Modèles d'arbre de Markov cachés

Yann Guédon[(1)] & Jean-Baptiste Durand[(2)]

[(1)] UMR CIRAD/CNRS/INRA/IRD/Université Montpellier II

Botanique et Bioinformatique de l'Architecture des Plantes

guedon@cirad.fr

[(2)] Laboratoire de Modélisation et Calcul

Institut d'Informatique et Mathématiques Appliquées de Grenoble

Jean-Baptiste.Durand@imag.fr

Contexte

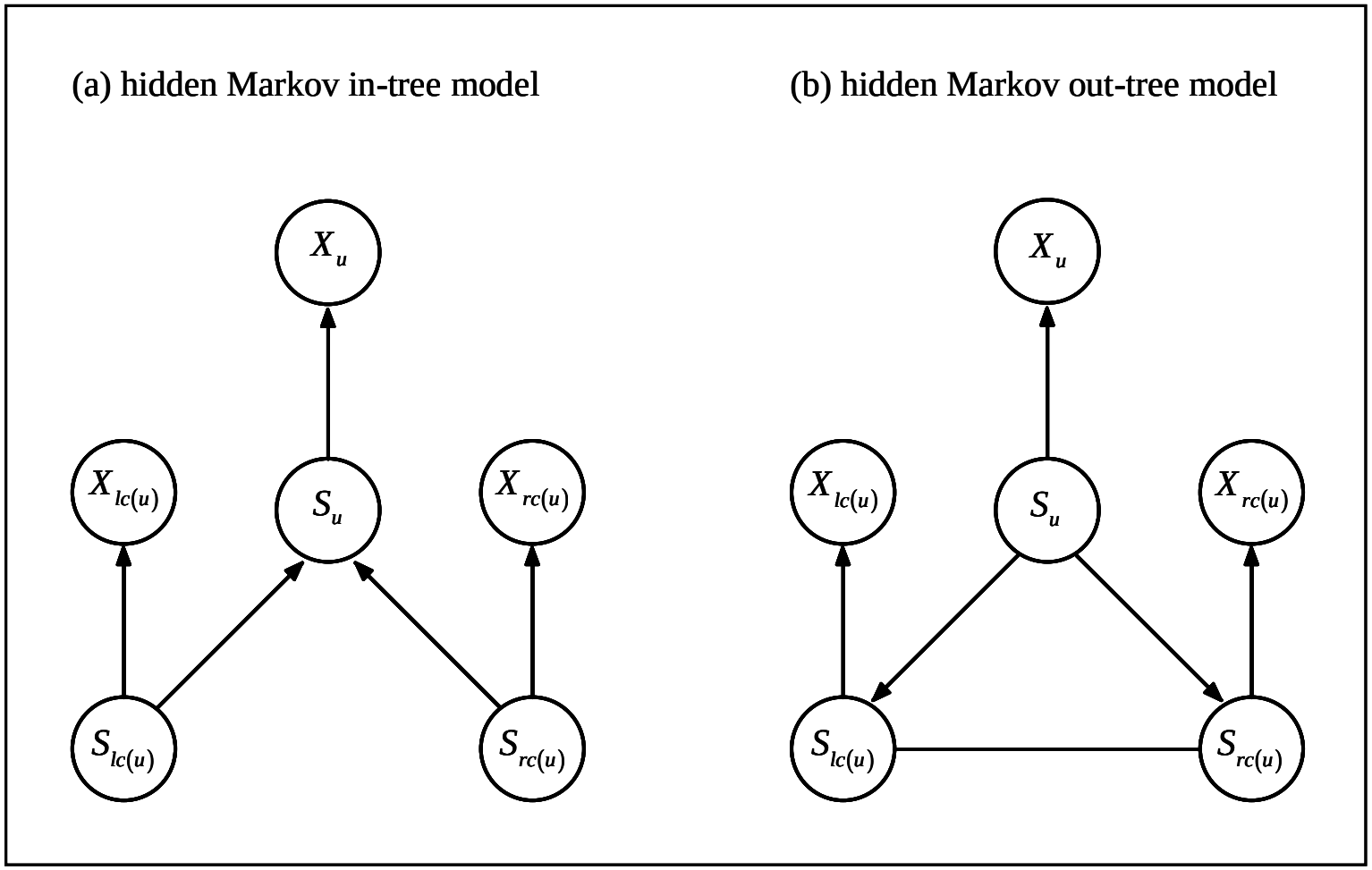
Arborescence : structure orientée non-symétrique,

2 cas :

- arborescence orientée depuis la racine,
- arborescence orientée depuis les feuilles.

