, \boldsymbol{X} = \boldsymbol{x})$ is 0.02, whereas $H(S_5|\boldsymbol{X} = \boldsymbol{x}) = 0.46$. Similarly, the conditional entropy $H(S_4|\bar{\boldsymbol{S}}_{c(4)}, \boldsymbol{X} = \boldsymbol{x}) = H(S_4|S_5, \boldsymbol{X} = \boldsymbol{x})$ is 0.02, whereas $H(S_4|\boldsymbol{X} = \boldsymbol{x}) = 0.46$, as illustrated by both entropy profiles in Figure 10a) and b). Since there practically is no uncertainty on the value of S_3, the mutual information between S_3 and S_4 given $\boldsymbol{X} = \boldsymbol{x}$ is very low.

The contribution of the vertices of the considered path $\mathcal{P}$ to the global state tree entropy is equal to 0.48 in the above example (that is, 0.08 per vertex on average), which is far less than for both models without the "length" variable. The global state tree entropy for this individual is 0.21 per vertex, against 0.20 per vertex in the whole dataset. This illustrates that incorporating the length variable into the HMT model strongly reduces uncertainty on the state trees. The mean marginal state entropy for this individual is 0.37 per vertex, which strongly overestimates the mean state tree entropy.

Case 2) Sterile shoots Then, focus is put on a path essentially composed by monocyclic, sterile shoots in the fourth individual (for which $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x}) = 47.5$). The path contains 5 vertices, referred to as $\{0, \ldots, 4\}$. Shoots 0 and 1 are long and highly branched, and thus are in state 0 with probability ≈ 1 (also, shoot 0 is bicyclic). Shoots 2 to 4 are monocyclic and sterile. Shoots 2 and 3 bear one branch, and can be in states 1 or 2 essentially. Shoot 4 is unbranched and from the Viterbi profile in Figure 11c), it can be in states 2, 3 or 5. This is summarized by the entropy profile in Figure 11b). Since there is no uncertainty on S_1, $H(S_2|S_1, \boldsymbol{X}_0 = \boldsymbol{x}_0) = H(S_2|S_1, \boldsymbol{X}_0 = \boldsymbol{x}_0)$, as shown in Figure 11 b). Moreover, from the Viterbi profile, only the following three configurations for (S_2, S_3) have non-negligible probabilities: $(2,1), (1,1)$ and $(2,2)$, and $S_2 = 2$ has highest probability. Since S_3 cannot be deduced from $S_2 = 2$,

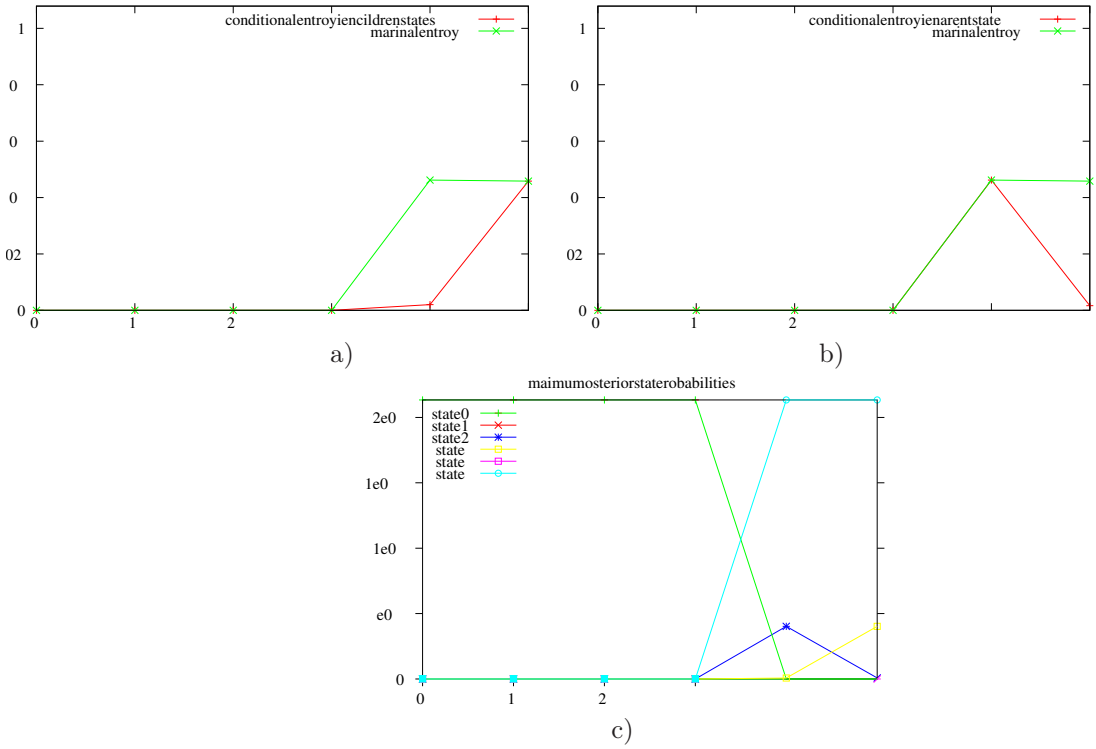


Figure 10: *Entropy profiles along a path containing a female shoot. a) Marginal and conditional entropy given children states. b) Marginal and conditional entropy given parent state. c) State tree restoration with the Viterbi upward-downward algorithm.*

