Quantifying and localizing state uncertainty in hidden Markov models using conditional entropy profiles

Jean-Baptiste Durand, Yann Guédon

Quantifying and localizing state uncertainty in hidden Markov models using conditional entropy profiles

Jean-Baptiste Durand, *Univ. Grenoble Alpes, LJK and Inria, Mistis, F-38000 Grenoble, France*, jean-baptiste.durand@imag.fr
Yann Guédon, *CIRAD, UMR AGAP and Inria, Virtual Plants, F-34095 Montpellier, France*, guedon@cirad.fr

Abstract. A family of graphical hidden Markov models that generalizes hidden Markov chain (HMC) and tree (HMT) models is introduced. It is shown that global uncertainty on the state process can be decomposed as a sum of conditional entropies that are interpreted as local contributions to global uncertainty. An efficient algorithm is derived to compute conditional entropy profiles in the case of HMC and HMT models. The relevance of these profiles and their complementarity with other state restoration algorithms for interpretation and diagnosis of hidden states is highlighted. It is also shown that classical smoothing profiles (posterior marginal probabilities of the states at each time, given the observations) cannot be related to global state uncertainty in the general case.

Keywords. Hidden Markov models, State inference, Conditional entropy.

1 Introduction

Hidden Markov models (HMMs) have been used frequently in sequence analysis for modeling various types of latent structures, such as homogeneous zones or noisy patterns (Ephraim & Mehrav, 2002). They have been extended from sequences to more general structures, particularly tree structures. In HMMs, inference for model parameters can be distinguished from inference for the state process given parameters. This work focuses on state process inference.

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 (Ephraim

& Mehrav, 2002). Such use of the state sequence relies on the assumption that uncertainty on the state process given observations should be reasonably low. 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 – typically, to compare histograms with conditional densities given the states. The use of restored states in the above-mentioned contexts makes assessment of state sequence uncertainty a critical step of the analysis.

Global quantification of such uncertainty has been addressed by Hernando *et al.* (2005). However, this is insufficient for detailed state interpretation: knowledge of the distribution of that global uncertainty along the structure is also of primary importance. Quantification of local state uncertainty given observed sequence $\boldsymbol{X} = \boldsymbol{x}$ for a known HMC model has been adressed by either enumeration of state sequences, or by state profiles, which are state sequences summarised in a $K \times T$ array, T being the sequence length and K the number of states (Guédon, 2007).

We here address quantification of state uncertainty in an HMM with observed process $\boldsymbol{X} = (X_v)_{v\in\mathcal{V}}$ indexed by a fixed Directed Acyclic Graph (DAG) $\mathcal{G}$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. This family of HMMs is referred to as graphical hidden Markov models (GHMMs). This family contains hidden Markov chain (HMC) and tree (HMT) models. Let $\boldsymbol{S} = (S_v)_{v\in\mathcal{V}}$ denote the associated hidden state process, S_v taking values in the set $\{0, \ldots, K-1\}$. Let $\boldsymbol{x}$ be a possible realization of $\boldsymbol{X}$. Let $\mathrm{pa}(v)$ denote the parent of vertex v and for any subset U of $\mathcal{V}$, let X_U (resp. $\boldsymbol{x}_U$) denote the family of random variables $(X_u)_{u\in U}$ (resp. observations $(x_u)_{u\in U}$). It is assumed that: $\boldsymbol{S}$ satisfies the Markovian factorization property associated with DAG $\mathcal{G}$, where the vertex set $\mathcal{V}$ is assimilated to the family of random variables $(S_v)_{v\in\mathcal{V}}$ (Lauritzen, 1996); the distribution of $\boldsymbol{S}$ is parametrized by the transition probabilities $p_{\boldsymbol{s}_{\mathrm{pa}(v)},k} = P(S_v = k|\boldsymbol{S}_{\mathrm{pa}(v)} = \boldsymbol{s}_{\mathrm{pa}(v)})$ and for the source vertices (vertices with no parent) u in $\mathcal{G}$, by the initial probabilities $(P(S_u = k))_k$; given $\boldsymbol{S}$, the random variables $(X_v)_v$ are independent and X_u is independent from $(S_v)_{v\neq u}$.

