Localizing the latent structure canonical uncertainty: entropy profiles for hidden Markov models

Jean-Baptiste Durand, Yann Guédon

Statistics and Computing manuscript No.
(will be inserted by the editor)

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

Jean-Baptiste Durand · Yann Guédon

Received: date / Accepted: date

Abstract This paper 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 multidimensional posterior state (or smoothed) probability profiles because of the marginalization that is intrinsic in the computation of these posterior probabilities. Here, we introduce a new type of profiles that 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 localise this uncertainty, (ii) these profiles are unidimensional 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. The use of entropy profiles is illustrated by sequence and tree data examples.

Keywords Conditional Entropy · Hidden Markov Chain Model · Hidden Markov Tree Model · Plant Structure Analysis

1 Introduction

Hidden Markov chain (HMC) models have been widely used in signal processing and pattern recognition, for the analysis of sequences with various types of

J.-B. Durand
Univ. Grenoble Alpes, Laboratoire Jean Kuntzmann and Inria, Mistis
51 rue des Mathématiques
B.P. 53, F-38041 Grenoble cedex 9, France Tel.: +33 4 76 63 57 09
Fax: +33 4 76 63 12 63
E-mail: jean-baptiste.durand@imag.fr

Y. Guédon
CIRAD, UMR AGAP and Inria, Virtual Plants
F-34095 Montpellier, France E-mail: guedon@cirad.fr

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 in the HMC case.

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 Durand *et al.* (2013) proposed to optimise 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 an HMC 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 a critical step of the analysis.

Not only is state restoration essential for model interpretation, it is generally used for model diagnostic and validation as well, for example by visualising some 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 an HMC 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 HMC model case, but the same reasoning applies to other families of latent structure models, including hidden semi-Markov models and hidden Markov tree (HMT) models.

Let $\boldsymbol{S} = (S_t)_{t=0,1,\ldots}$ denote the finite state process and J the number of states of an HMC model. Let $\boldsymbol{X} = (X_t)_{t=0,1,\ldots}$ denote its output (or *observed*) process, which takes value in an arbitrary set (countable or uncountable, uni- or multidimensional, ...). To simplify notations and without loss of generality – since this work focuses on conditional distributions of states given the outputs – we will use $P(X_t = x_t)$ as if X_t was a discrete random variable. Methods for exploring the state sequences that explain a given observed sequence $\boldsymbol{X} = \boldsymbol{x}$ for a known HMC model may be divided into three categories: (i) enumeration of state sequences, (ii) state profiles, which are state sequences summarised in a $J \times T$ array where T is 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 HMC model was proposed as a global measure of the state sequence uncertainty by Hernando *et al.* (2005). We assume here that this conditional entropy $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ 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 (2007b) and references therein. We identified some shortcomings of the proposed methods:

– The entropy of the state sequence is not a summary of the state profiles based on the posterior state (or smoothed) probabilities $\{P(S_t = j|\boldsymbol{X} = \boldsymbol{x})\}_{t=0,\ldots,T-1;j=0,\ldots,J-1}$, because of the marginalization that is intrinsic in the computation of these probabilities. We illustrate using examples the fact that these state profiles contain artifacts, that introduce confusion between (i) local state uncertainty due to overlap at $X_t = x_t$ of emission distributions for different states and (ii) mere propagation of uncertainty from past states to current state S_t.
– Due to their multidimensional nature, state profiles are difficult to visualise and interpret on trees except in the case of two-state models.

Our objective is to overcome these shortcomings, by proposing profiles satisfying the following properties:

(a) These profiles result from an additive decomposition of the the canonical measure of state sequence uncertainty along the sequence.
(b) Each term of the decomposition can be interpreted as a local contribution to the global state sequence uncertainty, and thus corresponds to a local state uncertainty.
(c) The profiles are unidimensional, and thus can be extended to more general supporting structures such as directed acyclic graphs (DAGs), and in particular, trees.

In the case of a hidden first-order Markov chain model, applying the chain rule (Cover and Thomas, 2006; Chapter 2) and the Markovian property, the global state sequence entropy can be decomposed as a conditional entropy profile and thus localised along the sequence

$$\begin{aligned} H(\mathbf{S}|\mathbf{X} = \mathbf{x}) &= H(S_0|\mathbf{X} = \mathbf{x}) + \sum_{t=1}^{T-1} H(S_t|S_{t-1}, \ldots, S_0, \mathbf{X} = \mathbf{x}) \\ &= H(S_0|\mathbf{X} = \mathbf{x}) + \sum_{t=1}^{T-1} H(S_t|S_{t-1}, \mathbf{X} = \mathbf{x}). \end{aligned} \quad (1)$$

The conditional entropy profile $\{H(S_t|S_{t-1}, \mathbf{X} = \mathbf{x})\}_{t=0,\ldots,T-1}$ benefits from all the properties of the entropy function: unique function satisfying the Shannon-Khinchin axioms, interpretation of conditional entropy as expected value of the entropies of the conditional distributions, averaged over the conditioning random variables (Cover and Thomas, 2006). We show that the posterior state probability profiles $\{P(S_t = j|\mathbf{X} = \mathbf{x})\}_{t=0,\ldots,T-1;j=0,\ldots,J-1}$ can be summarised as a marginal entropy profile $\{H(S_t|\mathbf{X} = \mathbf{x})\}_{t=0,\ldots,T-1}$ with $H(S_t|S_{t-1}, \mathbf{X} = \mathbf{x}) \leq H(S_t|\mathbf{X} = \mathbf{x})$ for $t = 0, \ldots, T-1$. Contrary to conditional entropy profiles (which can be referred to as hidden Markov entropy profiles), marginal entropy profiles do not reflect the state sequence uncertainty, as deduced from the conditional independence structure of the HMC model. We show using examples that marginal entropy profiles and consequently posterior state probability profiles give erroneous diagnostics and should not be used for localising latent state structure uncertainty. It should be noted that a similar approach, based on conditional entropy profiles, has been proposed by Guédon (2013) for localizing the segmentation uncertainty along a sequence in the case of a multiple change-point model.

One of the outcomes of this work is to derive efficient algorithms to compute the conditional entropy profile for HMC models. This approach is extended to

HMT models; in this case, the conditional entropy profile is used in a first stage to identify zones with high local contributions to global uncertainty. In a second stage, state profiles computed by the Viterbi algorithm and its variants (Durand *et al.*, 2004), or an adaption to trees of the forward-backward Viterbi algorithm of Brushe *et al.* (1998), are visualised on selected paths of interest within the state process. This allows for identification of alternative states at positions with ambiguous state value, and for better insight on how the model associates the states with observed data.

The remainder of this paper is organized as follows. Section 2 focuses on algorithms to compute conditional entropy profiles for HMC models. 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 conditional entropy profiles are derived in the case of HMT 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, HMC models are first defined. Then the classic forward-backward algorithm and the algorithm of Hernando *et al.* (2005) to compute the entropy of the state sequence that explains an observed sequence are presented. These algorithms form the basis of the proposed method to compute conditional entropy profiles (i.e., decomposition of 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),

and where for any $(s_t, x_t)_{t=0,1,\ldots,T-1}$

$$\begin{aligned} &P(X_0 = x_0, \ldots, X_{T-1} = x_{T-1} | S_0 = s_0, \ldots, S_{T-1} = s_{T-1}) \\ &= \prod_{t=0}^{T-1} P(X_t = x_t | S_t = s_t). \end{aligned}$$

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 $\boldsymbol{X}$.

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 denoted $\boldsymbol{X} = \boldsymbol{x}$ and $S_0^{T-1} = s_0^{T-1}$ is denoted $\boldsymbol{S} = \boldsymbol{s}$. 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 for computing the entropy of the state sequence that explains an observed sequence

The forward-backward algorithm aims at computing the posterior state (or 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 initialised at $t = 0$ and for $j = 0, \ldots, J-1$ as follows:

$$\begin{aligned} F_0(j) &= P(S_0 = j|X_0 = x_0) \\ &= \frac{b_j(x_0)}{N_0}\pi_j. \end{aligned} \tag{2}$$

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

$$\begin{aligned} F_t(j) &= P(S_t = j|X_0^t = x_0^t) \\ &= \frac{b_j(x_t)}{N_t}\sum_{i=0}^{J-1} p_{ij}F_{t-1}(i). \end{aligned} \tag{3}$$

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

$$\begin{aligned} N_t &= P(X_t = x_t|X_0^{t-1} = x_0^{t-1}) \\ &= \sum_{j=0}^{J-1} P(S_t = j, X_t = x_t|X_0^{t-1} = x_0^{t-1}), \end{aligned}$$

with

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

and

$$P(S_t = j, X_t = x_t|X_0^{t-1} = x_0^{t-1}) = b_j(x_t)\sum_{i=0}^{J-1} p_{ij}F_{t-1}(i).$$

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

$$L_{T-1}(j) = P(S_{T-1} = j|\boldsymbol{X} = \boldsymbol{x}) = F_{T-1}(j). \tag{4}$$

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

$$\begin{aligned} L_t(j) &= P(S_t = j|\boldsymbol{X} = \boldsymbol{x}) \\ &= \left\{\sum_{k=0}^{J-1} \frac{L_{t+1}(k)}{G_{t+1}(k)}p_{jk}\right\} F_t(j), \end{aligned} \tag{5}$$

where the predicted probability $G_{t+1}(k)$ is directly deduced from the filtered probabilities $F_t(j)$ for the different states j

$$\begin{aligned} G_{t+1}(k) &= P(S_{t+1} = k | X_0^t = x_0^t) \\ &= \sum_{j=0}^{J-1} p_{jk} F_t(j). \end{aligned}$$

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 (2) and (3) as a building block. It requires a forward recursion on entropies of partial state sequences S_0^t.
This algorithm is initialised at $t = 1$ and for $j = 0, \ldots, J-1$ as follows:

$$\begin{aligned} &H(S_0 | S_1 = j, X_0^1 = x_0^1) \\ &\quad = -\sum_{i=0}^{J-1} P(S_0 = i | S_1 = j, X_0^1 = x_0^1) \log P(S_0 = i | S_1 = j, X_0^1 = x_0^1). \end{aligned} \quad (6)$$