Topologie de l'arborescence fixée par les données.

Objectif : identification de zones homogènes (de nature arborescente), de ruptures dans des arborescences.


(a) hidden Markov in-tree model
X_u
$X_{lc(u)}$
S_u
$X_{rc(u)}$
$S_{lc(u)}$
$S_{rc(u)}$
(b) hidden Markov out-tree model
X_u
$X_{lc(u)}$
S_u
$X_{rc(u)}$
$S_{lc(u)}$
$S_{rc(u)}$

Définition d'un modèle d'arbre de Markov orienté depuis les feuilles (Markov in-tree model)

Arborescence binaire où $lc(u)$ et $rc(u)$ désignent les deux vertex fils du vertex u et $b(u)$ le vertex frère de u.

Un modèle d'arbre de Markov binaire, orienté depuis les feuilles à J états est défini par

- probabilités initiales (pour les feuilles) $\pi_j = P(S_u = j)$,
- probabilités de transition $p_{hij} = P\left(S_u = j | S_{lc(u)} = h, S_{rc(u)} = i\right)$.

J^2 lois de transition et $J^2(J-1)$ probabilités de transition indépendantes comme pour une chaîne de Markov d'ordre 2.

Dans le cas où l'arborescence est non-ordonnée, les probabilités de transition sont invariantes par permutation des vertex fils

$$\begin{aligned} p_{hij} &= P\left(S_u = j | S_{lc(u)} = h, S_{rc(u)} = i\right) \\ &= P\left(S_u = j | S_{lc(u)} = i, S_{rc(u)} = h\right) \\ &= p_{ihj}. \end{aligned}$$

Définition d'un modèle d'arbre de Markov orienté depuis la racine (Markov out-tree model)

Nous supposons que

$$\begin{aligned} & P\left(S_{lc(u)} = k, S_{rc(u)} = m | S_u = j\right) \\ \neq \; & P\left(S_{lc(u)} = k | S_u = j\right) P\left(S_{rc(u)} = m | S_u = j\right). \end{aligned}$$

Un modèle d'arbre de Markov binaire, orienté depuis la racine à J états est défini par

- probabilités initiales (pour la racine) $\pi_j = P\left(S_0 = j\right)$,
- probabilités de transition $p_{jkm} = P\left(S_{lc(u)} = k, S_{rc(u)} = m | S_u = j\right)$.

J lois de transition et $J\left(J^2-1\right)$ probabilités de transition indépendantes.

Dans le cas où l'arborescence est non-ordonnée, les probabilités de transition sont invariantes par permutation des vertex fils

$$\begin{aligned} p_{jkm} &= P\left(S_{lc(u)}=k, S_{rc(u)}=m | S_u=j\right) \\ &= P\left(S_{lc(u)}=m, S_{rc(u)}=k | S_u=j\right) \\ &= p_{jmk}. \end{aligned}$$

Définition d'un modèle d'arbre de Markov orienté depuis la racine (Markov out-tree model)

Nous supposons que

$$\begin{aligned} & P\left(S_{lc(u)}=k, S_{rc(u)}=m|S_u=j\right) \\ = \; & P\left(S_{lc(u)}=k|S_u=j\right) P\left(S_{rc(u)}=m|S_u=j\right). \end{aligned}$$

Un modèle d'arbre de Markov, orienté depuis la racine à J états est défini par

- probabilités initiales (pour la racine) $\pi_j = P\left(S_0 = j\right)$,
- probabilités de transition $p_{jk} = P\left(S_{c(u)}=k|S_u=j\right)$.

J lois de transition et $J\left(J-1\right)$ probabilités de transition indépendantes comme pour une chaîne de Markov d'ordre 1.

