Chaînes et champs de Markov cachées

Gilles Celeux et Yann Guédon
Inria Rhône-Alpes et Cirad

Remerciements à Jean-Baptiste Durand, Florence Forbes et Nathalie Peyrard

Modèle à structure cachée

$$\begin{array}{lccccccccc}
\text{données observées} & y_1 & & y_2 & & y_3 & & \dots & & y_n \\
 & \uparrow & & \uparrow & & \uparrow & & \dots & & \uparrow \\
\text{données cachées} & z_1 & \to & z_2 & \to & z_3 & \to & \dots & \to & z_n
\end{array}$$

- Les données observées $\mathbf{y} = (y_1, \dots, y_n)$ suivent une loi de paramètre θ et sont conditionnellement indépendantes sachant $\mathbf{z} = (z_1, \dots, z_n)$.
- les z_i sont à valeurs discrètes (classes) et soit sont indépendantes (mélange) soit constituent un processus markovien (chaine 1D, champ 2D).

La loi marginale d'une donnée y s'écrit

$$f(y) = \sum_{\ell=1}^{K} p_\ell \phi(y|\theta_\ell)$$

les $p_\ell, \ell = 1, \dots, K$ définissant la loi marginale de z.

Chaînes de Markov cachées

Soit un processus observé $y_1, ..., y_n$ à valeurs dans $\mathbf{R}^d$ et un processus caché $z_1, ..., z_n$ à valeurs dans $\{1, ..., k\}$ tels que

- les z_i forment une chaîne de Markov homogène, irréductible, de distribution stationnaire $(p_\ell, \ell = 1, \ldots, k)$
- les y_i sont indépendants conditionnellement aux z_i,
- si $z_i = \ell, y_i$ suit une loi de densité $\phi(y_i|\theta_\ell)$.

Le problème

- estimer $\theta_\ell, \ell = 1, \ldots, k$,
- caractériser le processus markovien : estimer la matrice de transition de la chaîne de Markov $\mathbf{T} = (t_{j\ell}, j = 1, ..., k; \ell = 1, ..., k)$ et sa probabilité stationnaire $(p_\ell, \ell = 1, ..., k)$,
- restaurer $\mathbf{z}$ au vu de $\mathbf{y}$: classification.

Domaines d'application

Les domaines d'application des CMC qui sont en fait un modèle de mélange avec une dépendance markovienne de la suite des étiquettes des composants sont nombreux.

- La reconnaissance de la parole est l'un des domaines-rois d'utilisation des CMC (Rabiner 1989, Guédon et Cocozza 1990).
- Dans la foulée, d'autres problèmes de reconnaissance de formes se traitent très bien par les CMC, comme la reconnaissance de gestes, par exemple, Durand et Martin (1999).
- Les CMC ont également connu un grand succès pour l'analyse des séquences génomiques (voir Churchill 1989, Muri 1998 par exemple).
- La modélisation de croissance de plantes (Guédon 1998)
- La modélisation des courbes de consommation électriques (action en cours avec EDF), la fiabilité des logiciels (action en cours avec le LMC),. . .

Un exemple de modélisation

Les courbes de consommation électriques des ménages ou des entreprises peuvent être considérés comme régies par une chaîne de Markov cachée.

- Les états cachées sont le type d'activité d'un ménage (sommeil, préparation de repas, moment de loisir, . . .)
- On suppose qu'à l'intérieur de l'un de ces états les variations aléatoires de consommation électrique sont indépendantes.
- Les états cachées expliquent toutes les fluctuations importantes de consommation.
- Le passage d'un état à l'autre se fait par une chaîne de Markov.

Une fois le modèle identifié, il permet facilement

- de faire de la prévision à relativement court terme,
- de simuler des courbes de consommation réalistes.