$H(S_3|S_2, \boldsymbol{X}_0 = \boldsymbol{x}_0)$ is rather high. Similarly, only the following three configurations for (S_3, S_4) have non-negligible probabilities: $(1,5), (1,2)$ and $(2,3)$ and $S_3 = 2$ has low probability, so that $H(S_4|S_3, \boldsymbol{X}_0 = \boldsymbol{x}_0)$ is rather high.

The profile $H(S_u|\boldsymbol{S}_{c(u)}, \boldsymbol{X} = \boldsymbol{x})$ in Figure 11 a) is interpreted as follows: the marginal entropy of S_2 is high (0.61), and S_2 cannot be deduced from S_3. However, S_2 can be deduced from a brother S_3' of S_3, such as $S_3' = 3$ implies $S_2 = 2$ and $S_3' = 5$ implies $S_2 = 1$ (as would be shown by entropy profiles including S_3'). Hence, $H(S_2|\boldsymbol{S}_{c(2)}, \boldsymbol{X} = \boldsymbol{x})$ is low. This results into high mutual information between S_2 and its children states given $\boldsymbol{X} = \boldsymbol{x}$, as illustrated in the profile Figure 11 d).

The contribution of this path to the global state tree entropy is 1.41 (that is, 0.28 per vertex on average), which is higher than the contribution of the path containing a female cone considered hereabove. This is also higher than the mean contribution in the whole branch (that is, 0.24 per vertex). This is explained by the lack of information brought by the observed variables (several successive sterile monocyclic shoots, which can be in states 1, 2, 3 or 5). The mean marginal state entropy for this individual is 0.37 per vertex, which strongly overestimates the mean state tree entropy. Note that the representation of state uncertainty using profiles of smoothed probabilities induces a perception of global uncertainty on the states along $\mathcal{P}$ equivalent to that provided by marginal entropy profiles. The discrepancy between the profile of partial state entropies along $\mathcal{P}$ and the profile of cumulative marginal entropies is highlighted in Figure 11e).