Définition d'un modèle d'arbre de Markov caché

Le processus d'observation $\{X_u\}$ discret univarié est relié au modèle d'arbre de Markov caché $\{S_u\}$ par les probabilités d'observation (ou d'émission)

$$b_j(y) = P(X_u = y | S_u = j)$$

Extension directe au cas multivarié incluant éventuellement des processus d'observation continus (avec par exemple des lois d'observation gaussiennes).

Application de l'algorithme EM

$\longrightarrow$ généralisation directe du cas des chaînes de Markov cachées.

Espérance conditionnelle de la log-vraisemblance des données complètes :

$$Q(\theta|\theta^{(k)}) = E\left\{\log f\left(\overline{S}_0, \overline{X}_0; \theta\right) |\overline{X}_0 = \overline{x}_0; \theta^{(k)}\right\}.$$

Modèle d'arbre de Markov caché orienté depuis les feuilles

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= Q_\pi\left(\left\{\pi_j\right\}_{j=0}^{J-1}|\theta^{(k)}\right) + \sum_{h,i} Q_p\left(\left\{p_{hij}\right\}_{j=0}^{J-1}|\theta^{(k)}\right) \\ &\quad + \sum_j Q_b\left(\left\{b_j(y)\right\}_{y=0}^{Y-1}|\theta^{(k)}\right). \end{aligned}$$

avec

$$Q_\pi\left(\left\{\pi_j\right\}_{j=0}^{J-1}|\theta^{(k)}\right) = \sum_j \sum_{u\in\text{feuilles}} P\left(S_u = j|\overline{X}_0 = \overline{x}_0;\theta^{(k)}\right)\log\pi_j,$$

$$\begin{aligned} &Q_p\left(\left\{p_{hij}\right\}_{j=0}^{J-1}|\theta^{(k)}\right) \\ = &\sum_j \sum_{u\in\text{vertex internes}} P\left(S_u = j, S_{lc(u)} = h, S_{rc(u)} = i|\overline{X}_0 = \overline{x}_0;\theta^{(k)}\right)\log p_{hij}. \end{aligned}$$

Modèle d'arbre de Markov caché orienté depuis la racine

$$Q(\theta|\theta^{(k)}) = Q_\pi\left(\left\{\pi_j\right\}_{j=0}^{J-1}|\theta^{(k)}\right) + \sum_j Q_p\left(\left\{p_{jkm}\right\}_{k,m=0}^{J-1}|\theta^{(k)}\right) + \sum_j Q_b\left(\left\{b_j(y)\right\}_{y=0}^{Y-1}|\theta^{(k)}\right).$$

avec

$$Q_\pi\left(\left\{\pi_j\right\}_{j=0}^{J-1}|\theta^{(k)}\right) = \sum_j P\left(S_0 = j|\overline{X}_0 = \overline{x}_0;\theta^{(k)}\right)\log\pi_j,$$

$$Q_p\left(\left\{p_{jkm}\right\}_{k,m=0}^{J-1}|\theta^{(k)}\right) = \sum_{k,m}\sum_{u\in\text{vertex internes}} P\left(S_{lc(u)} = k, S_{rc(u)} = m, S_u = j|\overline{X}_0 = \overline{x}_0;\theta^{(k)}\right)\log p_{jkm}$$

Dans les deux cas, pour les probabilités d'observation

$$Q_b\left(\left\{b_j\left(y\right)\right\}_{y=0}^{Y-1}|\theta^{(k)}\right)=\sum_y\sum_u P\left(X_u=y,S_u=j|\overline{X}_0=\overline{x}_0;\theta^{(k)}\right)\log b_j\left(y\right).$$

Algorithme “montant-descendant” pour un modèle d’arbre de Markov caché orienté depuis les feuilles

Principe

$$
\begin{aligned}
& P\left(S_u = j | \overline{X}_0 = \overline{x}_0\right) \\
= \; & \frac{P\left(\overline{X}_{0\setminus u} = \overline{x}_{0\setminus u} | S_u = j\right)}{P\left(\overline{X}_{0\setminus u} = \overline{x}_{0\setminus u} | \overline{X}_u = \overline{x}_u\right)} P\left(S_u = j | \overline{X}_u = \overline{x}_u\right) \\
= \; & \alpha_u(j)\, \beta_u(j)\,.
\end{aligned}
$$

Récurrence montante - Initialisation (feuilles de l'arborescence)