Méthodes d'estimation

Le modèle CMC est un modèle à données incomplètes. Les données complètes sont $\mathbf{x} = (x_1, ..., x_n)$, de densité $g(\mathbf{x}|\eta)$, où $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, avec :

- données observées : $\mathbf{y} = (y_1, ..., y_n)$, de densité $f(\mathbf{y}|\eta)$
- données manquantes : $\mathbf{z} = (z_1, ..., z_n)$, de densité conditionnelle $k(\mathbf{z}|\mathbf{y}, \eta) = \frac{g(\mathbf{x}|\eta)}{f(\mathbf{y}|\eta)}$

Données incomplètes $\rightarrow$ Développement d'algorithmes de restauration-maximisation. Partant de η^0, deux étapes :

- Restauration : "donner" des valeurs $\mathbf{z}^m$ aux données manquantes, sachant η^m.

- Maximisation : trouver η^{m+1} qui maximise la vraisemblance complétée $g(\mathbf{x}^m|\eta)$, où $\mathbf{x}^m = (\mathbf{y}, \mathbf{z}^m)$.

Stratégies de restauration

Les valeurs manquantes sont remplacées par:

- leur espérance conditionnelle sachant $\mathbf{y}$ et η^m : EM calculer à chaque itération η^{m+1} qui maximise $E[\log\ g(\mathbf{x}|.)|\mathbf{y}, \eta^m]$
  $\rightarrow$ CMC : algorithme de Baum-Welch.
- leur valeur la plus probable sachant $\mathbf{y}$ et η^m : MAP (maximum a posteriori).
  $\rightarrow$ CMC : algorithme de Viterbi.
- des tirage(s) aléatoire(s) suivant la loi conditionnelle $k(\mathbf{z}|\mathbf{y}, \eta^m)$: SEM $\rightarrow$ un tirage; MCEM $\rightarrow$ moyenne de r tirages.
  $\rightarrow$ CMC : algorithme EM à la Gibbs.

Un théorème rassurant

Théorème (Leroux 1992) *Si le mélange est identifiable, si sa matrice de transition* $\mathbf{T}$ *est irréductible, et sous des conditions de régularité classiques (à la Wald), la suite des estimateurs du maximum de vraisemblance tend p.s. vers les vraies valeurs des paramètres du modèle des CMC.*

L'algorithme *forward-backward* de Baum-Welch

Étape E l'étape de restauration revient à calculer les quantités

$$D_{j\ell}(i) = P(z_{i+1} = \ell, z_i = j \mid \mathbf{y}, \eta)$$

$\eta = (\mathbf{T}, \theta_1, \ldots, \theta_k)$ étant l'estimation courante des paramètres du modèle.
Ces calculs reposent sur l'égalité :

$$D_{j,\ell}(i) = \frac{F_i(j) t_{j\ell} \phi(y_{i+1}|\theta_\ell) B_{i+1}(\ell)}{f(\mathbf{y} \mid \eta)}$$

avec

$$F_i(j) = P(y_1, \ldots, y_i, z_i = j \mid \eta),$$

$$B_i(\ell) = P(y_{i+1}, \ldots, y_n \mid z_i = \ell, \eta)$$

et $f(\mathbf{y} \mid \eta) = \Sigma_{j=1}^{k} F_n(j)$.

La procédure avant-arrière

Le calcul des F_i se fait par la procédure avant suivante :

a) $F_1(\ell) = p_\ell \phi(y_1 \mid \theta_\ell)\ \ (\ell = 1, \ldots, k)$

b) $F_{i+1}(\ell) = \phi(y_{i+1} \mid \theta_\ell)\, \Sigma_{j=1}^{k}\, F_i(j) t_{j\ell}$.

Le calcul des B_i se fait par la procédure arrière suivante :

a) $B_n(j) = 1\ \ (j = 1, \ldots, k)$.

b) $B_i(j) = \Sigma_{\ell=1}^{k}\, t_{j\ell} \phi(y_{i+1} \mid \theta_\ell) B_{i+1}(\ell)$.

Difficulté numérique : problèmes d' "underflow" lors de la propagation des calculs.