Usually, profiles of smoothed probabilities $(P(S_v = k|\boldsymbol{X} = \boldsymbol{x}))_{v\in\mathcal{V}}$ with $k = 0, \ldots, K-1$ have been used for quantifying state uncertainty. This approach suffers from two main shortcomings: as will be shown later, perception of state uncertainty associated with those profiles leads to overestimating global uncertainty of $\boldsymbol{S}$ given $\boldsymbol{X} = \boldsymbol{x}$. Moreover, visualization of those multidimensional profiles is made difficult by the graphical nature of arbitrary DAGs $\mathcal{G}$, provided that $K > 2$. In our approach, entropy H is considered as the canonical measure of uncertainty. Thus, $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ quantifies state process uncertainty given observations. This entropy can be decomposed into a sum of entropies. Every term of that sum is associated with one vertex in $\mathcal{V}$. Hence, these entropies can be interpreted as local contributions to global uncertainty. Since these profiles are unidimensional, they can be drawn whatever the graphical structure $\mathcal{G}$.

In what follows, this decomposition is made explicit. Then efficient algorithms are given in the HMC and HMT model cases to compute the elements of the decomposition. It is shown using synthetic and real-case data that the obtained local entropy profiles are relevant for state uncertainty diagnosis and state interpretation. These algorithms are complementary with approaches that enumerate the L most likely state restorations (so-called generalized Viterbi algorithm), and with approaches that compute profiles of alternative states to the most likely state process value. This so-called Viterbi forward–backward algorithm formally solves the optimization problem

$$(\arg)\max_{(s_u)_{u\neq v}} P((S_v = s_v)_{u\neq v}, S_v = k|\boldsymbol{X} = \boldsymbol{x}).$$

It is also shown that usual smoothed probability profiles are not relevant for quantifying global state uncertainty, due to their inherent marginalization property.

2 Conditional entropy profiles

Let $\boldsymbol{X}$ be a GHMM as defined in Section 1. It is assumed that the associated hidden state process $\boldsymbol{S}$ satisfies the factorization associated with the Markov property on $\mathcal{G}$:

$$\forall \boldsymbol{s},\ P(\boldsymbol{S}=\boldsymbol{s}) = \prod_{v\in\mathcal{V}} P(S_v = s_v | \boldsymbol{S}_{\mathrm{pa}(v)} = \boldsymbol{s}_{\mathrm{pa}(v)}), \tag{1}$$

where $P(S_v = s_v|\boldsymbol{S}_{\mathrm{pa}(v)} = \boldsymbol{s}_{\mathrm{pa}(v)})$ refers to $P(S_s = s_s)$ if $\mathrm{pa}(v) = \emptyset$.

The decomposition of entropy $H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x})$ comes from the conditional distribution of $\boldsymbol{S}$ given $\boldsymbol{X}=\boldsymbol{x}$ also satisfying the factorization property of $\mathcal{G}$:

$$P(\boldsymbol{S}=\boldsymbol{s}|\boldsymbol{X}=\boldsymbol{x}) = \prod_{v} P(S_v = s_v | \boldsymbol{S}_{\mathrm{pa}(v)} = \boldsymbol{s}_{\mathrm{pa}(v)}, \boldsymbol{X}=\boldsymbol{x}),$$

with the same convention as before if $\mathrm{pa}(v) = \emptyset$.