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

Pour tous les vertex internes pris dans le sens montant,

$$\begin{aligned}\beta_u(j) &= P\left(S_u = j | \overline{X}_u = \overline{x}_u\right) \\ &= \frac{b_j(x_u)\sum_{h,i} p_{hij}\beta_{lc(u)}(h)\,\beta_{rc(u)}(i)}{N_u}.\end{aligned}$$

Le facteur de normalisation N_u est donné par

$$\begin{aligned} N_u &= P\left(X_u = x_u\right) \\ &= \sum_j b_j\left(x_u\right)\pi_j \end{aligned}$$

pour les feuilles, et

$$\begin{aligned} N_u &= P\left(X_u = x_u | \overline{X}_{c(u)} = \overline{x}_{c(u)}\right) \\ &= \sum_j b_j\left(x_u\right) \sum_{h,i} p_{hij} \beta_{lc(u)}\left(h\right) \beta_{rc(u)}\left(i\right) \end{aligned}$$

pour les vertex internes.

Récurrence descendante - Initialisation (racine de l'arborescence)

$$\alpha_0(j) = 1.$$

Pour tous les vertex restants pris dans le sens descendant,

$$\begin{aligned}
\alpha_u(j) &= \frac{P\left(\overline{X}_{0\setminus u} = \overline{x}_{0\setminus u} | S_u = j\right)}{P\left(\overline{X}_{0\setminus u} = \overline{x}_{0\setminus u} | \overline{X}_u = \overline{x}_u\right)} \\
&= \frac{\sum_m \alpha_{\rho(u)}(m)\, b_m\left(x_{\rho(u)}\right) \sum_k p_{jkm}\, \beta_{b(u)}(k)}{N_{\rho(u)}}.
\end{aligned}$$

Algorithme “montant-descendant” pour un modèle d’arbre de Markov caché orienté depuis la racine

Difficulté : récurrence montante dans le sens inverse de la structure.

$\longrightarrow$ Nécessité de calculer dans une passe préalable descendante les lois marginales $(P(S_u = j);\ j = 0, \ldots, J-1)$ pour chaque vertex u.

Récurrence montante - Initialisation (feuilles de l'arborescence)

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

Pour tous les vertex internes pris dans le sens montant,

$$\begin{aligned}\beta_u(j) &= P\left(S_u = j | \overline{X}_u = \overline{x}_u\right) \\ &= \left\{\sum_{k,m} \frac{\beta_{lc(u)}(k)\, \beta_{rc(u)}(m)\, p_{jkm}}{P\left(S_{lc(u)} = k\right) P\left(S_{rc(u)} = m\right)}\right\} \frac{b_j(x_u)\, P(S_u = j)}{N_u}.\end{aligned}$$

Le facteur de normalisation N_u est donné par

$$\begin{aligned} N_u &= P\left(X_u = x_u\right) \\ &= \sum_j b_j\left(x_u\right) P\left(S_u = j\right) \end{aligned}$$

pour les feuilles, et

$$\begin{aligned} N_u &= \frac{P\left(\overline{X}_u = \overline{x}_u\right)}{P\left(\overline{X}_{lc(u)} = \overline{x}_{lc(u)}\right) P\left(\overline{X}_{rc(u)} = \overline{x}_{rc(u)}\right)} \\ &= \sum_j \left\{ \sum_{k,m} \frac{\beta_{lc(u)}\left(k\right) \beta_{rc(u)}\left(m\right) p_{jkm}}{P\left(S_{lc(u)} = k\right) P\left(S_{rc(u)} = m\right)} \right\} b_j\left(x_u\right) P\left(S_u = j\right) \end{aligned}$$

pour les vertex internes.