Parade

Travailler sur des probabilités conditionnelles :

$$D_{j\ell}(i) = \frac{\tilde{F}_i(j) t_{j\ell} \tilde{B}_{i+1}(\ell)}{P(y_{i+1}|y_1,\ldots,y_i)}$$

avec

$$\tilde{F}_i(j) = P(z_i = j|y_1,\ldots,y_i,\eta)$$

et

$$\tilde{B}_i(\ell) = \frac{P(y_{i+1},\ldots,y_n|z_i=\ell,\eta)}{P(y_{i+1},\ldots,y_n|y_1,\ldots,y_i,\eta)}.$$

Et les $\tilde{F}_i$ et les $\tilde{B}_i$ se calculent par des procédures avant et arrière analogues aux précédentes.

Étape M : l'étape de maximisation conduit à calculer

$$t_{j\ell}^{m+1} = \frac{\Sigma_{i=1}^{n-1} D_{j\ell}^{m}(i)}{\Sigma_{i=1}^{n-1} \Sigma_{r=1}^{k} D_{jr}^{m}(i)}$$

Et (cas gaussien)

$$\mu_j^{m+1} = \sum_{i=1}^{n} w_j^{m+1}(i) y_i / \sum_{i=1}^{n} w_j^{m+1}(i)$$

où $w_j^{m+1}(i) = \Sigma_{\ell=1}^{k} D_{j\ell}^{m+1}(i)$.

...

Enfin, les proportions du mélange sont estimées par

$$p_j^{m+1} = \frac{1}{n} \sum_{i=1}^{n} w_j^{m+1}(i).$$

Algorithme de Viterbi

Étape C : l'étape de restauration est une étape de classification. Les sommes de la procédure avant-arrière sont remplacées par des max. On construit $\mathbf{z}^m$ par le principe du MAP : z_i maximise $f(z_i|\mathbf{y}, z_{j\neq i})$. Cela donne pour $1 < i < n$

$$z_i^m = \text{arg} \max_{\ell=1,\ldots,k} t^m_{z^m_{i-1}\ell}\phi(y_i|\theta^m_\ell)t^m_{\ell z^{m-1}_{i+1}},$$

et

$$z_1^m = \text{arg} \max_{\ell=1,\ldots,k} \phi(y_2|\theta^m_\ell)t^m_{\ell z^{m-1}_2},$$

$$z_n^m = \text{arg} \max_{\ell=1,\ldots,k} t^m_{z^m_{n-1}\ell}\phi(y_n|\theta^m_\ell).$$

Étape M : les termes de la matrice D^{m+1} sont calculés par la formule

$$D^{m+1}_{j\ell} = \frac{1}{n}\sum_{i=2}^{n} \mathbf{I}_{\{z^m_{i-1}=j,\ z^m_i=\ell\}}$$

D'où l'actualisation de la matrice de transition $\mathbf{T}$, des θ_ℓ et des proportions p_ℓ par les mêmes formules que pour l'algorithme EM.

EM à la Gibbs

But : proposer un algorithme stochastique qui évite de tomber dans le premier point fixe rencontré sans alourdir les calculs.
Moyen : Adapter l'échantillonnage de Gibbs au cadre fréquentiel pour simuler les données manquantes $\mathbf{z}$.

Étape de restauration : les z_i^m sont simulés selon la loi de probabilité conditionnelle de z_i sachant tout le reste. Pour $1 < i < n$

$$\begin{aligned} f(z_i|\mathbf{y}, \eta, z_{j\neq i}) &= f(z_i|y_i, \eta, z_{i-1}, z_{i+1}) \\ &\propto t_{z_{i-1},z_i}\, \phi(y_i|\theta_{z_i})\, t_{z_i,z_{i+1}}, \end{aligned}$$

et

$$\begin{aligned} f(z_1|\mathbf{y}, \eta, z_{j>1}) &\propto \phi(y_n|\theta_{z_n}) t_{z_1,z_2}, \\ f(z_n|\mathbf{y}, \eta, z_{j<n}) &\propto t_{z_{n-1},z_n} \phi(y_n|\theta_{z_n}). \end{aligned}$$