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

$$\begin{aligned} &H(S_0^{t-1} | S_t = j, X_0^t = x_0^t) \\ &= -\sum_{s_0, \ldots, s_{t-1}} P(S_0^{t-1} = s_0^{t-1} | S_t = j, X_0^t = x_0^t) \log P(S_0^{t-1} = s_0^{t-1} | S_t = j, X_0^t = x_0^t) \\ &= -\sum_{s_0, \ldots, s_{t-2}} \sum_{i=0}^{J-1} P(S_0^{t-2} = s_0^{t-2} | S_{t-1} = i, S_t = j, X_0^t = x_0^t) P(S_{t-1} = i | S_t = j, X_0^t = x_0^t) \\ &\times \left\{ \log P(S_0^{t-2} = s_0^{t-2} | S_{t-1} = i, S_t = j, X_0^t = x_0^t) + \log P(S_{t-1} = i | S_t = j, X_0^t = x_0^t) \right\} \\ &= -\sum_{i=0}^{J-1} P(S_{t-1} = i | S_t = j, X_0^{t-1} = x_0^{t-1}) \left\{ \sum_{s_0, \ldots, s_{t-2}} P(S_0^{t-2} = s_0^{t-2} | S_{t-1} = i, X_0^{t-1} = x_0^{t-1}) \right. \\ &\left. \times \log P(S_0^{t-2} = s_0^{t-2} | S_{t-1} = i, X_0^{t-1} = x_0^{t-1}) + \log P(S_{t-1} = i | S_t = j, X_0^t = x_0^t) \right\} \\ &= \sum_{i=0}^{J-1} P(S_{t-1} = i | S_t = j, X_0^{t-1} = x_0^{t-1}) \left\{ H(S_0^{t-2} | S_{t-1} = i, X_0^{t-1} = x_0^{t-1}) \right. \\ &\left. - \log P(S_{t-1} = i | S_t = j, X_0^{t-1} = x_0^{t-1}) \right\}, \end{aligned} \quad (7)$$

with

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

Using a similar argument as in (7), the termination step is given by

$$
\begin{aligned}
&H(S_0^{T-1}|\boldsymbol{X}=\boldsymbol{x}) \\
&= -\sum_{j=0}^{J-1} P(S_{T-1}=j|\boldsymbol{X}=\boldsymbol{x}) \left\{ \sum_{s_0,\ldots,s_{T-2}} P(S_0^{T-2}=s_0^{T-2}|S_{T-1}=j,\boldsymbol{X}=\boldsymbol{x}) \right. \\
&\left. \times\ \log P(S_0^{T-2}=s_0^{T-2}|S_{T-1}=j,\boldsymbol{X}=\boldsymbol{x}) + \log P(S_{T-1}=j|\boldsymbol{X}=\boldsymbol{x}) \right\} \\
&= \sum_{j=0}^{J-1} F_{T-1}(j) \left\{ H(S_0^{T-2}|S_{T-1}=j,\boldsymbol{X}=\boldsymbol{x}) - \log F_{T-1}(j) \right\}. \qquad (8)
\end{aligned}
$$

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

Remark. The forward recursion (7) can be interpreted as the chain rule

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

with

$$
\begin{aligned}
&H(S_0^{t-2}|S_{t-1},S_t=j,X_0^t=x_0^t) \\
&= \sum_{i=0}^{J-1} P(S_{t-1}=i|S_t=j,X_0^{t-1}=x_0^{t-1}) H(S_0^{t-2}|S_{t-1}=i,X_0^{t-1}=x_0^{t-1})
\end{aligned}
$$

and

$$
\begin{aligned}
&H(S_{t-1}|S_t=j,X_0^t=x_0^t) \\
&= -\sum_{i=0}^{J-1} P(S_{t-1}=i|S_t=j,X_0^{t-1}=x_0^{t-1}) \log P(S_{t-1}=i|S_t=j,X_0^{t-1}=x_0^{t-1}).
\end{aligned}
$$

2.3 Algorithms for computing conditional entropy profiles for hidden Markov chain models

In what follows, we derive algorithms to compute conditional entropy profiles $\{H(S_t|S_{t-1},\boldsymbol{X}=\boldsymbol{x})\}_{t=0,\ldots,T-1}$. As a byproduct, the global state sequence entropy $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})$ can be directly extracted.

We propose a first solution where the partial state sequence entropies $\left\{H(S_0^t|\boldsymbol{X}=\boldsymbol{x})\right\}_{t=0,\ldots,T-1}$ 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.

For $t = 0, \ldots, T-1$, we have

$$\begin{aligned}
&H(S_0^t|\boldsymbol{X}=\boldsymbol{x}) \\
&= -\sum_{s_0,\ldots,s_t} P(S_0^t = s_0^t|\boldsymbol{X}=\boldsymbol{x}) \log P(S_0^t = s_0^t|\boldsymbol{X}=\boldsymbol{x}) \\
&= -\sum_{j=0}^{J-1} P(S_t = j|\boldsymbol{X}=\boldsymbol{x}) \left\{ \sum_{s_0,\ldots,s_{t-1}} P(S_0^{t-1} = s_0^{t-1}|S_t = j, X_0^t = x_0^t) \right. \\
&\quad \left. \times \log P(S_0^{t-1} = s_0^{t-1}|S_t = j, X_0^t = x_0^t) + \log P(S_t = j|\boldsymbol{X}=\boldsymbol{x}) \right\} \\
&= \sum_{j=0}^{J-1} L_t(j) \left\{ H(S_0^{t-1}|S_t = j, X_0^t = x_0^t) - \log L_t(j) \right\}. \qquad (9)
\end{aligned}$$

This recursion relies on the relation

$$P(S_0^{t-1} = s_0^{t-1}|S_t = j, \boldsymbol{X}=\boldsymbol{x}) = P(S_0^{t-1} = s_0^{t-1}|S_t = j, X_0^t = x_0^t)$$

due to the time-reversed process of $(S_t, X_t)_{t=0,1,\ldots}$ being also a hidden first-order Markov chain.

In this way, the partial state sequence entropies $\left\{H(S_0^t|\boldsymbol{X}=\boldsymbol{x})\right\}_{t=0,\ldots,T-1}$ can be computed as a byproduct of the forward-backward algorithm where the usual forward recursion (3) and the recursion (7) proposed by Hernando *et al.* (2005) are mixed. The conditional entropies are then directly deduced by first-order differencing

$$\begin{aligned}
H(S_t|S_{t-1}, \boldsymbol{X}=\boldsymbol{x}) &= H(S_t|S_0^{t-1}, \boldsymbol{X}=\boldsymbol{x}) \\
&= H(S_0^t|\boldsymbol{X}=\boldsymbol{x}) - H(S_0^{t-1}|\boldsymbol{X}=\boldsymbol{x}). \qquad (10)
\end{aligned}$$

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

$$\begin{aligned}
&H(S_t|S_{t-1}, \boldsymbol{X}=\boldsymbol{x}) \\
&= -\sum_{i,j=0}^{J-1} P(S_t = j, S_{t-1} = i|\boldsymbol{X}=\boldsymbol{x}) \log P(S_t = j|S_{t-1} = i, \boldsymbol{X}=\boldsymbol{x}) \qquad (11)
\end{aligned}$$

with

$$\begin{cases} P(S_t = j|S_{t-1} = i, \boldsymbol{X}=\boldsymbol{x}) = L_t(j) p_{ij} F_{t-1}(i) / \{G_t(j) L_{t-1}(i)\} \text{ and} \\ P(S_t = j, S_{t-1} = i|\boldsymbol{X}=\boldsymbol{x}) = L_t(j) p_{ij} F_{t-1}(i) / G_t(j). \end{cases} \qquad (12)$$

These latter quantities are directly extracted during the backward recursion (5) of the forward-backward algorithm.

In summary, a first possibility is to compute the partial state sequence entropies $\left\{H(S_0^t|\boldsymbol{X}=\boldsymbol{x})\right\}_{t=0,\ldots,T-1}$ using the usual forward and backward recursions combined with (6), (7) and (9), 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 (10). A second possibility is to compute the profile of conditional entropies directly using the usual forward and backward recursions combined with (11) and

to deduce global state sequence entropy 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):

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

with

$$\begin{aligned} H(S_t|\boldsymbol{X} = \boldsymbol{x}) &= -\sum_{j=0}^{J-1} P(S_t = j|\boldsymbol{X} = \boldsymbol{x}) \log P(S_t = j|\boldsymbol{X} = \boldsymbol{x}) \\ &= -\sum_{j=0}^{J-1} L_t(j) \log L_t(j). \end{aligned}$$

The difference between the marginal and the conditional entropy is the mutual information $I(S_t; S_{t-1}|\boldsymbol{X} = \boldsymbol{x})$ 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 marginal entropy profile can be interpreted as a summary of the classic multidimensional posterior state probability profiles $\{P(S_t = j|\boldsymbol{X} = \boldsymbol{x})\}_{t=0,\ldots,T-1;j=0,\ldots,J-1}$. Hence, approximating the global state sequence uncertainty using the sum of marginal entropies $\sum_{t=0}^{T-1} H(S_t|\boldsymbol{X} = \boldsymbol{x})$ would result into the approximation error $\sum_{t=0}^{T-1}\{H(S_t|\boldsymbol{X} = \boldsymbol{x}) - H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x})\}$.