Proof. This property is proved by induction on the vertices of $\mathcal{G}$ (as would be proved factorization (1)). The random variables $(\boldsymbol{S}, \boldsymbol{X})$ satisfy the Markov property on DAG $\mathcal{G}'$ which edge set $\mathcal{E}'$ is defined as $a \in \mathcal{E}' \Leftrightarrow \{[a = (S_u, S_v) \text{ and } (u \in \mathrm{pa}(v))] \text{ or } a = (S_u, X_u)\}$. Let u in $\mathcal{G}$ be a sink vertex (vertex without children): then S_u is separated from $(S_v)_{v\neq u, v\notin \mathrm{pa}(u)}$ by $\boldsymbol{S}_{\mathrm{pa}(u)}$ in the moral graph of $\mathcal{G}'$. Thus, the following factorization holds:

$$P(\boldsymbol{S}=\boldsymbol{s}|\boldsymbol{X}=\boldsymbol{x}) = P(S_u = s_u|\boldsymbol{S}_{\mathrm{pa}(u)} = \boldsymbol{s}_{\mathrm{pa}(u)}, \boldsymbol{X}=\boldsymbol{x}) P((S_v)_{v\neq u} = (s_v)_{v\neq u}|\boldsymbol{X}=\boldsymbol{x}).$$

□

The additive decomposition of entropy is obtained by applying the chain rule (Cover & Thomas, 2006, chap. 2)

$$H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) = \sum_{v} H(S_v|\boldsymbol{S}_{\mathrm{pa}(v)}, \boldsymbol{X}=\boldsymbol{x}), \tag{2}$$

with the same convention as before if $\mathrm{pa}(v) = \emptyset$. As a consequence, the global state process uncertainty is decomposed as a sum of conditional entropies $(H(S_v|\boldsymbol{S}_{\mathrm{pa}(v)}, \boldsymbol{X}=\boldsymbol{x}))_{v\in\mathcal{V}}$, which define an entropy profile. Hence, each term of the sum is interpreted as a local uncertainty that contributes additively to global uncertainty.

In contrast, marginal entropies $(H(S_v|\boldsymbol{X}=\boldsymbol{x}))_{v\in\mathcal{V}}$ quantify uncertainty associated with smoothed probabilities $\xi_v(k) = P(S_v = k|\boldsymbol{X}=\boldsymbol{x})$ for $v \in \mathcal{V}$ and $0 \leq k < K$. These marginal entropies are upper bounds of the conditional entropies (Cover & Thomas (2006), chap. 2). Hence,

$$H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) \leq \sum_{v} H(S_v|\boldsymbol{X}=\boldsymbol{x}).$$

As a consequence, smoothed probability profiles do not represent uncertainty on the value of $\boldsymbol{S}$.

The particular case of HMC models is considered. Here $\mathcal{G}$ is a linear graph with T vertices, and for any $t < T$, $X_0 = x_0, \ldots, X_t = x_t$ is denoted by $X_0^t = x_0^t$. Here (2) can be rewritten as

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

with

$$H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}) = -\sum_{i,j} P(S_t = j, S_{t-1} = i|\boldsymbol{X} = \boldsymbol{x}) \log P(S_t = j|S_{t-1} = i, \boldsymbol{X} = \boldsymbol{x}).$$

This results from definition $H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}) = E[-\log P(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x})]$, where expectation is under $P(S_t, S_{t-1}|\boldsymbol{X} = \boldsymbol{x})$. The usual forward recursion computes $\alpha_t(j) = P(S_t = j|X_0^t = x_0^t)$ and $\gamma_t(j) = P(S_t = j|X_0^{t-1} = x_0^{t-1})$ for each time t and each state j and combine them in the backward recursion to yield the smoothed probabilities $\xi_t(j) = P(S_t = j|\boldsymbol{X} = \boldsymbol{x})$. Thus, computation of the conditional entropy profile $H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x})$ with $0 < t \leq T - 1$ can be integrated in the backward recursion by computing $P(S_t = j|S_{t-1} = i, \boldsymbol{X} = \boldsymbol{x}) = \xi_t(j)p_{ij}\alpha_{t-1}(i)/\{\gamma_t(j)\xi_{t-1}(i)\}$ where $p_{ij} = P(S_t = j|S_{t-1} = i)$ is the transition probability. This approach can be seen as an alternative to the algorithm of Hernando *et al.* (2005). It allows the computation of $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ with the same complexity in $\mathcal{O}(TK^2)$, but the advantage of our approach is to provide the conditional entropy profile.