Étape de maximisation : L'actualisation de η^m est identique à celle de l'algorihme de Viterbi (ou de l'algorithme SEM). Une précaution numérique :

$$p_{j,\ell}^{(m+1)} = \frac{\alpha + \Sigma_{i=2}^{n} \mathbf{I}_{\{z_{i-1}^m = j\}} \mathbf{I}_{\{z_i^m = \ell\}}}{\Sigma_{i=2}^{n} \mathbf{I}_{\{z_{i-1}^m = j\}} + n\alpha}$$

où α est ici un nombre positif à choisir petit devant n. De plus, cette modification dans le calcul de $\mathbf{T}^m$ permet d'assurer le théorème suivant.

Théorème : *La chaîne de Markov* $(\mathbf{z}^m)$ *engendrée par l'algorithme EM à la Gibbs est uniformément géométriquement ergodique et possède une distribution stationnaire unique.*

Estimation bayésienne

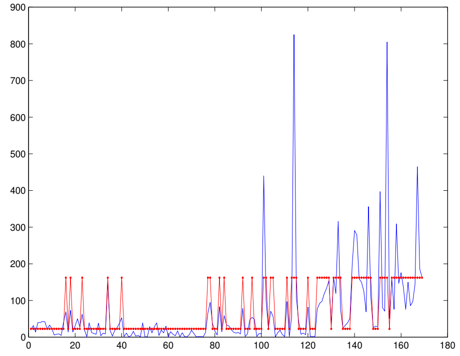
- L'identification par l'échantillonnage de Gibbs ne pose pas de problème particulier. La simulation de $\mathbf{z}$ se fait comme dans l'algorithme EM à la Gibbs. La simulation des paramètres se fait par un formulaire analogue à celui d'un mélange indépendant (Robert, Celeux et Diebolt 1993).

- Dans le cadre de l'analyse génomique, Muri (1998) a étudié le cas de mélanges de lois discrètes binomiales régis par une CMC.

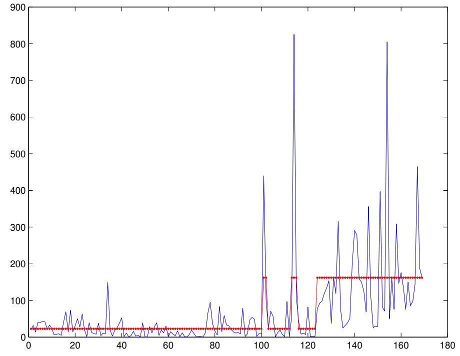
- La procédure de *reversible jump* a été étudié par Robert, Rydén et Titterington (2000) dans le cadre de CMC unidimensionnellles pour un mélange gaussien dont les composants ont des moyennes égales et connues.

Application en fiabilité des logiciels

Les données observées sont les temps inter-défaillances d'un logiciel qui subit des corrections. Le modèle utilisé est une CMC à deux états, dont les lois conditionnelles sachant l'état caché sont des lois exponentielles. Le modèle a été estimé par l'algorithme EM.



L'état caché le plus probable, superposé aux données



La chaîne cachée la plus probable, superposée aux données

Analogie avec les modèles à espace d'états

Une décomposition utile

$$\begin{aligned} L_i(j) &= P(z_i = j|\mathbf{y}, \eta) \\ &= \frac{P(y_{i+1}, \dots, y_n|z_i = j, \eta)}{P(y_{i+1}, \dots, y_n|y_1, \dots, y_i, \eta)} P(z_i = j|y_1, \dots, y_i, \eta) \\ &= \widetilde{B}_i(j)\, \widetilde{F}_i(j) \end{aligned}$$

avec
$\widetilde{F}_i(j)$: probabilité filtrée,
$L_i(j)$: probabilité lissée.