3 Entropy profiles for hidden Markov tree models

In this section, 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 fixed (i.e. non-random), finite or infinite directed acyclic graph (DAG) with vertex set $\mathcal{U}$. A GHM model is a probabilistic model for observed random variables $\boldsymbol{X} = (X_u)_{u\in\mathcal{U}}$ indexed by $\mathcal{U}$. The distribution of $\boldsymbol{X}$ depends on a hidden (i.e. unobserved) J-state process $\boldsymbol{S} = (S_u)_{u\in\mathcal{U}}$ 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 $\mathcal{U}$). 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 $(S_u)_{u\in E}$ (set of variables in $\boldsymbol{S}$ which index belongs to E). GHM models assume that the following factorization of $P_{\boldsymbol{S}}$ – associated with the local Markov property on $\mathcal{G}$, see Lauritzen (1996) – holds for any $\boldsymbol{s}$:

$$P(\boldsymbol{S} = \boldsymbol{s}) = \prod_{u\in\mathcal{U}} P(S_u = s_u|\boldsymbol{S}_{\mathrm{pa}(u)} = \boldsymbol{s}_{\mathrm{pa}(u)}),$$

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 $\mathrm{pa}(u) = \emptyset$. The local Markov property on $\mathcal{G}$ states that conditionally on $\boldsymbol{S}_{\mathrm{pa}(u)} = \boldsymbol{s}_{\mathrm{pa}(u)}$,

S_u does not depend on S_v for every vertex v such that there is no path from u to v (v is not a descendant of u). The output (or *observed*) process $\boldsymbol{X} = (X_u)_{u\in\mathcal{U}}$ is 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. Moreover, it is assumed that given $S_u = j$, $X_u = x$ has probabiliy $b_j(x)$. Process $\boldsymbol{X}$ is referred to as a GHM model with respect to DAG $\mathcal{G}$.

In the particular case where $\mathcal{G}$ is a rooted tree graph, $\boldsymbol{X}$ is called a hidden Markov out-tree model with conditionally-independent children states, given their parent state (or more shortly, an HMT 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 subtree rooted at vertex u, $\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u$ denote the observed 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), $\bar{\boldsymbol{X}}_{v\setminus u} = \bar{\boldsymbol{x}}_{v\setminus u}$ the subtree $\bar{\boldsymbol{x}}_v$ except the subtree $\bar{\boldsymbol{x}}_u$ (assuming that $\bar{\boldsymbol{x}}_u$ is a proper subtree of $\bar{\boldsymbol{x}}_v$), and finally $\bar{\boldsymbol{X}}_{b(u)} = \bar{\boldsymbol{x}}_{b(u)}$ the family of brother subtrees $(\bar{\boldsymbol{X}}_v)_{v\in c(\rho(u));v\neq u}$ of u (assuming that u is not the root vertex). Let $\mathcal{V}$ be a subtree of $\mathcal{T}$ and let $\bar{\boldsymbol{X}}_{\mathcal{V}} = \bar{\boldsymbol{x}}_{\mathcal{V}}$ denote the process $(X_u)_{u\in\mathcal{V}}$, i.e., the observed subtree indexed by $\mathcal{V}$. 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 (from Durand *et al.*, 2004).

Let $\mathcal{T}$ be a fixed tree with vertex set $\mathcal{U}$. A J-state HMT model $(\boldsymbol{S},\boldsymbol{X}) = (S_u,X_u)_{u\in\mathcal{U}}$ on $\mathcal{T}$ is defined by the following parameters:

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

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

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

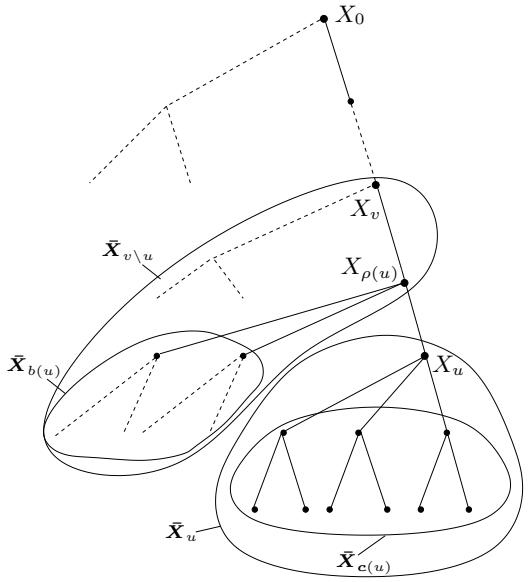

Fig. 1 *The notations used for indexing subtrees of observed tree* $\boldsymbol{X} = \bar{\boldsymbol{X}}_0$*. These notations transpose to the state tree* $\boldsymbol{S} = \bar{\boldsymbol{S}}_0$*.*

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 $\{(S_u, X_u) | u \in \mathcal{U}\}$.

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

$$P(\boldsymbol{S}_A = \boldsymbol{s}_A | \boldsymbol{X} = \boldsymbol{x}) = \prod_{v\in A} P(S_v = s_v | \boldsymbol{S}_{\mathrm{pa}(v)} = \boldsymbol{s}_{\mathrm{pa}(v)}, \boldsymbol{X} = \boldsymbol{x}). \tag{13}$$

Since the joint distribution of state vertices in different connected components $(\mathcal{G}_1, \ldots, \mathcal{G}_C)$ of $\mathcal{G}$ can be factorised as $\prod_c P(\boldsymbol{S}_{\mathcal{G}_c} = \boldsymbol{s}_{\mathcal{G}_c} | \boldsymbol{X} = \boldsymbol{x})$, equation (13) is proved separately for each connected component.

It is easily seen that if u is a source vertex of $\mathcal{G}$ (vertex without parents), both the right-hand and the left-hand sides of equation (13) 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 (13). ■

The following corollary is derived from Proposition 1:

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 $pa(u)=\emptyset$.

Proof This corollary results from

$$\begin{aligned} H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) =& E\left[\log P(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})|\boldsymbol{X}=\boldsymbol{x}\right] \\ =& \sum_u E\left[\log P(S_u|\boldsymbol{S}_{\mathrm{pa}(u)}, \boldsymbol{X}=\boldsymbol{x})|\boldsymbol{X}=\boldsymbol{x}\right] \\ =& \sum_u H(S_u|\boldsymbol{S}_{\mathrm{pa}(u)}, \boldsymbol{X}=\boldsymbol{x}).\blacksquare \end{aligned}$$

This result extends equation (1) 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 parents' states (or equivalently, given the non-descendant states, from the Markov property on $\mathcal{G}(\boldsymbol{S}, \boldsymbol{X})$). In practical applications of GHM models, the unidimensional profiles $\left\{H(S_u|\boldsymbol{S}_{\mathrm{pa}(u)}, \boldsymbol{X}=\boldsymbol{x})\right\}_{u\in\mathcal{U}}$ are computed in a first stage, to identify vertices in $\mathcal{U}$ associated with high local contributions to global uncertainty. In a second stage, the Viterbi algorithm and its variants are used to investigate alternative state restorations.

The remainder of this Section focuses on the derivation of algorithms to compute entropy profiles efficiently in HMT models.

3.2 Reminder: upward-downward algorithm

The upward-downward algorithm aims at computing the posterior state 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 recursion such that every vertex must be visited after its parent vertex (referred to as *downward recursion*). This recursion is initialised at the root vertex $u = 0$ and for $j = 0, \ldots, J-1$ as follows:

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

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

$$P(S_u = j) = \sum_{i=0}^{J-1} p_{ij} P(S_{\rho(u)} = i).$$

In the *upward recursion*,every vertex must be visited after its children vertices. It is initialised 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 given, for internal vertices u taken upwards and for $j = 0, \ldots, J-1$, by:

$$\begin{aligned}\beta_{\rho(u),u}(j) =& \frac{P(\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u | S_{\rho(u)} = j)}{P(\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)} \\ =& \sum_{k=0}^{J-1} \frac{\beta_u(k) p_{jk}}{P(S_u = k)}\end{aligned}$$

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

$$N_u = P(X_u = x_u) = \sum_{j=0}^{J-1} b_j(x_u) P(S_u = j)$$

for the leaf vertices, and

$$N_u = \frac{P(\bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)}{\prod_{v \in \boldsymbol{c}(u)} P(\bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v)} = \sum_{j=0}^{J-1} \left\{ \prod_{v \in \boldsymbol{c}(u)} \beta_{u,v}(j) \right\} b_j(x_u) P(S_u = j)$$

for the internal vertices.
The downward recursion is initialised 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 given, for vertices $u \neq 0$ taken downwards and for $j = 0, \ldots, J-1$, by:

$$\begin{aligned}\xi_u(j) =& P(S_u = j | \boldsymbol{X} = \boldsymbol{x}) \\ =& \frac{\beta_u(j)}{P(S_u = j)} \sum_{i=0}^{J-1} \frac{p_{ij} \xi_{\rho(u)}(i)}{\beta_{\rho(u),u}(i)}.\end{aligned} \tag{14}$$

3.3 Algorithms for computing conditional entropy profiles for hidden Markov tree models

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

$$H(\boldsymbol{S} | \boldsymbol{X} = \boldsymbol{x}) = H(S_0 | \boldsymbol{X} = \boldsymbol{x}) + \sum_{u \neq 0} H(S_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}). \tag{15}$$

This decomposition can be extended to the entropy of any state subtree, as shown in Proposition A1 in Appendix A. As a consequence of this decomposition, the canonical measure of state tree uncertainty can be localised along the observed tree, $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ representing the local contribution at vertex u to the global state tree entropy. Note that this decomposition generalises to HMT models decomposition (1) of HMC models.

We propose a first approach where conditional entropy profiles $\{H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})\}_{u\in\mathcal{U}}$ are directly computed during the downward recursion (14). As a byproduct, the global state tree entropy $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ is obtained by summation. Then we propose a second approach based on the direct computation of the conditional entropies of state subtrees. The conditional entropy profiles are deduced from the latter.