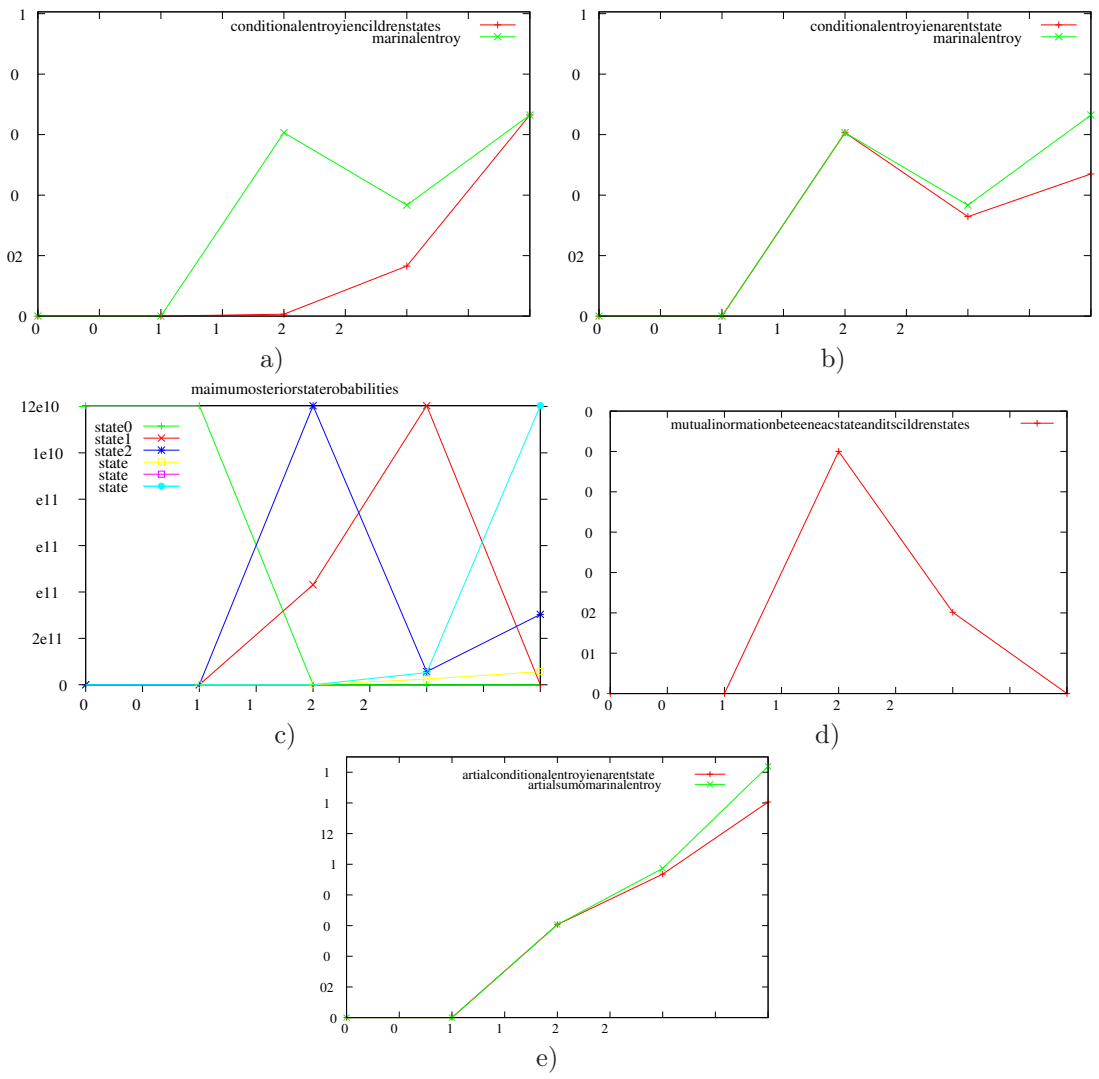


Figure 11: *Entropy profiles along a path containing mainly sterile monocyclic shoots. a) Marginal and conditional entropy given children states. b) Marginal and conditional entropy given parent state. c) State tree restoration with the Viterbi upward-downward algorithm. d) Mutual information between a state and its children states. e) Profiles of partial state sequence and of cumulative marginal entropies.*

Case 3) Male shoots Finally, a path with a terminal male shoot included in the fourth individual is analyzed. The path contains 5 vertices, referred to as $\{0, \ldots, 4\}$. Shoots 0 and 1 are long and highly branched, and thus must be in state 0 (also, shoot 0 is bicyclic). Thus, $H(S_u|\boldsymbol{X} = \boldsymbol{x}) = 0$ for $u = 0, 1$. Shoot 2 is long and unbranched, and thus must be in state 2. Shoot 3 bears one branch, and can be in states 1 or 2 essentially (since $S_1 = 2$ and $\hat{P}_{2,2}$ is low). As a male shoot, shoot 4 is in state 4 with a very high probability, or in state 5 otherwise and $H(S_4|\boldsymbol{X} = \boldsymbol{x}) = 0.08$. Moreover, $S_3 = 0$ if and only if $S_4 = 4$, thus $H(S_4|S_3, \boldsymbol{X} = \boldsymbol{x}) = 0 =$

$H(S_3|S_4, \boldsymbol{X} = \boldsymbol{x})$.

Finally, the contribution of this path to the total entropy is 0.09 (*i.e.* 0.02 per vertex on average), which is negligible. This result is typical of male shoots, which mainly are in state 4, and since state 4 can only be accessed to from state 3.