Récurrence descendante - Initialisation (racine de l'arborescence)

$$\alpha_0(j) = 1.$$

Pour tous les vertex restants pris dans le sens descendant,

$$\begin{aligned}
\alpha_u(j) &= \frac{P\left(\overline{X}_{0\setminus u} = \overline{x}_{0\setminus u} | S_u = j\right)}{P\left(\overline{X}_{0\setminus u} = \overline{x}_{0\setminus u} | \overline{X}_u = \overline{x}_u\right)} \\
&= \frac{1}{N_{\rho(u)} P\left(S_u = j\right)} \sum_k \frac{\beta_{b(u)}(k)}{P\left(S_{b(u)} = k\right)} \\
&\quad \times \sum_i p_{ijk} b_i\left(x_{\rho(u)}\right) \alpha_{\rho(u)}(i) P\left(S_{\rho(u)} = i\right).
\end{aligned}$$

Cas des états des vertex fils conditionnellement indépendants (Durand *et al.*, 2004)

Récurrence montante

$$
\begin{aligned}
\beta_u(j) &= \left\{ \sum_{k,m} \frac{\beta_{lc(u)}(k)\,\beta_{rc(u)}(m)\,p_{jkm}}{P\left(S_{lc(u)}=k\right)P\left(S_{rc(u)}=m\right)} \right\} \frac{b_j(x_u)\,P(S_u=j)}{N_u} \\
&= \left\{ \sum_{k} \frac{\beta_{lc(u)}(k)\,p_{jk}}{P\left(S_{lc(u)}=k\right)} \right\} \left\{ \sum_{m} \frac{\beta_{rc(u)}(m)\,p_{jm}}{P\left(S_{rc(u)}=m\right)} \right\} \frac{b_j(x_u)\,P(S_u=j)}{N_u} \\
&= \frac{\beta_{u,lc(u)}(j)\,\beta_{u,rc(u)}(j)\,b_j(x_u)\,P(S_u=j)}{N_u}.
\end{aligned}
$$

Récurrence descendante

$$
\begin{aligned}
\alpha_u(j) &= \frac{1}{N_{\rho(u)} P(S_u = j)} \sum_k \frac{\beta_{b(u)}(k)}{P\left(S_{b(u)} = k\right)} \\
&\quad \times \sum_i p_{ijk} b_i\left(x_{\rho(u)}\right) \alpha_{\rho(u)}(i) P\left(S_{\rho(u)} = i\right) \\
&= \frac{1}{N_{\rho(u)} P(S_u = j)} \sum_i \left\{ \underbrace{\sum_k \frac{\beta_{b(u)}(k) p_{ik}}{P\left(S_{b(u)} = k\right)}}_{\beta_{\rho(u),b(u)}(i)} \right\} \\
&\quad \times p_{ij} b_i\left(x_{\rho(u)}\right) \alpha_{\rho(u)}(i) P\left(S_{\rho(u)} = i\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{P\left(S_u = j\right)} \sum_i p_{ij} \left\{ \underbrace{\frac{\beta_{\rho(u),u}\left(i\right) \beta_{\rho(u),b(u)}\left(i\right) b_i\left(x_{\rho(u)}\right) P\left(S_{\rho(u)} = i\right)}{N_{\rho(u)}}}_{\beta_{\rho(u)}(i)} \right\} \\
&\quad \times \frac{\alpha_{\rho(u)}\left(i\right)}{\beta_{\rho(u),u}\left(i\right)} \\
&= \frac{1}{P\left(S_u = j\right)} \sum_i \frac{p_{ij} \beta_{\rho(u)}\left(i\right) \alpha_{\rho(u)}\left(i\right)}{\beta_{\rho(u),u}\left(i\right)}.
\end{aligned}
$$

Complexité en $O\left(J^2 n\right)$ au lieu de $O\left(J^3 n\right)$ dans le cas dépendant.

Algorithme “arrière-avant” pour une chaîne de Markov cachée

Principe

$$
\begin{aligned}
& P\left(S_t = j | X_0^{\tau-1} = x_0^{\tau-1}\right) \\
= & \frac{P\left(X_0^{t-1} = x_0^{t-1} | S_t = j\right)}{P\left(X_0^{t-1} = x_0^{t-1} | X_t^{\tau-1} = x_t^{\tau-1}\right)} P\left(S_t = j | X_t^{\tau-1} = x_t^{\tau-1}\right) \\
= & \alpha_t(j)\,\beta_t(j)\,.
\end{aligned}
$$

Récurrence arrière - $t = \tau - 1$:

$$\begin{aligned}\beta_{\tau-1}(j) &= P\left(S_{\tau-1} = j | X_{\tau-1} = x_{\tau-1}\right) \\ &= \frac{b_j(x_{\tau-1})\, P\left(S_{\tau-1} = j\right)}{N_{\tau-1}}.\end{aligned}$$