Direct computation of conditional entropy profiles Firstly, for every non-root vertex u, the conditional entropy

$$\begin{aligned} &H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}) \\ &= -\sum_{i,j} P(S_u = j, S_{\rho(u)} = i|\boldsymbol{X} = \boldsymbol{x}) \log P(S_u = j|S_{\rho(u)} = i, \boldsymbol{X} = \boldsymbol{x}) \end{aligned} \tag{16}$$

is directly extracted during the downward recursion (14), similarly to (12) for HMC models, with

$$\begin{cases} P(S_u = j|S_{\rho(u)} = i, \boldsymbol{X} = \boldsymbol{x}) = \beta_u(j)p_{ij}/\{P(S_u = j)\beta_{\rho(u),u}(i)\} \text{ and} \\ P(S_u = j, S_{\rho(u)} = i|\boldsymbol{X} = \boldsymbol{x}) = \beta_u(j)p_{ij}\xi_{\rho(u)}(i)/\{P(S_u = j)\beta_{\rho(u),u}(i)\}. \end{cases} \tag{17}$$

The global state tree entropy $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ is obtained by summation of conditional entropies using (15). The time complexity of the algorithm is in $\mathcal{O}(J^2n)$. As in HMC models, the marginal entropy profile $\{H(S_u|\boldsymbol{X} = \boldsymbol{x})\}_{u\in\mathcal{U}}$ can be viewed as pointwise upper bounds on the conditional entropy profile $\{H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})\}_{u\in\mathcal{U}}$.

Computation of conditional entropies 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. As in the case of HMC models, the global state tree entropy $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ is obtained at the final step of the upward recursion. This approach, described in detail in Appendix B, relies on the computation of the entropies $H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)$ for $u \in \mathcal{U}, u \neq 0$ and for $j = 0, \ldots, J-1$, using an upward recursion.

Partial state tree entropies $H(\bar{\boldsymbol{S}}_u|\boldsymbol{X} = \boldsymbol{x})$ can be deduced from the quantities $H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)$ in the downward recursion. Finally, the conditional entropy profiles $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ are extracted from the latter entropies. The time complexities of the algorithms are in $\mathcal{O}(J^2n)$.

4 Applications

To illustrate the practical ability of entropy profiles to provide localised information on the latent state structure uncertainty, three examples are considered.

1. The first one consists in a synthetic example of HMC model that illustrates how to quantify the roles of emission distributions and the Markovian structure in assessing global state sequence uncertainty.
2. The second one consists in the HMC analysis of the earthquake dataset, published by Zucchini & MacDonald (2009). The third one consists in 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 emphasised.

4.1 Synthetic example of HMC model

A family of 2-state HMC models with known parameters is considered. The transition probability matrix is

$$P = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix}$$

where $\varepsilon \in [0, 0.5]$ is a known parameter. The initial probabilities are $\pi_0 = \pi_1 = 0.5$, which correspond to the stationary state distribution in the ergodic cases ($\varepsilon > 0$). The observed variables X_t take values in $\{0, 1, 2\}$ and the emission distributions are defined by the following emission probability matrix (with states in rows and observations in columns):

$$B = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \end{bmatrix}.$$

Let $\boldsymbol{x} = x_0^{T-1}$ be defined as $x_t = 1$ for $t = 0, \ldots, T-1$. For every $t = 0, \ldots, T-1$, the posterior state probabilities are $P(S_t = 0|\boldsymbol{X} = \boldsymbol{x}) = P(S_t = 1|\boldsymbol{X} = \boldsymbol{x}) = 0.5$. Thus at every time t, the marginal entropy is $\log 2$, independently of ε, and the sum of marginal entropies is $T \log 2$. In contrast, the global state sequence entropy is an increasing function of ε, with $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x}) = \log 2$ for $\varepsilon = 0$ and $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x}) = T \log 2$ for $\varepsilon = 0.5$. In the case where $\varepsilon = 0$, every state is equal to the initial state S_0 and the only uncertainty in the whole state sequence is related to the value of S_0. Thus, the global state sequence entropy is the initial state entropy $\log 2$. In this particular case, all the other states could be deduced from one single known state. The profile of entropies conditional on the next state $H(S_t|S_{t+1}, \boldsymbol{X} = \boldsymbol{x})$ could also be used to quantify local uncertainty (see also discussion in Section 5). This approach would place whole state sequence uncertainty to the value of S_{T-1}, which is correct but inconsistent with the chosen parameterisation of the model. In the case where $\varepsilon = 0.5$, $\boldsymbol{S}$ is a zero-order Markov chain and the global state sequence entropy is also the sum of marginal entropies $T \log 2$.

This interpretation and the associated quantification of state sequence uncertainty are consistent with the conditional entropy profiles depicted in Figure 2.

In contrast, the marginal entropy profiles only translate the fact that both states are equally likely at every time t. In our case, they are identical to the entropy profiles of a zero-order Markov chain. While the conditional entropy profiles highlight that new observations $X_t = 1$ do not increase state sequence uncertainty, the marginal entropy profiles reflect propagation of state uncertainty from a state S_t to its neighbour states S_{t-1} and S_{t+1}. As a consequence, except in the case of independent states, the value of marginal entropy cannot be interpreted in terms of local contributions to global state uncertainty.

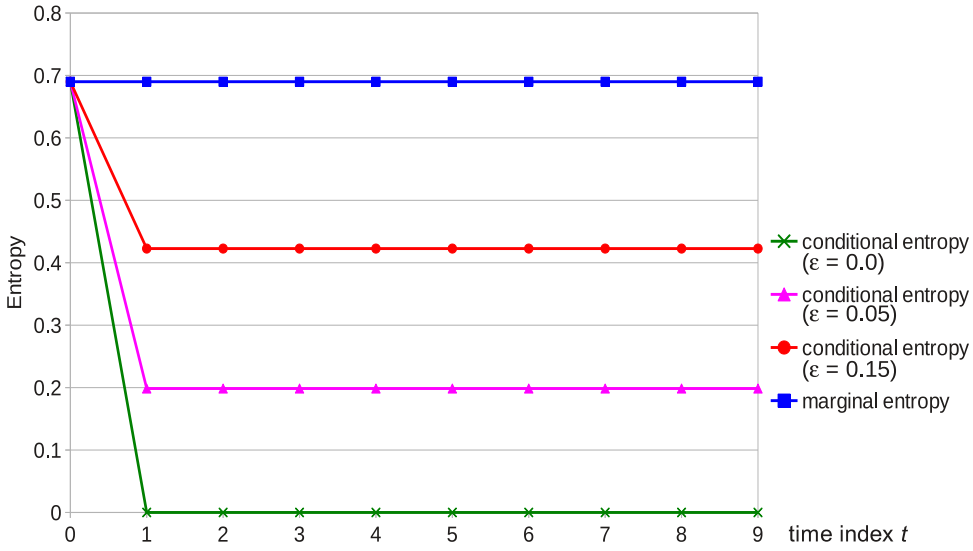

Fig. 2 *Entropy profiles for 2-states HMC models with state transition probabilities* $\varepsilon = 0.0$, $\varepsilon = 0.05$ *and* $\varepsilon = 0.15$.

4.2 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 3.

A 3-state stationary HMC model with Poisson emission distributions was estimated on the basis of this earthquake count sequence using the EM algorithm – see Zucchini & MacDonald (2009). The estimated parameters of the Poisson emission 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 3 as step functions, the level of the segments being either the parameter $\widehat{\lambda}_j$ of the Poisson emission distributions corresponding to the restored state j or the empirical mean estimated for the segment. The state profiles computed by the forward backward algorithm $\{P(S_t = j|\boldsymbol{X} = \boldsymbol{x})\}_{t=0,\ldots,T-1;j=0,\ldots,J-1}$ are shown in Figure 4. The entropy of the state sequence that explains the observed sequence for the estimated HMC model

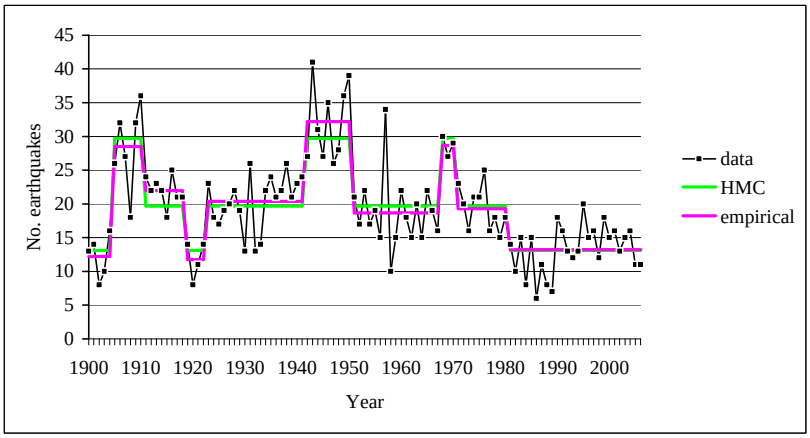

Fig. 3 *Earthquake data: Restored state sequence represented as step functions, the level of the segments being either the parameter $\widehat{\lambda}_j$ of the Poisson emission distributions corresponding to the restored state j or the empirical mean estimated for the segment.*

is bounded from above by the sum of the marginal entropies

$$H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) = \sum_t H(S_t|S_{t-1},\boldsymbol{X}=\boldsymbol{x}) = 14.9$$
$$< \sum_t H(S_t|\boldsymbol{X}=\boldsymbol{x}) = 19.9.$$

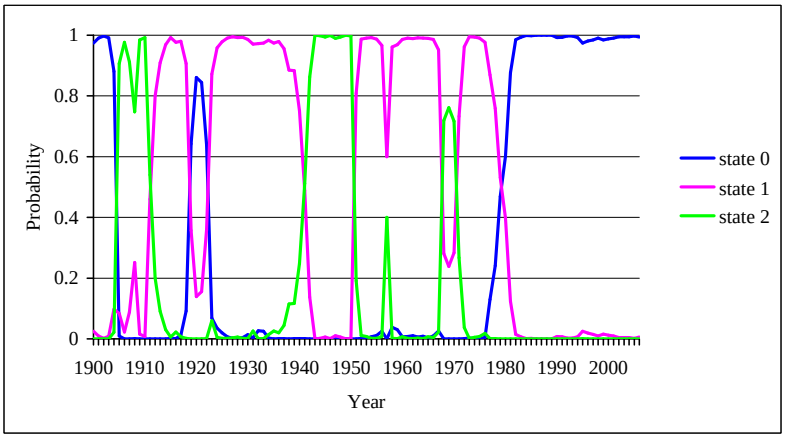

Fig. 4 *Earthquake data: State profiles computed by the forward-backward algorithm.*

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 5.

