Quantification de l'incertitude sur le processus d'états dans des modèles de Markov cachés

Jean-Baptiste DURAND[1] Yann GUÉDON[2]

[1]Laboratoire Jean Kuntzmann et INRIA, Mistis (Grenoble)

[2]CIRAD, UMR AGAP et INRIA, Virtual Plants (Montpellier)

Journées MSTGA, 7 juin 2012

Plan

- Modèles de Markov cachés et restauration des états.
- Quantification de l'incertitude : profils d'états et entropie.
- Algorithmes de calcul de l'entropie par décomposition.
- Conclusion et perspective en sélection de modèles.

Modèles de Markov cachés

Chaînes de Markov cachées

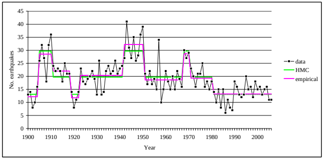


FIGURE: Nombre annuel de tremblements de terre

- Processus observé $(X_t)_{t=0,1,\ldots}$ et caché $(S_t)_{t=0,1,\ldots}$.
- $(S_t)_{t=0,1,\ldots}$ chaîne de Markov à J états, de matrice de transition $P = (p_{ij})_{i,j}$.
- X_t conditionnellement indépendante de toutes les autres variables sachant $S_t = j$, de loi $b_j(x_t)$ (exemple : gaussienne multivariée $\mathcal{N}(\mu_j, \Sigma_j)$).

Modèles de Markov cachés graphiques

- Processus indexé par les sommets d'un DAG $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ (sommets et arcs), fixé ($\mathcal{G}$ pas aléatoire).
- Processus observé $\boldsymbol{X} = (X_u)_{u \in \mathcal{U}}$ et caché $\boldsymbol{S} = (S_u)_{u \in \mathcal{U}}$.
- $\boldsymbol{S}$ satisfait une propriété de Markov sur $\mathcal{G}$ (indépendance conditionnelle entre S_u et ses non-descendants sachant les parents $\boldsymbol{S}_{\mathrm{pa}(u)}$).

Propriété (Factorisation)

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

(facteur $P(S_u = s_u)$ si $\mathrm{pa}(u) = \emptyset$).

- X_u conditionnellement indépendante de toutes les autres variables sachant $S_u = j$, de loi $b_j(x_u)$.

Exemple : arbres de Markov cachés

- $\mathcal{G} = \mathcal{T}$ une arborescence de racine $u = 0$.
- Enfants $\bar{\boldsymbol{S}}_{c(u)}$ conditionnellement indépendants sachant S_u.
- Application : pousses u d'une plante caractérisées par X_u – longueur, nombre de feuilles, de fleurs, etc.
- Interprétation des états : types de pousses représentatifs de différents stades de croissance.

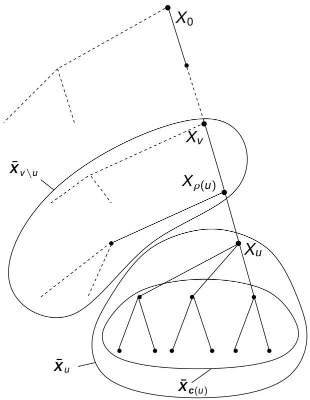


FIGURE: Arborescence et notations

Restauration des états cachés

Importance de pouvoir «estimer» (*restaurer*) les états :

- Interprétation du phénomène modélisé.
- Utilisation des états dans un post-traitement (exemple : alignement de séquences, arborescences, ...).
- Diagnostic du modèle («histogrammes conditionnels»).

Définition (Restauration suivant le principe du MAP)

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} P(\boldsymbol{S} = \boldsymbol{s} | \boldsymbol{X} = \boldsymbol{x})$$

est appelé graphe d'états restauré suivant le principe du Maximum A Posteriori (MAP).

Calcul par un algorithme dit de Viterbi (programmation dynamique).
Utilisation pertinente des états restaurés $\Leftrightarrow$ restauration peu ambigüe.

Restauration : illustration

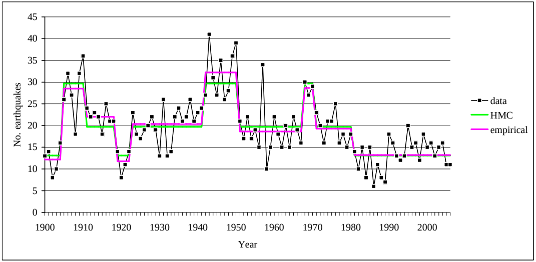


FIGURE: Nombre annuel de tremblements de terre

- Chaîne de Markov cachée à lois conditionnelles de Poisson.
- Instants de sauts déterminés par l'algorithme de Viterbi.
- Niveau des sauts représentés : paramètres des lois de Poisson (vert) / moyennes empiriques pour chaque valeur d'état restauré (rose).