Procédure avant seule : algorithme de filtrage
$\rightarrow$ permet de calculer la vraisemblance des données $P(y_1, \dots, y_n, \eta)$,
Procédure avant-arrière : algorithme de lissage.

Analyse des motifs de ramification de branches de cyprès

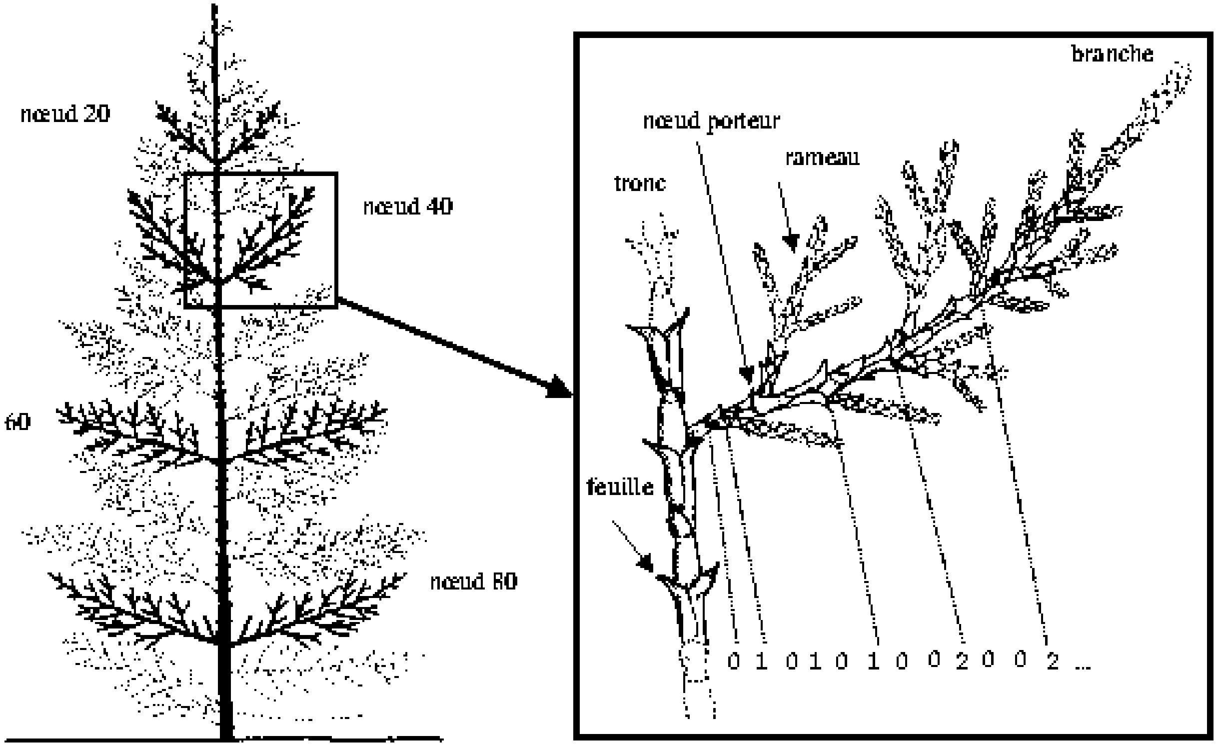
Contexte expérimental

Quatre paires de branches sont mesurées noeud par noeud de la base au sommet sur 35 cyprès du Tassili agés de 3 ans. La phyllotaxie est de 2 (2 feuilles opposées par noeud) et le nombre de rameaux portés est donc inférieur ou égal à 2.

Données

- 276 séquences de longueur cumulée 7192.
- 3 valeurs possibles : 0, 1 ou 2 rameaux portés
- Les motifs 101, 2001 et 2002 sont les motifs les plus fréquents.