In Figure 5, the mutual information $I(S_{t-1}; S_t|\mathbf{X} = \mathbf{x})$ between S_{t-1} and S_t, given $\mathbf{X} = \mathbf{x}$ is represented, that is, the difference between the marginal and the conditional entropy at each time t:

$$I(S_{t-1}; S_t|\mathbf{X} = \mathbf{x}) = H(S_t|\mathbf{X} = \mathbf{x}) - H(S_t|S_{t-1}, \mathbf{X} = \mathbf{x}).$$

This mutual information is highly variable as a function of t and the dates where this mutual information is high tend to be aggregated (between 1912 and 1913, 1920 and 1922, 1939 and 1941, 1969 and 1970, 1978 and 1981 where $I(S_{t-1}; S_t|\mathbf{X} = \mathbf{x}) > 0.12$); see Figure 5. In these segments $[t, t']$ of high mutual information, the canonical measure of state uncertainty $H(S_t^{t'}|\mathbf{X} = \mathbf{x})$ is far less than suggested by the marginal entropies, and consequently by the posterior state probabilities.

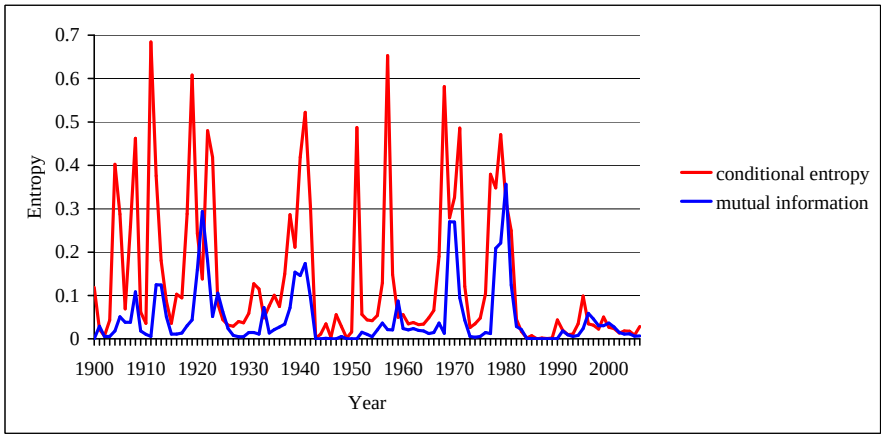

Fig. 5 *Earthquake data: Profiles of conditional entropies and of mutual information.*

4.3 Analysis of the structure of Aleppo pines

The aim of this study was to build a model of the architectural development of Aleppo pines. Its has been shown previously that HMC and HMT models can be used to analyse the architectural development of fruit and forest trees over several years on the basis of retrospective measurements of annual shoot characteristics; see Durand *et al.* (2005) and Guédon *et al.* (2007a). In HMT models, variables X_u are observed along a given tree $\mathcal{T}$ with vertex set $\mathcal{U}$ used as index for the observed process $\boldsymbol{X} = (X_u)_{u \in \mathcal{U}}$.

This type of model enables typical successions of annual shoots to be identified and characterised within tree structures. These successions generally extend from long polycyclic highly branched annual shoots in proximal positions to sterile or reproductive short monocyclic unbranched annual shoots in distal positions in the Aleppo pine case. Regarding biology, one strength of this approach is the capability to infer complex dynamical information on the basis of retrospective measurements.

The data set is composed of seven branches of Aleppo pines (*Pinus Halepensis* Mill., *Pinaceae*) planted in the south of France (Clapiers, Hérault). The branches came 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. A given year of growth can be divided into three periods of potential growth, referred to as *(growth) cycles*. For a given annual shoot, if growth occurred during the first cycle only, it is said to be *monocyclic*. Otherwise, growth occurred during the first cycle and at least one more cycle, and the annual shoot is said to be *polycyclic*. Moreover, a given annual shoot may or not bear sexual organs: female cones, male cones, or no cone at all. In the latter case, the annual shoot it is said to be *sterile*. Five variables were recorded for each annual shoot: length (in cm), number of branches per tier, number of growth cycles beyond the first one and presence or absence of female cones and of male cones. On these seven branches, a total of 836 annual shoots was measured.

Competing models An HMT model was estimated on the basis of the seven branches, to identify categories of annual shoots with comparable values for the variables, and to characterise the succession of the categories within the branches. The set of parameters θ (including the initial and transition probabilities and the emission distribution parameters) was estimated using the EM algorithm for HMT models – see Crouse *et al.*(1998). The branches were considered as mutually independent random realizations of a same HMT model (the trees have same parameter set but different structures.) Except for the length variable, the emission distributions were multinomial distributions $\mathcal{M}(1; p_1, \ldots, p_N)$, where N denotes the number of possible values for this variable. Four families of parametric discrete distributions were considered for the emission distributions associated with the length variable: uniform, Poisson, binomial and negative binomial families of distributions, each with an additional shift parameter. The family associated with the maximum likelihood of the parameters was selected (in our case, negative binomial distributions for each 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 model selection criteria. We resorted to ICL-BIC (McLachlan & Peel, 2000, chap. 6) to select this number. ICL-BIC is defined by

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

where n is the number of vertices in $\boldsymbol{X}$, $P_\theta(\boldsymbol{x})$ the likelihood of parameter θ, $\hat{\theta}_J$ the estimated parameters for a J-state HMT model, d_J the number of independent model parameters, and the entropy $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ is computed as shown in Section 3. ICL-BIC incorporates the aim of obtaining non-ambiguous state restoration in model selection.

The maximal number of possible states was set to 10, and a 6-state model was selected (with an ICL-BIC value of -10,704) followed by 5-state and 4-state models (with respective values of BIC -10,742 and -10,764).

Entropy profiles in the 6-state HMT model The estimated transition matrix of the 6-state HMT model 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 Markov tree is initialised in state 0 with probability 1. It can be seen from $\hat{P}$ that the Markov tree has transient state 0, transient class $\{1, 2, 3\}$, transient state 4 and absorbing state 5. Hence, only states 1, 2 and 3 can be visited more than once along a path within a tree. State transitions and an interpretation of the hidden states are provided in Figure 6. In particular, the states are ordered by decreasing length, except state 5, which corresponds to slightly longer shoots than state 4.

To quantify the separability between consecutive states i and j (i.e. such that $p_{ij} > 0$), we computed distances between the corresponding discrete emission distributions for each observed variable in the form $1 - \sum_y \min\{b_i(y), b_j(y)\}$. This distance, which is one minus the overlap between the two emission distributions, is between 0 (full overlap, i.e. identical distributions) and 1 (no overlap). For quantitative variables (shoot length in our case), this distance is also the sup-norm distance between the two emission distributions (i.e. the maximum absolute difference between the cumulative distribution functions) in the case of non-crossing cumulative distribution functions

$$1 - \sum_y \min\{b_i(y), b_j(y)\} = \sup_y \left| \sum_{x=0}^{y} b_j(x) - \sum_{x=0}^{y} b_i(x) \right|.$$

In our case of multivariate observations, we only give the highest distance among the five distances computed for each pair of states and for each variable; see Table 1. This highest distance is assumed to reflect the separability between consecutive states. It appears that the 6 states are well separated, except pairs $(2, 3)$ and $(3, 5)$. However, unbranched, monocyclic, sterile shoots 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	$0 \to 1$	$0 \to 2$	$0 \to 3$	$1 \to 2$	$2 \to 3$	$3 \to 1$	$3 \to 4$	$3 \to 5$	$4 \to 5$
sup-norm distance	0.85	0.80	0.89	0.78	0.20	0.98	0.99	0.48	0.97
most separated variable	br.	fem.	fem.	br.	br.	br.	male	len.	male

Table 1 *Distances between emission distributions for pairs of successive states, for the variable that achieves the maximal distance among the five variables (br. – branching; fem. – female cones; len. – length; male – male cones.)*

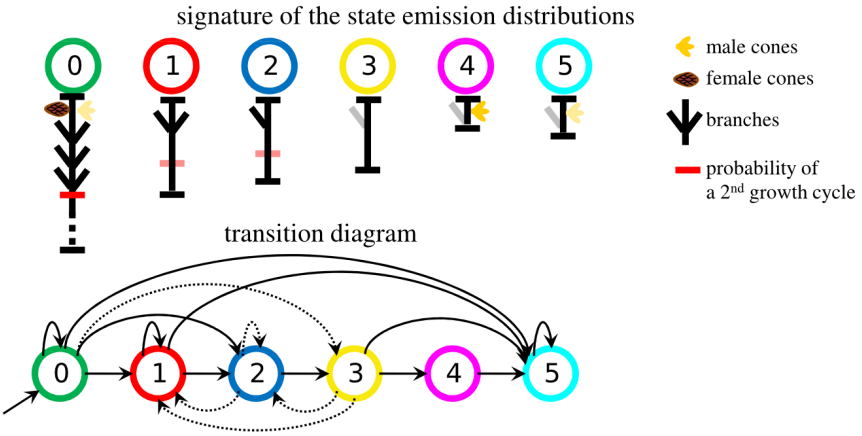

Fig. 6 *6-state HMT model: transition diagram and symbolic representation of the state signatures (conditional mean values of the variables given the states, depicted by typical shoots). The separation between growth cycle is represented by a horizontal red segment, which intensity is proportional to the probability of occurrence of a second growth cycle. 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).*

To analyse how state ambiguity due to unbranched, monocyclic, sterile shoots affects state restoration, entropy profiles were computed for each individual (namely, each branch). Firstly, the annual shoots were represented using a colourmap, which is a mapping between colours and the values of conditional entropies $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ (see Figure 7a)). Vertices with lowest conditional entropy are represented in blue, whereas those with highest conditional entropy are in red.

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 7b. This representation shows where the states are located within the tree; for example state 0 is located on the main axis (main stem) and at the basis of lateral axis. Moreover, in conjunction with Figure 7a, it highlights some states for which the restoration step is not much ambiguous (in our example, states 0, 3, 4, and 1 to a least extent). Thus, these states with low conditional entropy correspond to vertices with the highest number of branches, female or male cones. On the contrary, the vertices with highest conditional entropy are mostly unbranched, monocyclic and sterile, and are located at peripheral parts of the plant.