In the case of HMTs indexed by tree $\mathcal{G} = \mathcal{T}$ the smoothed probabilities $\xi_v(k) = P(S_v = k|\boldsymbol{X} = \boldsymbol{x})$ are computed for $v \in \mathcal{T}$ by an upward–downward algorithm. A numerically stable iterative algorithm was proposed by Durand *et al.* (2004). It relies on an upward recursion, initialized at the leaf vertices of $\mathcal{T}$. The computed quantities are $\beta_v(k) = P(S_v = k|\bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v)$ and $\beta_{\mathrm{pa}(v),v}(k) = P(\bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v|S_{\mathrm{pa}(v)} = k)/P(\bar{\boldsymbol{X}}_v = \bar{\boldsymbol{x}}_v)$ for each vertex v and each state j, where $\bar{\boldsymbol{X}}_v$ denotes the subtree rooted in v. These quantities are computed as a function of β_u and $\beta_{\mathrm{pa}(u),u}$ for the children u of v. The algorithm complexity is in $\mathcal{O}(K^2)$ per iteration. The smoothed probabilities are computed using a downward recursion initialized at the root vertex of $\mathcal{T}$. In this recursion, the $\xi_v(k)$ are computed as a function of $\xi_{\mathrm{pa}(v)}$, β_v and $\beta_{\mathrm{pa}(v),v}$. The complexity is in $\mathcal{O}(K^2)$ per iteration as well. Similarly to the HMC case, adding the computation of

$$H(S_v|S_{\mathrm{pa}(v)}, \boldsymbol{X} = \boldsymbol{x}) = -\sum_{i,j} P(S_v = j, S_{\mathrm{pa}(v)} = i|\boldsymbol{X} = \boldsymbol{x}) \log P(S_v = j|S_{\mathrm{pa}(v)} = i, \boldsymbol{X} = \boldsymbol{x})$$

to the downward recursion, with $P(S_v = j|S_{\mathrm{pa}(v)} = i, \boldsymbol{X} = \boldsymbol{x}) = \beta_v(j)p_{ij}/\{P(S_v = j)\beta_{\mathrm{pa}(v),v}(i)\}$ and $p_{ij} = P(S_v = j|S_{\mathrm{pa}(v)} = i)$, allows for extracting conditional entropy profiles, while keeping the complexity per iteration of the algorithm in $\mathcal{O}(K^2)$.

3 Applications

Synthetic examples

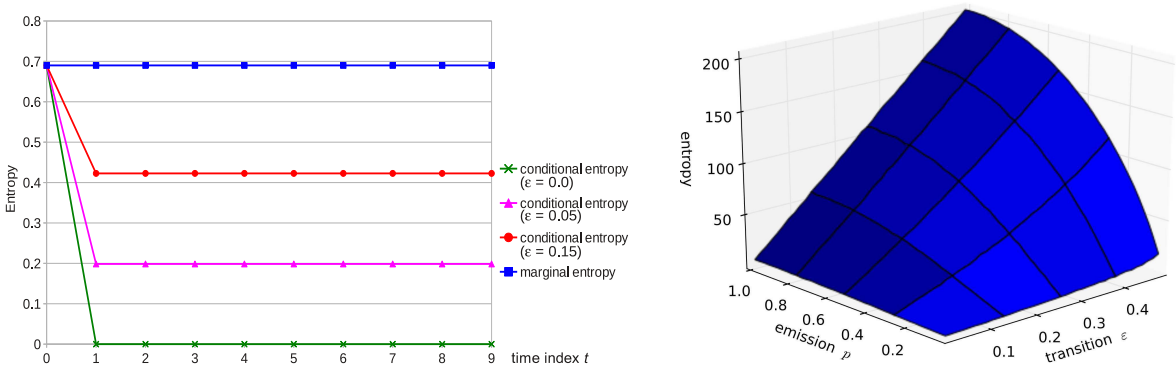
A two-state HMC family is considered. Its transition probability matrix is parametrized by $\varepsilon = P(S_t = 1|S_{t-1} = 0) = P(S_t = 0|S_{t-1} = 1)$, $\varepsilon \in [0, 0.5]$. The initial state distribution π is $P(S_0 = 0) = P(S_0 = 1) = 0.5$. The observation process takes values in $\{0, 1, 2\}$ and the emission distributions (conditional probabilities of observations given the states) are $P(X_t = 0|S_t = 0) = 1 - p$; $P(X_t = 1|S_t = 0) = p$; $P(X_t = 1|S_t = 1) = p$; $P(X_t = 2|S_t = 1) = 1 - p$ where $p \in [0, 1]$ is an additional parameter.