N.B. Différence exliquée par la méthode d'estimation (algorithme EM).

Profils d'états

Probabilités lissées pour les chaînes de Markov cachées

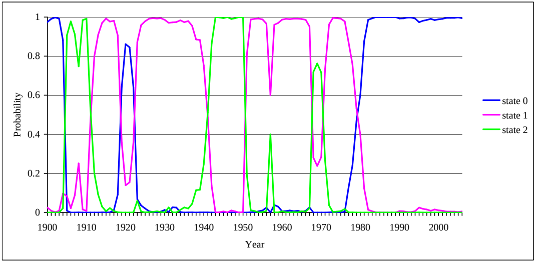


FIGURE: Profil d'états : probabilités lissées

- Probabilités lissées : $P(S_t = j | \boldsymbol{X} = \boldsymbol{x})$, représentées en fonction de t (une courbe par état).
- Tableau $T \times J$ calculé par des récursions avant-arrière de complexité $\mathcal{O}(TJ^2)$ [Ephraim and Merhav(2002)].
- Représentation «honnête» de l'incertitude sur $\boldsymbol{S}$?
- Transposition à des graphes ?

Profils de Viterbi

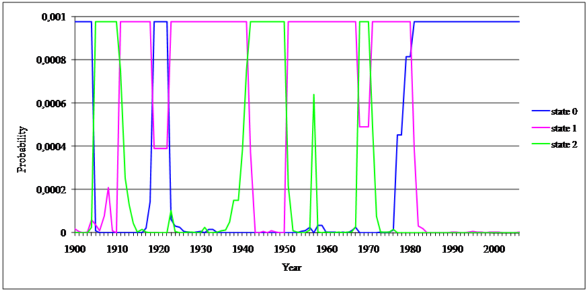


FIGURE: Profil d'états : Viterbi

$$\max_{(s_{t'})_{t' \neq t}} P((S_{t'} = s_{t'})_{t' \neq t}, S_t = j | \boldsymbol{X} = \boldsymbol{x}),$$

représentées en fonction de t (une courbe par état).

- Interprétation : probabilité des restaurations possibles avec $S_t = j$.
- Calcul en $\mathcal{O}(TJ^2)$ [Brushe et al.(1998)].

L meilleures séquences

rang	proba	S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ell = 1$	0.4	0	0	0	0	1	1	1	1	0	0	0	0
$\ell = 2$	0.15	0	0	0	0	0	0	0	0	0	0	0	0
$\ell = 3$	0.02	0	0	0	0	0	1	1	1	0	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$\ell-$ième solution du problème

$$\arg\max_{\boldsymbol{s}} P(\boldsymbol{S} = \boldsymbol{s} | \boldsymbol{X} = \boldsymbol{x}).$$

- Incertitude sur les états de $t = 4$ à 7.
- Mise en évidence par un profil de lissage.
- $S_4 \Rightarrow S_5, S_6, \ldots$ (voire : $S_7 \Rightarrow S_6, S_5, \ldots$).
- Type d'information écrasé par le profil de lissage.

Profils pour arbres de Markov cachés

- Divers algorithmes (lissage, Viterbi, etc.) transposables aux arbres de Markov cachés [Durand et al.(2004)].
- Complexité $\mathcal{O}(TJ^2)$ toujours.
- Particularité des arbres : récursion descendante dépendante de la récursion ascendante.
- Visualisation de profils sur aborescences : pénible !
- Pertinence des profils de lissage ? (idem chaîne de Markov cachée).
- Illustration : Viterbi, profils de lissage, ...
  Représentation par niveau de couleurs.
  Valeurs bleues les plus faibles, valeurs rouges les plus élevées.

Entropie

- S variable aléatoire à valeurs dans $\{0, \ldots, J-1\}$.

Définition (Entropie)

$$H(S) = -E[\log P(S)] = -\sum_j P(S=j) \log P(S=j) \geq 0.$$

- $H(S) = 0$ si et seulement si S suit une loi de Dirac.
- $H(S) = \log J$ si et seulement si S suit la loi uniforme.
- Entropie : mesure canonique de l'incertitude.

Entropies marginale et totale.