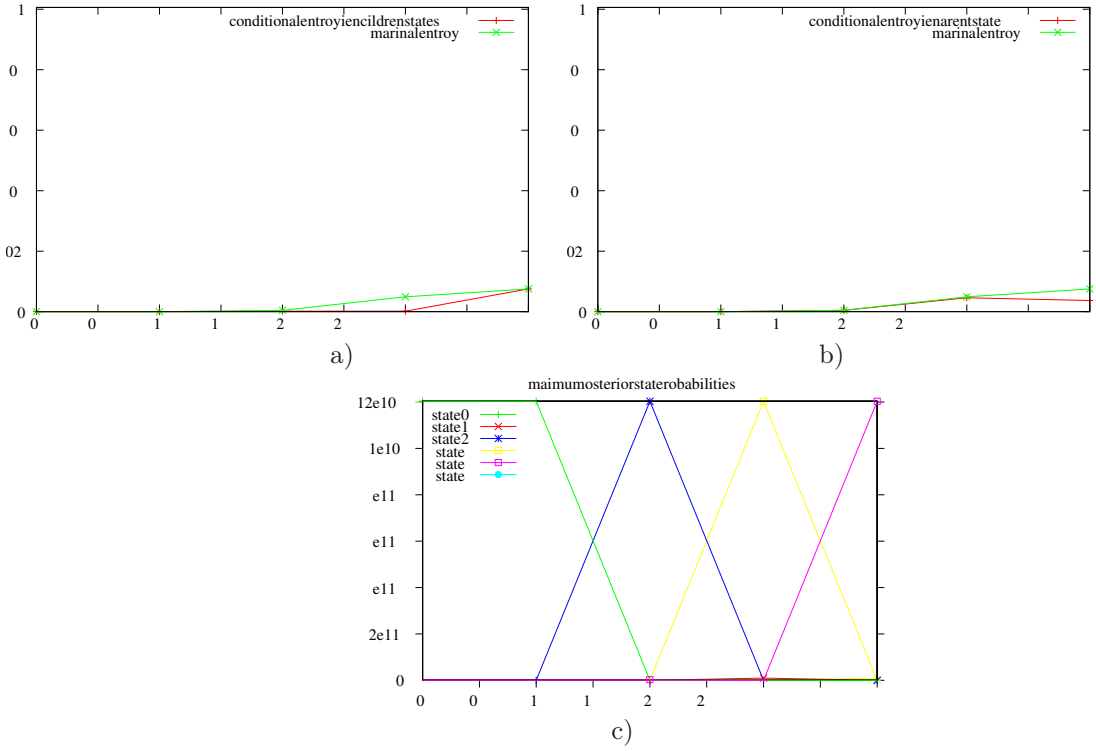


Figure 12: *Entropy profiles along a path containing path with a terminal male shoot. a) Marginal and conditional entropy given children states. b) Marginal and conditional entropy given parent state. c) State tree restoration with the Viterbi upward-downard algorithm.*

4.2.5 Comparison between entropy profiles conditioned on parent or children states

As discussed in Section 3, the following inequality is satisfied, regarding entropy profiles:

$$G(\mathcal{T}) = \sum_{u \in \mathcal{T}} H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}) \leq M(\mathcal{T}) = \sum_{u \in \mathcal{T}} H(S_u|\boldsymbol{X} = \boldsymbol{x}),$$

that is, the global state tree entropy is bounded from above by the sum of marginal entropies.

Let $C(\mathcal{T})$ be defined as

$$C(\mathcal{T}) = \sum_{u \in \mathcal{T}} H(S_u|\boldsymbol{S}_{c(u)}, \boldsymbol{X} = \boldsymbol{x}).$$

On the one hand, we have $C(\mathcal{T}) \leq M(\mathcal{T})$. On the other hand, by Proposition 2, $G(\mathcal{T}) \leq C(\mathcal{T})$. To assess the overestimation of state uncertainty induced by using the profiles based on $H(S_u|\boldsymbol{S}_{c(u)}, \boldsymbol{X} = \boldsymbol{x})$ or $H(S_u|\boldsymbol{X} = \boldsymbol{x})$ instead of $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$, these quantities were

computed for each tree in the dataset, using the 6-state HMT model with the "length" variable given in Section 4.2.4. The ratio $(C(\mathcal{T}) - G(\mathcal{T}))/G(\mathcal{T})$ and $(M(\mathcal{T}) - G(\mathcal{T}))/G(\mathcal{T})$ are given in Table 1.

Tree $\mathcal{T}$ number	$\frac{C(\mathcal{T}) - G(\mathcal{T})}{G(\mathcal{T})}$	$\frac{M(\mathcal{T}) - G(\mathcal{T})}{G(\mathcal{T})}$
1	10.1 %	69.1 %
2	30.9 %	78.0 %
3	22.4 %	76.4 %
4	16.2 %	56.0 %
5	6.5 %	85.2 %
6	19.1 %	73.5 %
7	26.6 %	85.1 %

Table 1: *Comparison between entropy conditioned on parent state, children states, and marginal entropy.* $(C(\mathcal{T}) - G(\mathcal{T}))/G(\mathcal{T})$ *represents the relative distance between conditional entropy given the children states and conditional entropy given the parent state (taken as reference).* $(M(\mathcal{T}) - G(\mathcal{T}))/G(\mathcal{T})$ *represents the relative distance between marginal entropy and conditional entropy given the parent state.*

It can be seen from Table 1 that $C(\mathcal{T})$ is much closer from $G(\mathcal{T})$ than $M(\mathcal{T})$ is. As a consequence, profiles based on $H(S_u|\boldsymbol{S}_{c(u)}, \boldsymbol{X} = \boldsymbol{x})$ provide moderate amplification of the perception of state uncertainty in our example. By contrast, $M(\mathcal{T})$ is a poor approximation of the global state tree entropy. As a consequence, the smoothed probability profiles are irrelevant to quantify uncertainty related to the state tree.