In a first experiment, p is fixed at 0.5 and the considered observed sequence is $x_t = 1$ for $t = 0, \ldots, T - 1$. The smoothed probabilities are $\xi_t(0) = \xi_t(1) = 0.5$ for $t = 0, \ldots, T - 1$. Thus, for any value of ε, marginal entropy is $\log 2$ and the sum of these entropies over t is $T \log 2$. In

contrast, global entropy of the hidden state sequence is a strictly increasing function of ε. Its minimum $\log 2$ is reached for $\varepsilon = 0$, whereas its maximum $T \log 2$ is reached for $\varepsilon = 0.5$.

Marginal and conditional entropy profiles are represented in Figure 1 a). For $\varepsilon = 0$, the conditional entropy profile is interpreted as follows: global uncertainty is $\log 2$, which corresponds to uncertainty concerning the first state only. Given this first state, every subsequent state is deterministic and does not contribute to global uncertainty. The marginal entropy profile highlights equiprobability of both states at each time t given the observations. The same statement would hold under an independent mixture assumption for $(X_t)_{t\geq 0}$. Marginal entropy results from uncertainty concerning state S_t due to observing $X_t = x_t$, but also to propagation of uncertainty from past states. As a consequence, marginal entropy cannot be interpreted in terms of local contributions to global uncertainty. In contrast, conditioning by the past state in entropy withdraws the effect of uncertainty propagation.

In a second experiment, the effect of p and ε on global state entropy is assessed by simulating 100 sequences of length $T = 300$ for each $p \in [0, 1]$ and each $\varepsilon \in [0, 0.5]$ on a regular grid with 40×40 points. The mean global entropy over the 100 sequences is represented in Figure 1 b). As expected, entropy increases with the emission distribution overlapping ($p \to 1$) and as the rows of the transition probability matrix tend to π ($\varepsilon \to 0.5$), so that maximal entropy $T \log 2$ is obtained in the independence case $\varepsilon = 0.5$ with full overlapping $p = 1.0$.



a) b)

Figure 1. a) Marginal and conditional entropy profiles for a 2-state HMC model with transition probabilities $\varepsilon = 0.0$, $\varepsilon = 0.05$ and $\varepsilon = 0.15$. b) Mean global state entropy for simulated sequences as a function of transition probability ε and emission probability p.

Analysis of the structure of Aleppo pines

The aim of this study was to build a model of the architectural development of Aleppo pines. The dataset contained seven branches of Aleppo pines, issued from different individuals. They were described at the scale of annual shoots v (segment of stem established within a year). Each branch was assimilated with a (mathematical) tree. Each tree vertex v (shoot) was characterized through one observed 5-dimensional vector X_v composed of the: number of growth cycles (from 1 to 3), presence of male sexual organs (binary variable), presence of female sexual organs (binary variable), length in cm, number of branches per tier. The parameters were estimated by maximum likelihood using the EM algorithm. The number of states was chosen by the ICL-BIC criterion (see Section 4), leading to selection of a 6-state HMT model. The Markov

tree is initialized in state 0 with probability one. A summary of the state transitions and an interpretation of the hidden states are provided in Figure 2.