- Quantification de l'incertitude des états $\boldsymbol{S}$ au vu de $\boldsymbol{x}$ (les états restaurés reflètent-ils bien $\boldsymbol{S}$?)
- Entropie marginale au sommet u (ou à l'instant u) : entropie

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

de la loi $P(S_u|\boldsymbol{X}=\boldsymbol{x})$.
- Entropie totale de $\boldsymbol{S}$:

$$H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) = -\sum_{\boldsymbol{s}} P(\boldsymbol{S}=\boldsymbol{s}|\boldsymbol{X}=\boldsymbol{x}) \log P(\boldsymbol{S}=\boldsymbol{s}|\boldsymbol{X}=\boldsymbol{x}) \qquad (J^T \text{ termes}).$$

Entropies marginale et totale.

- Entropie marginale $H(S_u|\boldsymbol{X} = \boldsymbol{x})$ facile à calculer en $\mathcal{O}(TJ^2)$. Peut être représentée même sur des arborescences.
- Mais : perception d'une **incertitude surestimée** par le profil d'entropie marginale $H(S_u|\boldsymbol{X} = \boldsymbol{x})$, $u = 0, 1, \ldots$.

$$\sum_u H(S_u|\boldsymbol{X} = \boldsymbol{x}) \geq H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$$

  (inégalité très large en général).
- Traduit le fait que la connaissance d'un état réduit (en cascade) l'incertitude sur les autres états.
- Problème : quelle quantité traduit une incertitude localisée, qui soit une contribution locale à $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$?

Décomposition de l'entropie dans les modèles de Markov cachés graphiques (MMCG)

Processus d'états conditionnellement markovien dans les MMCG, au sens où

Proposition

Soit $(\boldsymbol{S}, \boldsymbol{X})$ un MMCG par rapport au DAG $\mathcal{G}$. Alors pour tout $\boldsymbol{x}$, la loi de $\boldsymbol{S}$ sachant $\boldsymbol{X} = \boldsymbol{x}$ vérifie la propriété de Markov sur $\mathcal{G}$ et pour tout $\boldsymbol{s}$,

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

où $P(S_u = s_u | \boldsymbol{S}_{pa(u)} = \boldsymbol{s}_{pa(u)}, \boldsymbol{X} = \boldsymbol{x})$ représente $P(S_u = s_u | \boldsymbol{X} = \boldsymbol{x})$ si $pa(u) = \emptyset$.

Formule de décomposition de l'entropie totale

Par application d'une règle classique («règle de chaînage») de calcul d'entropie d'une loi factorisée, on obtient ce

Corollaire

Pour tout $\boldsymbol{x}$,

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

où $H(S_u|\boldsymbol{S}_{pa(u)}, \boldsymbol{X}=\boldsymbol{x})$ représente $H(S_u|\boldsymbol{X}=\boldsymbol{x})$ si $pa(u)=\emptyset$.

On en déduit au passage :

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

(surestimation de l'incertitude par le profil d'entropies marginales.)

Profils d'entropie : conclusion

Profil d'entropie $H(S_u|\boldsymbol{S}_{\mathrm{pa}(u)}, \boldsymbol{X} = \boldsymbol{x})$, $u = 0, 1, \ldots$: représentation de l'incertitude relative à $\boldsymbol{S}$ sans amplification.

- $H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x})$ pour les chaînes.
- $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ pour les arbres (un seul parent).
- Intérêt supplémentaire en pratique : comprendre la restauration des états (voir application).
- Comment les calculer efficacement ?
- Quid de $H(S_{t-1}|S_t, \boldsymbol{X} = \boldsymbol{x})$ pour les chaînes et $H(S_u|S_{c(u)}, \boldsymbol{X} = \boldsymbol{x})$ pour les arbres ?

Algorithme de Hernando *et al. (2005)*

Un algorithme dû à [Hernando et al.(2005)] permet de calculer récursivement les

$$H(S_0, \ldots, S_t | S_{t+1} = j, X_0 = x_0, \ldots, X_{t+1} = x_{t+1}) = H(S_0^t | S_{t+1} = j, X_0^{t+1} = x_0^{t+1}) :$$

- pour les chaînes de Markov cachées ;
- en $\mathcal{O}(TJ^2)$;
- sans faire apparaître les $H(S_t | S_{t-1}, \boldsymbol{X} = \boldsymbol{x})$;
- noter le conditionnement par l'état futur (surprenant) ;
- permet tout de même de calculer $H(\boldsymbol{S} | \boldsymbol{X} = \boldsymbol{x}) = H(S_0^{T-1} | \boldsymbol{X} = \boldsymbol{x})$.