5 Conclusion and discussion

5.1 Concluding remarks

This work illustrates the relevance of using entropy profiles to assess state uncertainty in graphical hidden Markov models. It has been shown that global state entropy can be decomposed additively along the graph structure. In the particular case of HMC and HMT models, we provided algorithms to compute the local contribution of each vertex to this entropy.

Used jointly with the Viterbi algorithm and its variants, these profiles allow deeper understanding on how the model assigns states to vertices – compared to plain Viterbi state restoration and smoothed probability profiles. In particular, these profiles may highlight zones of connected vertices where marginal state uncertainty is not only related to the observed value at each vertex, but where concurrent subtrees are plausible restorations in this zone. Such situations are characterized by high mutual information between neighboring states.

Equivalent algorithms remain to be derived for trees with conditional dependency between children states given parent state (in particular, for trees oriented from the leaf vertices toward the root), and in the case of the DAG structures mentioned in Section 3.1.

5.2 Connexion with model selection

Selection of the number of states In the perspective of model selection, entropy computation can also appear as a valuable tool. If irrelevant states are added to a graphical hidden

Markov model, global state entropy is expected to increase. This principle can be extended to adding irrelevant variables (that is, variables that are independent from the states or conditionally independent from the states given other variables). If the model parameters were known, adding such variables would not change the state conditional distribution. However, since the parameters are estimated from a finite sample, estimation induces perturbations in this conditional distribution in the context of irrelevant variables, and the global state entropy tends to increase. This intuitive statement explains why several model selection criteria based on state entropy were proposed. Among these is the Normalized Entropy Criterion introduced by Celeux & Soromenho (1996) in independent mixture models. It is defined for a mixture with J components as

$$\mathrm{NEC}(J) = \frac{H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})}{\log f_{\hat{\theta}_J}(\boldsymbol{x}) - \log f_{\hat{\theta}_1}(\boldsymbol{x})}$$

if $J > 1$, and has to be minimized. Here, θ_J denotes the parameters of a J-component mixture model, f_{θ_J} its probability density function and $\hat{\theta}_j$ the maximum likelihood estimator of θ_J. Note that $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})$ also depends on $f_{\hat{\theta}_j}$. The number of independent model parameters in θ_J will be denoted by d_J. For $J = 1$, NEC is defined as a ratio between the entropy of a mixture model with different variances and equal proportions and means, and the difference between the log-likelihoods of this model and a model with one component.

The ICL-BIC is also a criterion relying on global state entropy, and must be maximized. It was introduced by McLachlan & Peel (2000, chap. 6) and is defined by

$$\text{ICL-BIC}(J) = 2\log f_{\hat{\theta}_J}(\boldsymbol{x}) - 2H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) - d_J \log(n)$$

where n is the number of vertices in $\boldsymbol{X}$.

Although both criteria were originally defined in the context of independent mixtures, their generalization to graphical hidden Markov models is rather straightforward. By favoring models with small state entropy and high log-likelihood, they aim at selecting models such that the uncertainty of the state values is low, whilst achieving good fit to the data. In practice, they tend to select models with well-separated components in the case of independent mixture models (McLachlan & Peel, 2000, chap. 6).

Criterion	Number of states			
	4	5	6	7
BIC	-10,545	-10,558	**-10,541**	-10,558
NEC	0.48	0.37	**0.32**	0.46
ICL-BIC	-10,764	-10,742	**-10,704**	-10,814

Table 2: *Value of three model selection criteria: BIC, NEC and ICL-BIC, to select the number of states in the Aleppo pines dataset.*

A similar criterion based on minimization of a contrast combining the loglikelihood and state entropy in the context of independent mixture models was proposed by Baudry *et al.* (2008). Selection of the number of mixture components was achieved by a slope heuristic.

Applied to the Allepo pines dataset in Section 4, BIC would assess the 4-state and the 6-state HMT models as nearly equally suited to the dataset, and in practice the modeller could prefer the more parsimonious 4-state model (see Table 2). In contrast, NEC and ICL-BIC would select the 6-state HMT model, since it achieves a better state separation than the 4-state model, for equivalent fit (as assessed by BIC).

Let us note however that the BIC, NEC and ICL-BIC criteria are not suitable for variable selection, since the log-likelihoods of models with different number of variables cannot be compared.

Selection of variables To decide which variables are relevant for the identification of hidden Markov chain or tree models with interpretable states, global state entropy can be regarded as a diagnostic tool. Adding irrelevant variables in the model expectedly leads to increasing the state entropy; consequently, if adding a variable results into a reduction of the state entropy, this variable can be considered as relevant. Moreover, the state space does not depend on the number of observed variables. This makes the values of $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})$ and $H(\boldsymbol{S}|\boldsymbol{Y}=\boldsymbol{y})$ comparable, even if the observed processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ differ by their numbers of variables.