As a first step, profiles of conditional entropies were represented using a colormap (mapping between entropy values and color intensities) – see Figure 3 a). This step highlighted location of the vertices with least ambiguous states along the branch main axes, and location of the vertices with most ambiguous states at the peripheral parts of branches. Then, state profiles were drawn along paths extending from the root vertex to leaf vertices. These paths were chosen so as to contain vertices with high conditional entropies. On the one hand, a detailed analysis of state uncertainty along the paths were obtained by Viterbi upward–downward profiles. This provided local alternative state values to the most likely tree states given by the Viterbi algorithm. On the other hand, the generalized Viterbi algorithm was used to characterize how clusters of neighbor vertices had simultaneous state changes in alternative state configurations. These results highlighted that the paths with most ambiguous states were composed of successions of unbranched, sterile shoots with one single growth cycle.

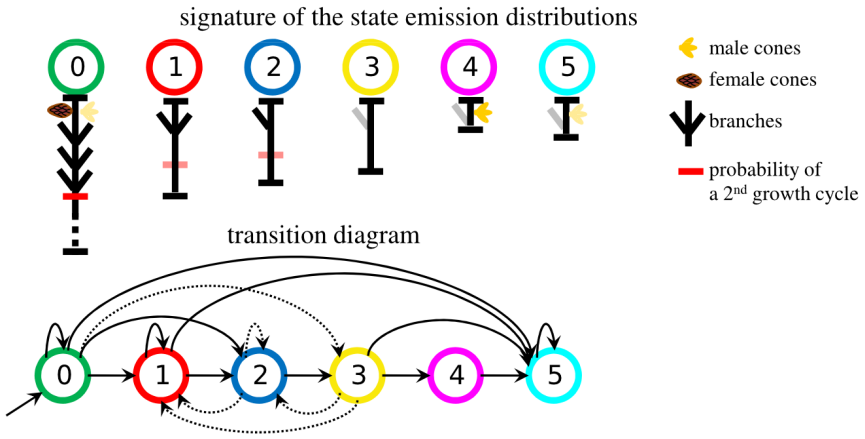


Figure 2. 6-state HMT model: transition diagram and symbolic representation of the state signatures (conditional mean values of the variables given the states, depicted by typical shoots). The separation between growth cycle is represented by a horizontal red segment, which intensity is proportional 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).

The application of this methodology is illustrated below on a path containing successive monocyclic, sterile shoots. This path belongs to the fourth individual (for which $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x}) = 47.5$). It is composed by 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 profiles in Figure 3c), it can be in states 2, 3 or 5. This is summarized by the conditional entropy profile in Figure 3b).

This conditional entropy profile can be further interpreted, in relation with mutual information $I(S_u; S_{\mathrm{pa}(u)}|\boldsymbol{X} = \boldsymbol{x})$. 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 an extension of (2) to subgraphs of $\mathcal{T}$, the contribution of the vertices of the considered path $\mathcal{P}$ to global state tree entropy can be computed as $\sum_{u\in\mathcal{P}} H(S_u|S_{\mathrm{pa}(u)}, \boldsymbol{X}=\boldsymbol{x})$ 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 3b).

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 3b). The mean marginal state entropy for this individual is 0.37 per vertex, which strongly overestimates the global state tree entropy per vertex (0.24).