A two-step analysis was performed to identify locations characterised by particularly high state uncertainty. The profile of conditional entropies in Figure 7a was used in a first step to select zones of vertices with high conditional entropies. In a second step, local alternatives to the Viterbi restoration were identified, using the so-called *upward-downward Viterbi profiles* as a complement to the entropy profiles. They rely on the following quantities

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

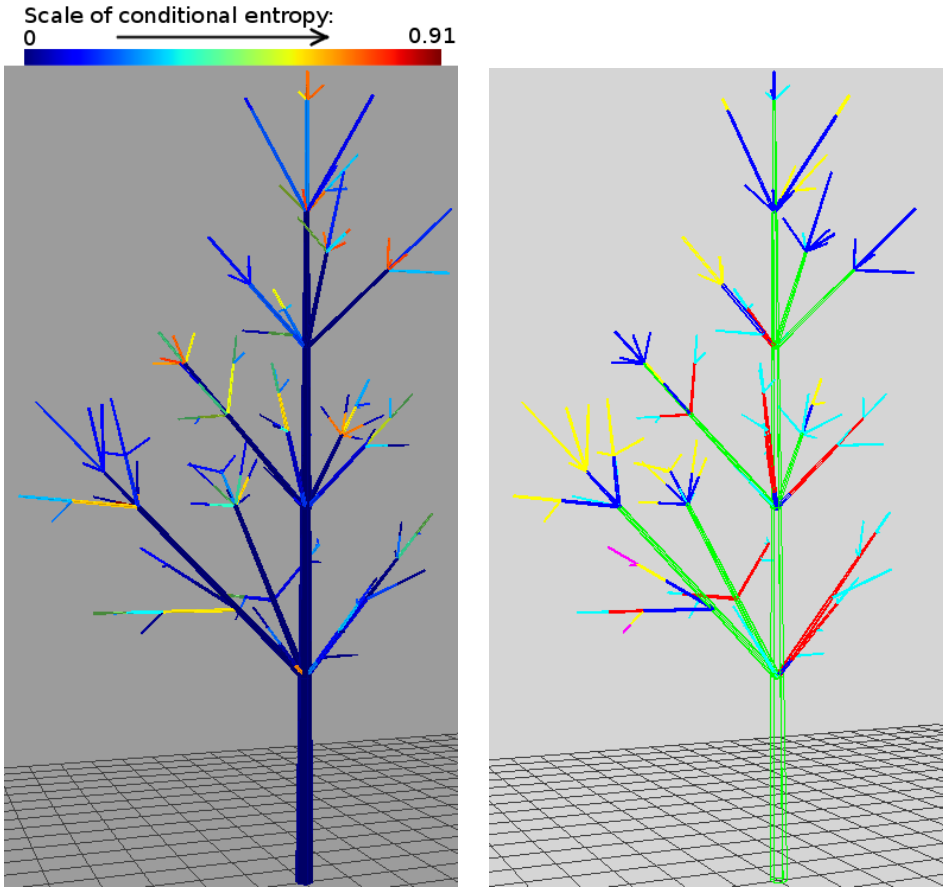

Fig. 7 *Conditional entropy and state tree restoration for a given branch. a) Conditional entropy* $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ *using a colourmap. Blue corresponds to lowest entropy and red to highest entropy. b) State tree restoration. The correspondence between states and colors is as follows: state 0 - green ; state 1 - red ; state 2 - blue ; state 3 - yellow ; state 4 - magenta ; state 5 - cyan.*

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. They were used by Guédon (2007b) as diagnostic tools for localization of state uncertainty in the context of hidden (semi-) Markov chains. This analysis leads to detailed understanding of the roles of the model and the observed variables to yield especially high or low state uncertainty, since both cases can be informative. In application of this methodology, two paths (extracted from two distinct individuals) were chosen for the contrasted situations they yielded. The detailed analysis of a path containing successive monocyclic, sterile shoots is provided hereafter. The analysis of a path containing a female shoot is given in Appendix C.

A path essentially composed by monocyclic, sterile shoots is considered within the fourth individual (for which $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) = 47.5$). The path contains 5 vertices, referred to as $\{0,\dots,4\}$. 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 8b), it can be in states 2, 3 or 5. This is summarised by the entropy profiles in Figure 8a).

This conditional entropy profile can be further interpreted, with contrasted interpretations according to whether mutual information (represented in Figure 8c)) is positive or null, in cases where marginal entropy remains positive. On the one hand, $I(S_1;S_2|\boldsymbol{X}=\boldsymbol{x}) = 0$. This results from state S_1 being known. Thus, conditioning by S_1 does not provide further information on its children state S_2. On the other hand, $I(S_3;S_4|\boldsymbol{X}=\boldsymbol{x}) = 0.2$. Uncertainty associated with the posterior distribution of S_4 is high, since $H(S_4|\boldsymbol{X}=\boldsymbol{x}) = 0.67$. However, knowledge of its parent state S_3 would reduce the uncertainty on S_4: if $S_3 = 1$ then $S_4 = 5$; if $S_3 = 2$ then $S_4 = 2$ (or less likely, $S_4 = 3$) and if $S_3 = 3$ then $S_4 = 5$ (or less likely, $S_4 = 2$).

Using Proposition A1 in Appendix A, the contribution of the vertices of the considered path $\mathcal{P}$ to the global state tree entropy can be computed as:

$$H(S_0|\boldsymbol{X}=\boldsymbol{x}) + \sum_{\substack{u\in\mathcal{P}\\u\neq 0}} H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}), \tag{18}$$

and is equal to 1.41 in the above example (that is, 0.28 per vertex on average). The global state tree entropy for this individual is 0.24 per vertex, against 0.20 per vertex in the whole dataset. 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 contribution of $\mathcal{P}$ to the global state tree entropy corresponds to the sum of the heights of every point of the profile of conditional entropies in Figure 8a).

Note that the representation of state uncertainty using profiles of posterior state probabilities induces a perception of global uncertainty on the states along $\mathcal{P}$ equivalent to that provided by marginal entropy profile in Figure 8a). The mean marginal state entropy for this individual is 0.37 per vertex, which strongly overestimates the mean state tree entropy.

This example shows that detailed insight can be brought by joint use of entropy profiles and the Viterbi algorithm with its variants. The assignment of states to vertices performed by the model is a global operation; however, the local role of the observed data and the neighbouring states can be understood precisely, by combination of the Viterbi algorithm and conditional entropy profiles.

Comparison between marginal and conditional entropy profiles 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.

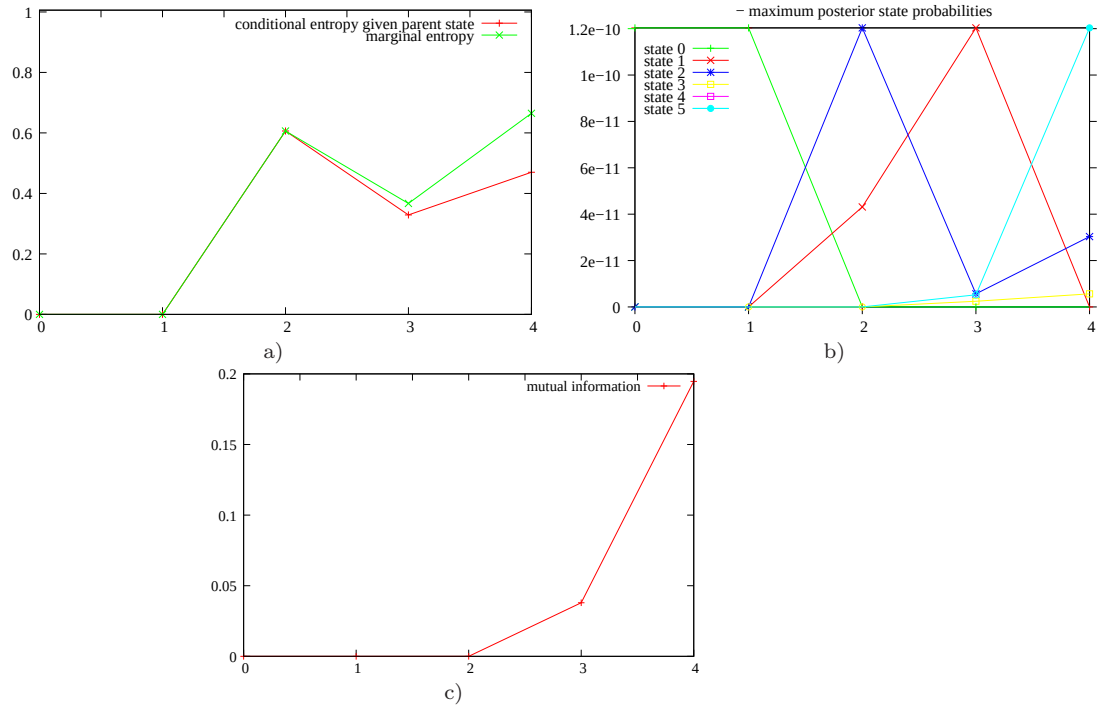

Fig. 8 *Entropy profiles along a path containing mainly sterile monocyclic shoots. a) Profiles of conditional and of marginal entropies. b) State tree restoration with the Viterbi upward-downward algorithm. c) Mutual information between a state and its parent state.*

To assess the overestimation of state uncertainty induced by using the profiles based on the marginal entropies $H(S_u|\boldsymbol{X} = \boldsymbol{x})$ instead of $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$, these quantities were computed for each tree in the dataset. The ratios $(M(\mathcal{T}) - G(\mathcal{T}))/G(\mathcal{T})$ are given in Table 2.