To illustrate this principle, the following experiment was conducted: ten samples of size 836 (same size as the dataset in the application of section 4) were simulated independently, using a Bernoulli distribution with parameter $p = 0.5$. They where also simulated independently from the five other variables described in the application, and were successively added to the Aleppo pines dataset.

After the addition of the i^{th} Bernoulli variable $Y_i = (Y_{i,u})_{1\leq u\leq 836}$, a 6-state HMT model was estimated on the $i+5$-dimensional dataset, and the total state entropy H_i was computed. This procedure was repeated ten times (*i.e.*, samples $(Y_{i,j,u})_{1\leq u\leq 836}$ were simulated for additional variables $i = 1,\ldots,10$ and for replications $j = 1,\ldots,10$). Thus, 10×10 values of $H_{i,j}$ were computed. For a given value of i, the observed variable was a $i+5$-dimensional vector. For $1\leq j\leq 10$, let $H_{0,j} = H_0$ be the state entropy yielded by the 6-state model in Section 4.2.4 using the original dataset. Its value does not depend on j. Only three values in $(H_{i,j})_{1\leq i\leq 10,1\leq j\leq 10}$ were below $H_{0,j}$. To assess the increase in state entropy related to the inclusion of irrelevant variables, the following regression model was considered:

$$H_{i,j} = \alpha i + \beta + \varepsilon_{i,j}$$

where the residuals $(\varepsilon_{i,j})_{i,j}$ were assumed independent and Gaussian with mean 0 and variance σ^2. The test of the null hypothesis $\mathcal{H}_0 : \alpha = 0$ against the alternative $\mathcal{H}_1 : \alpha\in\mathbb{R}$ had *P-value* 10^{-3}. The maximum likelihood estimate of α was $\hat{\alpha} = 3.4$. This result highlights that state entropy significantly increased with the number of additional variables.

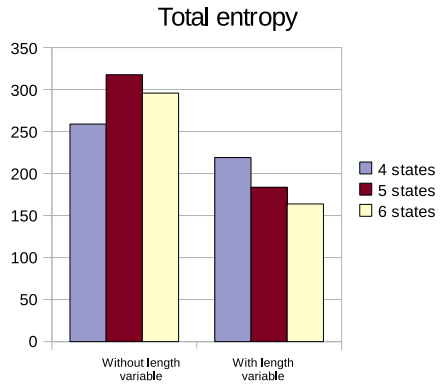


Figure 13: *Global state entropy of the whole forest of state trees, for models with 4 to 6 states, including or not the length variable.*

It can be seen from Figure 13 that global state entropy (computed on the whole forest of state trees) was lowest for the 6-state HMT model including the "length" variable. Combined with Table 2, this figure confirms that this HMT model is the most relevant for the Allepo pine dataset, since the information criteria BIC, NEC and ICL-BIC selected 6-state HMT models, and since removing the "length" variable from this model increased state entropy. This 6-state HMT model also has the most relevant interpretation, as illustrated in Section 4.

This highlights the potential benefit of using entropy-based criteria in model selection for hidden Markov models.

A Proof of propositions

A.1 Algorithms for computing entropy profiles conditioned on the future in the case of hidden Markov chain models

Algo

This algorithm to compute $H(S_{t+1}^{T-1}|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1})$ for $j = 0, \ldots, J-1$ and $t = 0, \ldots, T-1$ is initialized at $t = T-1$ and for $j = 0, \ldots, J-1$ as follows:

$$H\left(S_{T-1}|S_{T-2} = j, X_{T-1} = x_{T-1}\right)$$
$$= -\sum_k P\left(S_{T-1} = k|S_{T-2} = j, X_{T-1} = x_{T-1}\right) \log P\left(S_{T-1} = k|S_{T-2} = j, X_{T-1} = x_{T-1}\right).$$

The backward recursion is achieved, for $t = T-2, \ldots, 0$ and for $j = 0, \ldots, J-1$, using:

$$\begin{aligned}
&H\left(S_{t+1}^{T-1}|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) \\
&= -\sum_{s_{t+1}, \ldots, s_{T-1}} P\left(S_{t+1}^{T-1} = s_{t+1}^{T-1}|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) \log P\left(S_{t+1}^{T-1} = s_{t+1}^{T-1}|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) \\
&= -\sum_{s_{t+2}, \ldots, s_{T-1}} \sum_k P\left(S_{t+2}^{T-1} = s_{t+2}^{T-1}|S_{t+1} = k, S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) P\left(S_{t+1} = k|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) \\
&\times \left\{\log P\left(S_{t+2}^{T-1} = s_{2+1}^{T-1}|S_{t+1} = k, S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) + \log P\left(S_{t+1} = k|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right)\right\} \\
&= -\sum_k P\left(S_{t+1} = k|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) \sum_{s_{t+2}, \ldots, s_{T-1}} P\left(S_{t+2}^{T-1} = s_{t+2}^{T-1}|S_{t+1} = k, X_{t+2}^{T-1} = x_{t+2}^{T-1}\right) \\
&\times \left\{\log P\left(S_{t+2}^{T-1} = s_{2+1}^{T-1}|S_{t+1} = k, X_{t+2}^{T-1} = x_{t+2}^{T-1}\right) + \log P\left(S_{t+1} = k|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right)\right\} \\
&= \sum_k P\left(S_{t+1} = k|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) \left\{H\left(S_{t+2}^{T-1}|S_{t+1} = k, X_{t+2}^{T-1} = x_{t+2}^{T-1}\right)\right. \\
&\quad \left. - \log P\left(S_{t+1} = k|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right)\right\}, \qquad (30)
\end{aligned}$$