Calcul des $H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x})$ pour les chaînes de Markov cachées

On note $L_t(j) = P(S_t = j|\boldsymbol{X} = \boldsymbol{x})$.

- Calcul en utilisant l'algorithme de Hernando *et al. (2005)*

$$H\left(S_0^t|\boldsymbol{X} = \boldsymbol{x}\right) = \sum_j L_t\left(j\right) H\left(S_0^{t-1}|S_t = j, X_0^t = x_0^t\right) + H\left(S_t|\boldsymbol{X} = \boldsymbol{x}\right).$$

puis

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

(profils d'entropie partielle).

Variante pour le calcul des $H(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x})$

- Calcul direct des $H\left(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}\right)$ en utilisant les probabilités avant et arrière.
- Puis utiliser

$$H\left(S_0^t|\boldsymbol{X} = \boldsymbol{x}\right) = \sum_{r=0}^{t} H(S_r|S_{r-1}, \boldsymbol{X} = \boldsymbol{x}).$$

Conditionnement par l'état passé

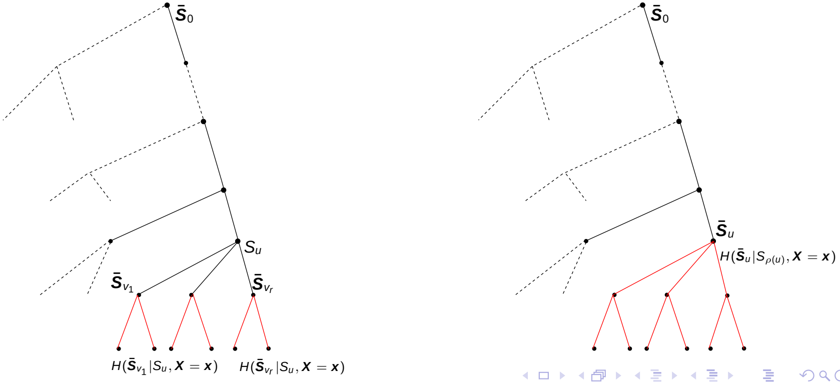
- Possibilité d'un calcul parcourant la séquence dans le sens contraire (cf. analogie arbres de Markov cachés).
- Calcul de $H(S_{t+1}^{T-1}|S_t = j, X_{t+1}^{T-1} = x_{t+1}^{T-1})$ par un algorithme *à la Hernando*, puis $H\left(S_t^{T-1}|\boldsymbol{X} = \boldsymbol{x}\right)$ et enfin $H\left(S_t|S_{t+1}, \boldsymbol{X} = \boldsymbol{x}\right)$.
- Calcul de $H\left(S_t|S_{t-1}, \boldsymbol{X} = \boldsymbol{x}\right)$ et $H\left(S_t|S_{t+1}, \boldsymbol{X} = \boldsymbol{x}\right)$: quantifier l'apport d'information provenant des états passé **et** futur.

Transposition aux arbres de Markov cachés

- Calcul des $H(S_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$ en utilisant les $L_u(j)$.
- Conditionnement cohérent avec l'orientation de l'arborescence.
- Incorporation dans un calcul récursif (*ascendant*) des $H(\bar{\boldsymbol{S}}_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$.
- Incorporation avec les entropies marginales dans le calcul d'entropies partielles $H(\bar{\boldsymbol{S}}_u | \boldsymbol{X} = \boldsymbol{x})$.
- Variante possible, comme pour les chaînes de Markov cachées (*à la Hernando*).
- Calcul possible des $H(\bar{\boldsymbol{S}}_{0\setminus u} | \boldsymbol{X} = \boldsymbol{x})$ par une récursion descendante.
- Calculs en $\mathcal{O}(TJ^2)$.

Schéma de la récursion ascendante

$$H(\bar{\boldsymbol{S}}_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}) = \sum_{v \in c(u)} H(\bar{\boldsymbol{S}}_v | S_u, \boldsymbol{X} = \boldsymbol{x}) + H(S_u | S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x}).$$



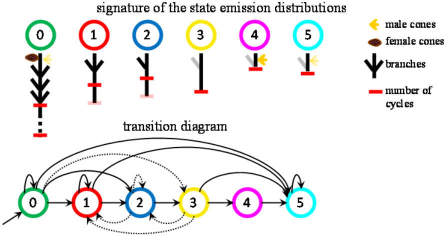
Profils conditionnés par les enfants

- Quantifier également l'apport d'information provenant de l'amont.
- Calcul des $H(S_u | \mathbf{S}_{c(u)}, \mathbf{X} = \mathbf{x})$ en $\mathcal{O}(TJ^{c+1})$, c nombre maximal d'enfants.
- Pas d'interprétation comme contribution locale à l'entropie totale :

$$\sum_u H(S_u | \mathbf{S}_{c(u)}, \mathbf{X} = \mathbf{x}) \geq H(\mathbf{S} | \mathbf{X} = \mathbf{x}).$$

Données et modèle

- 7 branches issues de 7 arbres différents.
- 836 pousses annuelles.
- Variables : longueur, nombre de branches, polycyclisme, cônes mâles, femelles.
- Étude d'un modèle à 6 états.