Tree $\mathcal{T}$ number	1	2	3	4	5	6	7
$\frac{M(\mathcal{T}) - G(\mathcal{T})}{G(\mathcal{T})}$	69.1 %	78.0 %	76.4 %	56.0 %	85.2 %	73.5 %	85.1 %

Table 2 *Relative distance $(M(\mathcal{T}) - G(\mathcal{T}))/G(\mathcal{T})$ between sum of marginal entropies $M(\mathcal{T})$ and global state tree entropy $G(\mathcal{T})$.*

It can be seen from Table 2 that $M(\mathcal{T})$ is a poor approximation of the global state tree entropy. As a consequence, the posterior state probability profiles cannot be used to quantify local contributions to global state uncertainty.

5 Concluding remarks

This article proposes a new methodology to assess state uncertainty in GHM models. It has been shown that global state entropy can be decomposed additively along the graph structure. Each element of the decomposition can be interpreted in terms of local state uncertainty, corresponding to conditional entropy. This makes it relevant to draw profiles of conditional entropies, indexed by the graph vertices. In the particular case of HMC and HMT models, we provided efficient algorithms to compute these profiles.

Used jointly with the Viterbi algorithm and its variants, these profiles allow deeper understanding of the local roles of the model parameters, the neighbouring states and the observed data, concerning state uncertainty. This leads to a much more efficient approach than the plain Viterbi algorithm and the posterior state probability profiles to analyse alternative state restorations, which may involve zones of connected vertices. Such situations are characterised by high mutual information between connected vertices. Moreover, we showed using examples that the posterior state probability profiles introduce confusion between (i) local state uncertainty due to overlap of emission distributions for different states and (ii) mere propagation of uncertainty from past to future states. Contrary to conditional entropy profiles, they suggest strong local contributions to global state uncertainty in zones where such uncertainty is in fact far more limited.

In Durand & Guédon (2012), algorithms similar to those presented in Section 2 have been proposed for computing profiles of entropies conditional on the next state in the HMC model case. This provides the other possible decomposition of the global state sequence entropy:

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

The interpretation of the profile of entropies conditional on the next state $\{H(S_t|S_{t+1},\boldsymbol{X}=\boldsymbol{x})\}_{t=0,\ldots,T-1}$ is far less obvious than that of the profile of entropies conditional on the previous state $\{H(S_t|S_{t-1},\boldsymbol{X}=\boldsymbol{x})\}_{t=0,\ldots,T-1}$ except in the case of reversible processes. In the same way, profiles of entropies conditional on the children states $\left\{H(S_u|\boldsymbol{S}_{c(u)},\boldsymbol{X}=\boldsymbol{x})\right\}_{u\in\mathcal{U}}$ were obtained in the HMT model case. Contrarily to the profile of entropies conditional on the parent state $\left\{H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\right\}_{u\in\mathcal{U}}$, the profile of entropies conditional on the children states does not constitute a decomposition of global state tree entropy and we proved (Durand & Guédon, 2012) that

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

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.

Stronger connexions can be hypothesised between entropy and the so-called generalised Viterbi algorithm, which enumerates the L state sequences or trees $\boldsymbol{s}$ with highest probabilities. In particular, the notion of typical set allows the cardinality of a subset of possible values of $\boldsymbol{s}$ with given minimal probability to be

bounded by some functions of the entropy (Cover and Thomas, 2006). This could lead to a bound on the value of L to be used in the generalised Viterbi algorithm.

In the perspective of model selection, entropy computation may also appear as a valuable tool. If irrelevant states are added to a GHM 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). This results from perturbations in the state conditional distribution induced by estimation from a finite sample. This intuitive statement explains why several model selection criteria based on a compromise between log-likelihood and state entropy were proposed. Among these is the Normalised Entropy Criterion introduced by Celeux & Soromenho (1996) in independent mixture models, and ICL-BIC criterion introduced by McLachlan & Peel (2000, chap. 6). Their generalization to GHM models is rather straightforward. By favouring models with small state entropy and high log-likelihood, these criteria aim at selecting models such as the uncertainty of the state values is low, whilst achieving good fit to the data.

Appendix.

A Computation of the global entropy of state subtrees in hidden Markov tree models

Proposition A1 *Let $\mathcal{V}$ be a subtree of $\mathcal{T}$ with root vertex r. Then for any possible value $\bar{s}_{\mathcal{V}}$ of $\bar{\boldsymbol{S}}_{\mathcal{V}}$ and for any $\boldsymbol{x}$,*

$$P(\bar{\boldsymbol{S}}_{\mathcal{V}} = \bar{s}_{\mathcal{V}} | \boldsymbol{X} = \boldsymbol{x}) = P(S_r = s_r | \boldsymbol{X} = \boldsymbol{x}) \prod_{\substack{u \in \mathcal{V} \\ u \neq r}} P(S_u = s_u | S_{\rho(u)} = s_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$$

and

$$H(\bar{\boldsymbol{S}}_{\mathcal{V}} | \boldsymbol{X} = \boldsymbol{x}) = H(S_r | \boldsymbol{X} = \boldsymbol{x}) + \sum_{\substack{u \in \mathcal{V} \\ u \neq r}} H(S_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}).$$

Proof This is proved by induction on the vertices in $\mathcal{V}$. The induction step is as follows: let ℓ be a leaf vertex of $\mathcal{V}$. Then for any possible value $\bar{s}_{\mathcal{V}}$ of $\bar{\boldsymbol{S}}_{\mathcal{V}}$,

$$\begin{aligned} &P(\bar{\boldsymbol{S}}_{\mathcal{V}} = \bar{\boldsymbol{s}}_{\mathcal{V}} | \boldsymbol{X} = \boldsymbol{x}) \\ &= P(S_\ell = s_\ell | \bar{\boldsymbol{S}}_{\mathcal{V}\backslash\{\ell\}} = \bar{\boldsymbol{s}}_{\mathcal{V}\backslash\{\ell\}}, \boldsymbol{X} = \boldsymbol{x}) P(\bar{\boldsymbol{S}}_{\mathcal{V}\backslash\{\ell\}} = \bar{\boldsymbol{s}}_{\mathcal{V}\backslash\{\ell\}} | \boldsymbol{X} = \boldsymbol{x}) \\ &= P(S_\ell = s_\ell | S_{\rho(\ell)} = s_{\rho(\ell)}, \boldsymbol{X} = \boldsymbol{x}) P(\bar{\boldsymbol{S}}_{\mathcal{V}\backslash\{\ell\}} = \bar{\boldsymbol{s}}_{\mathcal{V}\backslash\{\ell\}} | \boldsymbol{X} = \boldsymbol{x}) \end{aligned}$$

since S_ℓ is conditionally independent from the other vertices in $\mathcal{V}$ given $S_{\rho(\ell)}$ and $\boldsymbol{X}$.

The induction step is completed by observing that $\mathcal{V}\backslash\{\ell\}$ is a subtree of $\mathcal{T}$.

The decomposition of the entropy of $\bar{\boldsymbol{S}}_{\mathcal{V}}$ yielded by the chain rule

$$H(\bar{\boldsymbol{S}}_{\mathcal{V}} | \boldsymbol{X} = \boldsymbol{x}) = H(S_r | \boldsymbol{X} = \boldsymbol{x}) + \sum_{\substack{u \in \mathcal{V} \\ u \neq r}} H(S_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$$

is proved similarly as Corollary 1. ■

B Direct computation of global state tree entropy in hidden Markov tree models

Direct computation of global state tree entropy is based on recursive computation of the entropies of children state subtrees given each state. This recursion relies on conditional independence properties between hidden and observed variables in HMT models, and particularly the following relations: for any internal, non-root vertex u and for $j = 1, \ldots, J$,

$$\begin{aligned} &P(\bar{\boldsymbol{S}}_{c(u)} = \bar{\boldsymbol{s}}_{c(u)} | S_u = j, \bar{\boldsymbol{S}}_{0\backslash u} = \bar{\boldsymbol{s}}_{0\backslash u}, \boldsymbol{X} = \boldsymbol{x}) \\ &\qquad = P(\bar{\boldsymbol{S}}_{c(u)} = \bar{\boldsymbol{s}}_{c(u)} | S_u = j, S_{\rho(u)} = s_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}) \\ &\qquad = P(\bar{\boldsymbol{S}}_{c(u)} = \bar{\boldsymbol{s}}_{c(u)} | S_u = j, \boldsymbol{X} = \boldsymbol{x}) \\ &\qquad = \prod_{v \in c(u)} P(\bar{\boldsymbol{S}}_v = \bar{\boldsymbol{s}}_v | S_u = j, \boldsymbol{X} = \boldsymbol{x}) \\ &\qquad = \prod_{v \in c(u)} P(\bar{\boldsymbol{S}}_v = \bar{\boldsymbol{s}}_v | S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \qquad (19) \\ &\qquad = P(\bar{\boldsymbol{S}}_{c(u)} = \bar{\boldsymbol{s}}_{c(u)} | S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u), \qquad (20) \end{aligned}$$

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

Entropies $H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)$ can be computed for any $u \in \mathcal{U}, u \neq 0$ and for $j = 0, \ldots, J-1$, by an upward algorithm initialised at the leaf vertices u by

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

As a consequence from (20), we have for any state j, $H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) = H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \boldsymbol{X} = \boldsymbol{x})$. Thus, it is deduced from (19) that

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

Moreover, for any $v \in c(u)$ with $c(v) \neq \emptyset$ and for $j = 0, \ldots, J-1$,

$$\begin{aligned} &H(\bar{\boldsymbol{S}}_v|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &= - \sum_{\bar{\boldsymbol{s}}_{c(v)}, s_v} P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}, S_v = s_v|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &\quad \times \log P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}, S_v = s_v|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &= - \sum_{\boldsymbol{s}_{c(v)}} \sum_{k=0}^{J-1} P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}|S_v = k, S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &\quad \times \{\log P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}|S_v = k, S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) + \log P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u)\} \\ &= - \sum_{k=0}^{J-1} P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \left\{ \sum_{\boldsymbol{s}_{c(v)}} P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}|S_v = k, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right. \\ &\quad \left. \times \ \log P(\bar{\boldsymbol{S}}_{c(v)} = \boldsymbol{s}_{c(v)}|S_v = k, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) + \log P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right\} \\ &= \sum_{k=0}^{J-1} P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \left\{ H(\bar{\boldsymbol{S}}_{c(v)}|S_v = k, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right. \\ &\quad \left. - \log P(S_v = k|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right\}. \end{aligned} \tag{22}$$

Thus, the recursion of the upward algorithm is given by

$$\begin{aligned} &H(\bar{\boldsymbol{S}}_{c(u)}|S_u = j, \bar{\boldsymbol{X}}_u = \bar{\boldsymbol{x}}_u) \\ &= \sum_{v \in c(u)} \left\{ \sum_{s_v} P(S_v = s_v|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \left[H(\bar{\boldsymbol{S}}_{c(v)}|S_v = s_v, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right. \right. \\ &\quad \left. \left. - \log P(S_v = s_v|S_u = j, \bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v) \right] \right\}, \end{aligned} \tag{23}$$

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 (17).