with

$$\begin{aligned}
&P\left(S_{t+1} = k|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1}\right) \\
&= \frac{P\left(X_{t+1}^{T-1} = x_{t+1}^{T-1}, S_{t+1} = k|S_t = j\right)}{P\left(X_{t+1}^{T-1} = x_{t+1}^{T-1}|S_t = j\right)} \\
&= \frac{L_{t+1}(k)\, p_{jk}/G_{t+1}(k)}{\sum_m L_{t+1}(m)\, p_{jm}/G_{t+1}(m)}.
\end{aligned}$$

Using a similar argument as in (30), the termination step is given by

$$\begin{aligned}
&H\left(S_0^{T-1}|\boldsymbol{X} = \boldsymbol{x}\right) \\
&= -\sum_j P\left(S_0 = j|\boldsymbol{X} = \boldsymbol{x}\right) \left\{\sum_{s_1, \ldots, s_{T-1}} P\left(S_1^{T-1} = s_1^{T-1}|S_0 = j, X_1^{T-1} = x_1^{T-1}\right)\right. \\
&\left. \times \log P\left(S_1^{T-1} = s_1^{T-1}|S_0 = j, X_1^{T-1} = x_1^{T-1}\right) + \log P\left(S_0 = j|\boldsymbol{X} = \boldsymbol{x}\right)\right\} \\
&= \sum_j L_0(j) \left\{H\left(S_1^{T-1} = s_1^{T-1}|S_0 = j, X_1^{T-1} = x_1^{T-1}\right) - \log L_0(j)\right\}.
\end{aligned}$$

A.2 Entropy profiles conditioned on the children states for hidden Markov tree models

A proof of Proposition 2 is given, in the case of binary trees for the sake of simplicity.

Proof Let $lc(u)$ and $rc(u)$ denote the two children of vertex u. Applying the chain rule on the children of the root vertex, we can write

$$\begin{aligned} H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) = H\left(S_0|\bar{\boldsymbol{S}}_{c(0)}, \boldsymbol{X}=\boldsymbol{x}\right) + H\left(S_{lc(0)}|\bar{\boldsymbol{S}}_{c(lc(0))}, \bar{\boldsymbol{S}}_{rc(0)}, \boldsymbol{X}=\boldsymbol{x}\right) \\ + H\left(S_{rc(0)}|\bar{\boldsymbol{S}}_{c(lc(0))}, \bar{\boldsymbol{S}}_{c(rc(0))}, \boldsymbol{X}=\boldsymbol{x}\right) + H\left(\bar{\boldsymbol{S}}_{c(lc(0))}, \bar{\boldsymbol{S}}_{c(rc(0))}|\boldsymbol{X}=\boldsymbol{x}\right). \end{aligned}$$

This decomposition is indeed not unique and we can choose to extract the conditional entropy corresponding to $rc(0)$ before the conditional entropy corresponding to $lc(0)$. Applying the property that deconditioning augments entropy (Cover & Thomas, 2006, chap. 2)

$$\begin{aligned} H\left(S_{lc(0)}|\bar{\boldsymbol{S}}_{c(lc(0))}, \bar{\boldsymbol{S}}_{rc(0)}, \boldsymbol{X}=\boldsymbol{x}\right) \leq H\left(S_{lc(0)}|\bar{\boldsymbol{S}}_{c(lc(0))}, \boldsymbol{X}=\boldsymbol{x}\right), \\ H\left(S_{rc(0)}|\bar{\boldsymbol{S}}_{c(lc(0))}, \bar{\boldsymbol{S}}_{c(rc(0))}, \boldsymbol{X}=\boldsymbol{x}\right) \leq H\left(S_{rc(0)}|\bar{\boldsymbol{S}}_{c(rc(0))}, \boldsymbol{X}=\boldsymbol{x}\right), \end{aligned}$$

we obtain

$$\begin{aligned} H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) \leq H\left(S_0|\bar{\boldsymbol{S}}_{c(0)}, \boldsymbol{X}=\boldsymbol{x}\right) + H\left(S_{lc(0)}|\bar{\boldsymbol{S}}_{c(lc(0))}, \boldsymbol{X}=\boldsymbol{x}\right) \\ + H\left(S_{rc(0)}|\bar{\boldsymbol{S}}_{c(rc(0))}, \boldsymbol{X}=\boldsymbol{x}\right) + H\left(\bar{\boldsymbol{S}}_{c(lc(0))}, \bar{\boldsymbol{S}}_{c(rc(0))}|\boldsymbol{X}=\boldsymbol{x}\right). \end{aligned}$$