$t = \tau - 2, \ldots, 0$:

$$\begin{aligned}\beta_t(j) &= P\left(S_t = j | X_t^{\tau-1} = x_t^{\tau-1}\right) \\ &= \left\{\sum_k \frac{\beta_{t+1}(k)\, p_{jk}}{P\left(S_{t+1} = k\right)}\right\} \frac{b_j(x_t)\, P\left(S_t = j\right)}{N_t}.\end{aligned}$$

Le facteur de normalisation N_t est donné par

$$\begin{aligned} N_{\tau-1} &= P\left(X_{\tau-1} = x_{\tau-1}\right) \\ &= \sum_j b_j\left(x_{\tau-1}\right) P\left(S_{\tau-1} = j\right) \end{aligned}$$

pour $t = \tau - 1$

$$\begin{aligned} N_t &= \frac{P\left(X_t^{\tau-1} = x_t^{\tau-1}\right)}{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1}\right)} \\ &= \sum_j \left\{ \sum_k \frac{\beta_{t+1}(k)\, p_{jk}}{P\left(S_{t+1} = k\right)} \right\} b_j\left(x_t\right) P\left(S_t = j\right) \end{aligned}$$

pour $t < \tau - 1$.

Récurrence avant - $t = 0$:

$$\alpha_0(j) = 1.$$

$t = 1, \ldots, \tau - 1$:

$$\begin{aligned}
\alpha_t(j) &= \frac{P\left(X_0^{t-1} = x_0^{t-1} | S_t = j\right)}{P\left(X_0^{t-1} = x_0^{t-1} | X_t^{\tau-1} = x_t^{\tau-1}\right)} \\
&= \frac{1}{N_{t-1} P(S_t = j)} \sum_i p_{ij} b_i(x_{t-1}) \alpha_i(t-1) P(S_{t-1} = i).
\end{aligned}$$

Remarques :

– Tous les vertex de la sous-arborescence $\overline{x}_u$ sont visités avant sa racine x_u dans la récurrence montante,
– Certains des vertex de $\overline{x}_{0\setminus u}$ n'ont pas encore été visités dans la récurrence descendante à l'étape u.

Dans la première passe obligatoire ("avant" ou "arrière" dans le cas de séquences, "montante" dans le cas d'arborescences) - qui peut être précédée par une passe préalable de calcul des lois marginales suivant la direction de la structure - il est seulement possible de fusionner de l'information.

$\longrightarrow$ Ceci implique que le demi-degré intérieur (nombre d'arcs entrants) ou extérieur (nombre d'arcs sortants) de chaque vertex est égal à 1.

Pour l'application à l'analyse de la structure des plantes

Objectif : limiter le nombre de probabilités de transition indépendantes.

- Vertex fils non-ordonnés ou partiellement ordonnés (vertex fils correspondant à l'entité suivante - construite par le même méristème - distingué des autres vertex fils correspondant aux entités portées),

- Succession d'états transitoires et état final absorbant (variable d'état ordinale) ; ordre de grandeur du nombre d'états : 5,

- Prise en compte uniquement des entités jouant un rôle dans la construction de la structure de la plante (par opposition aux entités jouant principalement un rôle d'assimilation),

– Combinaison des probabilités de transition estimées du type $p_{hij} = P\left(S_u = j | \; S_{c(u)} = \{h, i\}\right)$ et $p_{ghij} = P\left(S_u = j | \; S_{c(u)} = \{g, h, i\}\right)$ dans le cas de modèles d'arbre de Markov cachés orientés depuis les feuilles, ou $p_{jkm} = P\left(S_{c(u)} = \{k, m\} \mid S_u = j\right)$ et $p_{jkmn} = P\left(S_{c(u)} = \{k, m, n\} \mid S_u = j\right)$ dans le cas de modèles d'arbre de Markov cachés orientés depuis la racine.

Les modèles d'arbre de Markov cachés orientés depuis les feuilles permettent de prendre en compte des arborescences où seules les parties périphériques de la plante sont observables du fait de l'effacement des marqueurs morphologiques lié à l'élagage et à la croissance cambiale (en diamètre).