- Pousses monocycliques, stériles, courtes et non ramifiées : forte incertitude sur l'état.

état	«proba»
0	0.001
2	0.261
3	0.367
5	0.371

Profils d'entropie et d'états

- Localisation de l'incertitude : $H(S_u|S_{\rho(u)}, \boldsymbol{X} = \boldsymbol{x})$
- Visualisation des profils le long de chemins.
- Pousse mâle (entropie du chemin : 0.09 = 0.02/sommet).
- Pousse femelle (entropie du chemin : 0.48 = 0.08/sommet).
- Pousse stérile (entropie du chemin : 1.41 = 0.28/sommet).

modèle

Écarts entre profils

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

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

Comparaison de $C(\mathcal{T})$ et $M(\mathcal{T})$ avec la référence $G(\mathcal{T})$:

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

Perspectives :

sélection de variables

- Idée : l'ajout de variables pertinentes par rapport aux états diminue l'entropie.
- L'ajout de variables indépendantes des états :
  - ne change pas l'entropie à paramètres connus ;
  - augmente l'entropie si les paramètres sont estimés.
- Vérifié empiriquement sur des variables simulées.
- Pénaliser la log-vraisemblance par l'entropie ?

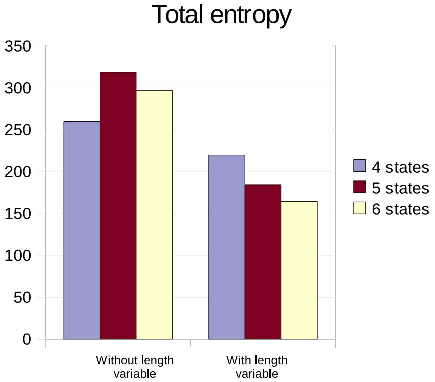


FIGURE: Entropie totale suivant le nombre d'états, et l'inclusion ou non de la variable *longueur*.

Perspectives :

sélection du nombre d'états J

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

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

d_J nombre de paramètres algébriquement indépendants du modèle à J états.

Critère	Nombre d'états J			
	4	5	6	7
BIC	-10,545	-10,558	**-10,541**	-10,558
NEC	0.48	0.37	**0.32**	0.46
ICL-BIC	-10,764	-10,742	**-10,704**	-10,814

TABLE: Sélection du nombre d'états cachés par les critères d'information BIC, NEC et ICL-BIC, sur le jeu de données « pins d'Alep ».

(ICL-BIC défini par [McLachlan and Peel(2000)], chap. 6.)
(NEC défini par [Celeux and Soromenho(1996)], avec cas particulier pour $J = 1$.)

Conclusion

- Profils standards de lissage : plutôt néfastes en général.
- Privilégier les profils d'entropie $H(S_u|\boldsymbol{S}_{\mathrm{pa}(u)}, \boldsymbol{X} = \boldsymbol{x})$, $u = 0, 1, \ldots$
- Algorithmes à intégrer sans surcoût excessif aux récursions usuelles.
  ⚠ Si on conditionne dans le bon sens !
- Entropie totale $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ calculable en une seule récursion.
- À compléter par des profils le long de chemins : entropie(s), Viterbi, ...

Bibliographie

NEC pour $J = 1$

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

est défini pour des modèles gaussiens dans le cas où $J = 1$ par « le ratio de l'entropie sous un modèle avec mêmes moyennes et variances différentes, et de la différence de la log-vraisemblance sous ce modèle et de celle d'un modèle à $J = 1$ état.»

retour

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

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

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

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

J.-B. Durand, P. Gonçalvès, and Y. Guédon.

Computational methods for hidden Markov tree models – an application to wavelet trees.
IEEE Transactions on Signal Processing, 52(9) :2551–2560, September 2004.

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

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

G.J. McLachlan and D. Peel.
Finite Mixture Models.
Wiley Series in Probability and Statistics. John Wiley and Sons, 2000.