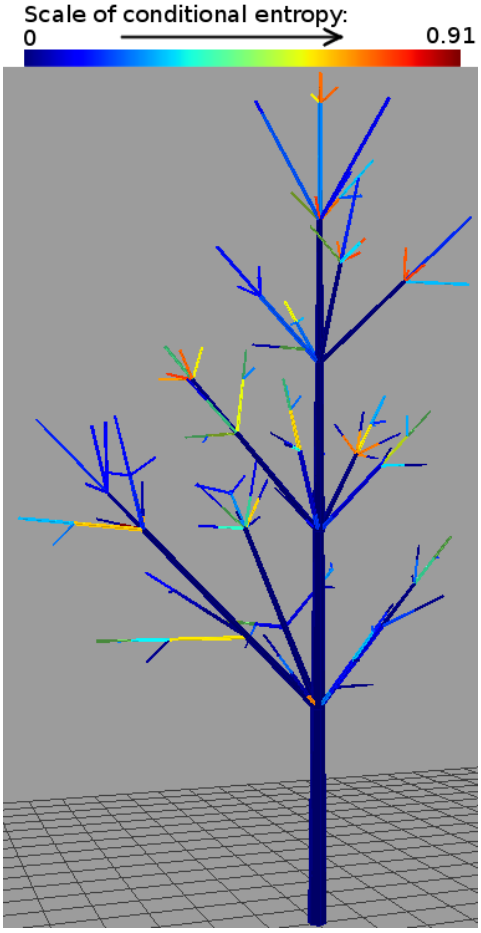


Figure 3 a)

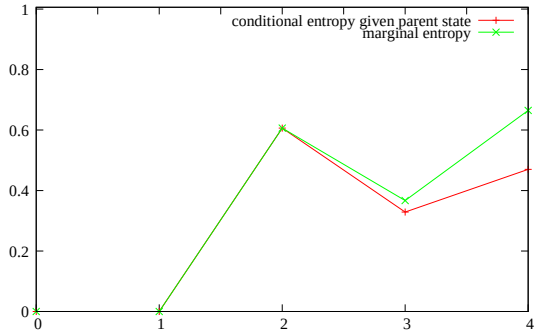


Figure 3 b)

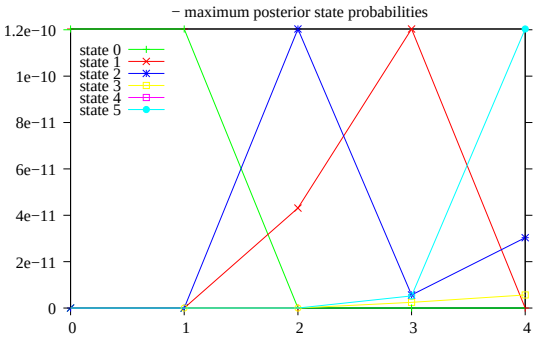


Figure 3 c)

Figure 3. a) Conditional entropy profiles $H(S_u|S_{\mathrm{pa}(u)}, \boldsymbol{X}=\boldsymbol{x})$ for vertices u associated with one of the seven branches. Blue corresponds to lowest and red to highest conditional entropies. b) Profiles of conditional and of marginal entropies along a path containing mainly sterile monocyclic shoots. c) State tree restoration with the Viterbi upward-downward algorithm.

4 Concluding remarks

In this work, conditional entropy profiles are proposed to assess both local and global state uncertainty in GHMMs. As shown in the examples, these profiles allow deeper understanding of the local roles of the model parameters, the neighbouring states and the observed data, concerning state uncertainty. These profiles are a valuable tool to analyse alternative state restorations, which may involve zones of connected vertices. Such situations are characterised by high mutual information between connected vertices. Moreover, the examples highlight 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 the perspective of model selection, entropy may also be useful. If irrelevant states or variables are added to GHMMs, global state entropy is expected to increase. This 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 introduced by McLachlan & Peel (2000, chap. 6). Their generalization to GHMMs 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.

Bibliography

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

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

[3] Durand, J.-B., Gonçalvès, P., and Guédon, Y. (Sept. 2004) *Computational Methods for Hidden Markov Tree Models – An Application to Wavelet Trees.* IEEE Transactions on Signal Processing, **52**, 9, 2551–2560.

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

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

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

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

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