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

$$\begin{aligned} H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x}) =& H(\bar{\boldsymbol{S}}_{c(0)}|S_0, \boldsymbol{X} = \boldsymbol{x}) + H(S_0|\boldsymbol{X} = \boldsymbol{x}) \\ =& \sum_{j=0}^{J-1} \beta_0(j) \left\{ H\left(\bar{\boldsymbol{S}}_{c(0)}|S_0 = j, \boldsymbol{X} = \boldsymbol{x}\right) - \log \beta_0(j) \right\}. \end{aligned}$$

Using similar arguments as in (22), 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 $j = 0, \ldots, J-1$) as

follows:

$$\begin{aligned} H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x}) =& H(\bar{\boldsymbol{S}}_{c(u)}|S_u,\boldsymbol{X}=\boldsymbol{x}) + H(S_u|\boldsymbol{X}=\boldsymbol{x}) \\ =& \sum_j \xi_u(j)\left\{H\left(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\boldsymbol{X}=\boldsymbol{x}\right) - \log \xi_u(j)\right\} \\ =& \sum_j \xi_u(j)\left\{H\left(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u\right) - \log \xi_u(j)\right\}, \end{aligned} \tag{24}$$

where the $\{\xi_u(j)\}_{j=0,\ldots,J-1}$ are directly extracted from the downward recursion (14).

The profile of conditional entropies $H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})$ is deduced from

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

where

$$\begin{aligned} H(\bar{\boldsymbol{S}}_{b(u)}|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}) =& \sum_{v\in b(u)} H(\bar{\boldsymbol{S}}_v|S_{\rho(v)},\boldsymbol{X}=\boldsymbol{x}) \\ =& \sum_{v\in b(u)} \left\{\sum_j \xi_{\rho(v)}(j) H(\bar{\boldsymbol{S}}_v|S_{\rho(v)}=j,\boldsymbol{X}=\boldsymbol{x})\right\}, \end{aligned}$$

and where for any brother vertex v of u, $H(\bar{\boldsymbol{S}}_v|S_{\rho(v)}=j,\boldsymbol{X}=\boldsymbol{x})$ is given by (22). Since

$$H(\bar{\boldsymbol{S}}_v|S_{\rho(v)},\boldsymbol{X}=\boldsymbol{x}) = \sum_j H(\bar{\boldsymbol{S}}_v|S_{\rho(v)}=j,\boldsymbol{X}=\boldsymbol{x})\xi_{\rho(v)}(j)$$

and since

$$\begin{aligned} H(\bar{\boldsymbol{S}}_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}) =& H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}) + H(\bar{\boldsymbol{S}}_c(u)|S_u,\boldsymbol{X}=\boldsymbol{x}) \\ =& H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x}) + \sum_j \xi_u(j) H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\boldsymbol{X}=\boldsymbol{x}), \end{aligned} \tag{25}$$

$H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})$ is directly extracted from the partial state entropies $H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\boldsymbol{X}=\boldsymbol{x})$ and $H(\bar{\boldsymbol{S}}_{\rho(u)}|\boldsymbol{X}=\boldsymbol{x})$ and from the marginal entropy $H(S_{\rho(u)}|\boldsymbol{X}=\boldsymbol{x})$.

In summary, the partial subtrees entropies $\left\{H(\bar{\boldsymbol{S}}_{c(u)}|S_u=j,\bar{\boldsymbol{X}}_u=\bar{\boldsymbol{x}}_u)\right\}_{u\in\mathcal{U};\ j=0,\ldots,J-1}$ are firstly computed using (23). The partial state tree entropies $\left\{H(\bar{\boldsymbol{S}}_u|\boldsymbol{X}=\boldsymbol{x})\right\}_{u\in\mathcal{U}}$ and then the profile of conditional entropies $\left\{H(S_u|S_{\rho(u)},\boldsymbol{X}=\boldsymbol{x})\right\}_{u\in\mathcal{U}}$ are deduced from these entropies and the posterior state probabilities, using (24) and (25). The time complexity of the algorithm is in $\mathcal{O}(J^2n)$.

C Application of HMT model to Aleppo pines: path containing female shoots

A path containing a female shoot is considered. This path corresponds to the main axis of the third individual (for which $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) = 29.6$). 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. Shoots 4 and 5 are unbranched, monocyclic, sterile shoots.

The contribution of the vertices of the considered path $\mathcal{P}$ to the global state tree is equal to 0.48 (that is, 0.08 per vertex on average). The global state tree entropy for this individual is 0.21 per vertex, against 0.20 per vertex in the whole dataset. The mean marginal state entropy for this individual is 0.37 per vertex, which strongly overestimates the mean state tree entropy.

Since a female shoot necessarily is in state 0, $H(S_2|\boldsymbol{X}=\boldsymbol{x}) = 0$ (no uncertainty). The states of shoots 0 and 1 can be deduced from S_2 using the transition matrix $\hat{P}$, thus their

marginal entropy is null. Since shoot 3 is bicyclic, it is in state 0 with a very high probability ($H(S_3|\boldsymbol{X}=\boldsymbol{x}) \approx 0$). Uncertainty remains concerning the states of shoots 4 and 5, which thus have high marginal entropies. However, S_5 can be deduced from S_4 using $\hat{P}$ and inversely, which results into high mutual information between S_4 and S_5 given $\boldsymbol{X}=\boldsymbol{x}$. This is illustrated by conditional and marginal entropy profiles in Figure 9.

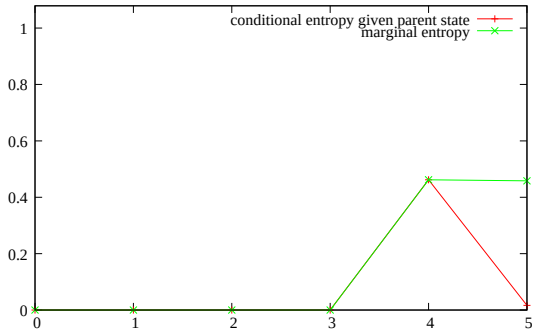

Fig. 9 *Path containing a female shoot: Profiles of conditional and of marginal entropies.*

Acknowledgements The authors are indebted to Yves Caraglio for useful comments on modelling the Aleppo pines dataset and for providing this dataset.

References

1. Brushe, G., Mahony, R., Moore, J.: A Soft Output Hybrid Algorithm for ML/MAP Sequence Estimation. IEEE Trans. Inf. Theory **44**(7), 3129–3134 (1998)
2. Cadre, J.P.L., Tremois, O.: Bearing-Only Tracking for Maneuvering Sources. IEEE Trans. Aerosp. Electron. Syst. **34**(1), 179–193 (1998)
3. Cappé, O., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer Series in Statistics. New York: Springer (2005)
4. Celeux, G., Soromenho, G.: An entropy criterion for assessing the number of clusters in a mixture model. J. Classif. **13**(2), 195–212 (1996)
5. Cover, T., Thomas, J.: Elements of Information Theory, 2nd edition. Hoboken, NJ: Wiley (2006)
6. Crouse, M., Nowak, R., Baraniuk, R.: Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. IEEE Trans. Signal Process. **46**(4), 886–902 (1998)
7. Devijver, P.A.: Baum's forward-backward Algorithm Revisited. Pattern Recognit. Lett. **3**, 369–373 (1985)
8. Durand, J.-B., Girard, S., Ciriza, V., Donini, L.: Optimization of power consumption and device availability based on point process modelling of the request sequence. Appl. Stat. **62**(2), 151–162 (2013)
9. Durand, J.-B., Gonçalvès, P., Guédon, Y.: Computational methods for hidden Markov tree models – an application to wavelet trees. IEEE Trans. Signal Process. **52**(9), 2551–2560 (2004)
10. Durand, J.-B., Guédon, Y., Caraglio, Y., Costes, E.: Analysis of the Plant Architecture via Tree-structured Statistical Models: the Hidden Markov Tree Models. New Phytol. **166**(3), 813–825 (2005)
11. Durand, J.-B., Guédon, Y.: Localizing the Latent Structure Canonical Uncertainty: Entropy Profiles for Hidden Markov Models. Available: hal.inria.fr/hal-00675223/en, Inria technical report (2012)

12. Ephraim, Y., Merhav, N.: Hidden Markov processes. IEEE Trans. Inf. Theory **48**, 1518–1569 (2002)
13. Guédon, Y., Caraglio, Y., Heuret, P., Lebarbier, E., Meredieu, C.: Analyzing growth components in trees. J. Theor. Biol. **248**(3), 418–447 (2007a)
14. Guédon, Y.: Exploring the state sequence space for hidden Markov and semi-Markov chains. Comput. Stat. Data An. **51**(5), 2379–2409 (2007b)
15. Guédon, Y.: Segmentation uncertainty in multiple change-point models. Stat. Comput., in press (2013)
16. Hernando, D., Crespi, V., Cybenko, G.: Efficient computation of the hidden Markov model entropy for a given observation sequence. IEEE Trans. Inf. Theory **51**(7), 2681–2685 (2005)
17. Lauritzen, S.: Graphical Models. Clarendon Press, Oxford, United Kingdom (1996)
18. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics. John Wiley and Sons (2000)
19. Zucchini, W., MacDonald, I.: Hidden Markov Models for Time Series: *An Introduction Using R.* Chapman & Hall/CRC: Boca Raton FL (2009)