Applying the same decomposition recursively from the root to the leaves and upper bounding on each internal vertex completes the proof by induction. ■

References

[Baudry et al.(2008)Baudry, Celeux, and Marin] J.-P. Baudry, G. Celeux, and J.-M. Marin. Selecting Models Focussing on the Modeller's Purpose. In Paula Brito, editor, *Compstat2008. Porto (Portugal)*. Physica-Verlag, August 2008.

[Boucheron and Gassiat(2007)] S. Boucheron and E. Gassiat. *An information-theoretic perspective on order estimation*, pages 565–602. O. Cappé, E. Moulines and T. Rydén, Springer, New York, 2007.

[Brushe et al.(1998)Brushe, Mahony, and Moore] G.D. Brushe, R.E. Mahony, and J.B. Moore. A Soft Output Hybrid Algorithm for ML/MAP Sequence Estimation. *IEEE Transactions on Information Theory*, 44(7):3129–3134, November 1998.

[Cadre and Tremois(1998)] J.-P. Le Cadre and O. Tremois. Bearing-Only Tracking for Maneuvering Sources. *IEEE Transactions on Aerospace and Electronic Systems*, 34(1):179–193, January 1998.

[Cappé et al.(2005)Cappé, Moulines, and Rydén] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. New York: Springer, 2005.

[Celeux and Durand(2008)] G. Celeux and J.-B. Durand. Selecting Hidden Markov Model State Number with Cross-Validated Likelihood. *Computational Statistics*, 23:541–564, 2008.

[Celeux and Soromenho(1996)] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

[Ciriza et al.(2011)Ciriza, Donini, Durand, and Girard] V. Ciriza, L. Donini, J.-B. Durand, and S. Girard. Optimal timeouts for power management under renewal or hidden Markov processes for requests. Submitted to Applied Statistics. Also in technical report, 2011. URL `http://hal.inria.fr/hal-00412509/en`.

[Cover and Thomas(2006)] T.M. Cover and J.A. Thomas. *Elements of Information Theory, 2nd edition*. Hoboken, NJ: Wiley, 2006.

[Crouse et al.(1998)Crouse, Nowak, and Baraniuk] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Transactions on Signal Processing*, 46(4):886–902, April 1998.

[Devijver(1985)] P. A. Devijver. Baum's forward-backward Algorithm Revisited. *Pattern Recognition Letters*, 3:369–373, 1985.

[Durand et al.(2004)Durand, Gonçalvès, and Guédon] J.-B. Durand, P. Gonçalvès, and Y. Guédon. Computational methods for hidden Markov tree models – an application to wavelet trees. *IEEE Transactions on Signal Processing*, 52(9):2551–2560, September 2004.

[Ephraim and Merhav(2002)] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48:1518–1569, June 2002.

[Guédon(2007)] Y. Guédon. Exploring the state sequence space for hidden Markov and semi-Markov chains. *Computational Statistics & Data Analysis*, 51(5):2379–2409, 2007.

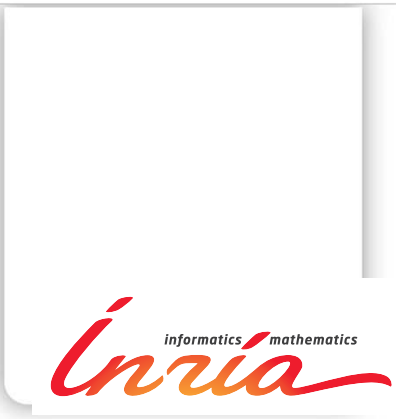
[Hernando et al.(2005)Hernando, Crespi, and Cybenko] D. Hernando, V. Crespi, and G. Cybenko. Efficient computation of the hidden Markov model entropy for a given observation sequence. *IEEE Transactions on Information Theory*, 51(7):2681–2685, July 2005.

[Lauritzen(1996)] S.L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, United Kingdom, 1996.

[McLachlan and Peel(2000)] G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley and Sons, 2000.

[Zucchini and MacDonald(2009)] W. Zucchini and I.L. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC: Boca Raton FL, 2009.

Contents



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399