Données expérimentales



©Yves Caraglio

Résultats

Estimation de l'ordre de la chaîne de Markov à 3 états

ordre	paramètres	$2\log L_r$	BIC(r)
0	2	-12175.6	-12193.3
1	5	-9885.1	-9929.5
2	13	-9164	-9279.4
3	31	-8842.8	-9118.1
4	56	-8642.6	-9139.9
5	98	-8466.8	-9337.1

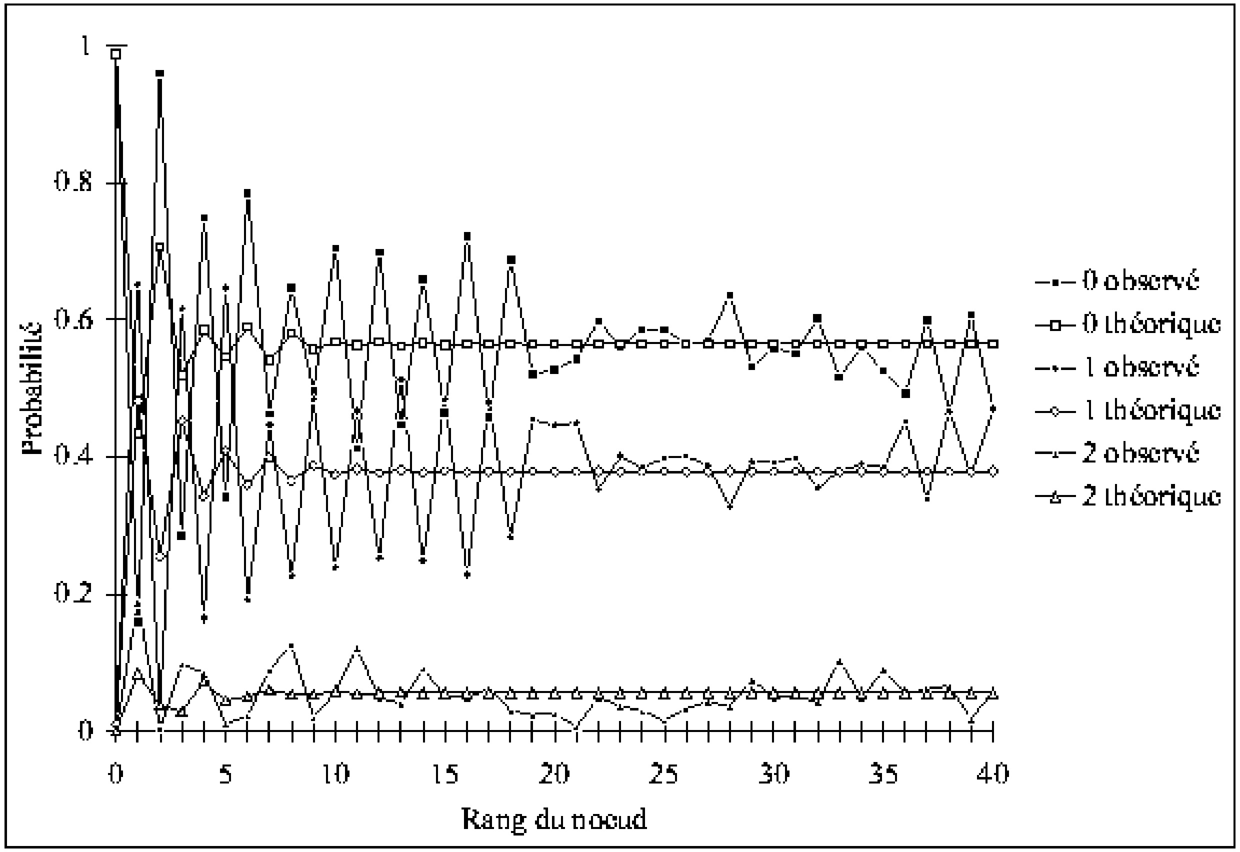
Mauvais ajustement des lois marginales fonction du paramètre d'index $\longrightarrow$ estimation d'une chaîne de Markov cachée d'ordre 3 (en fait bruitée car l'espace d'observations $=$ l'espace d'états).

Comparaison des deux chaînes

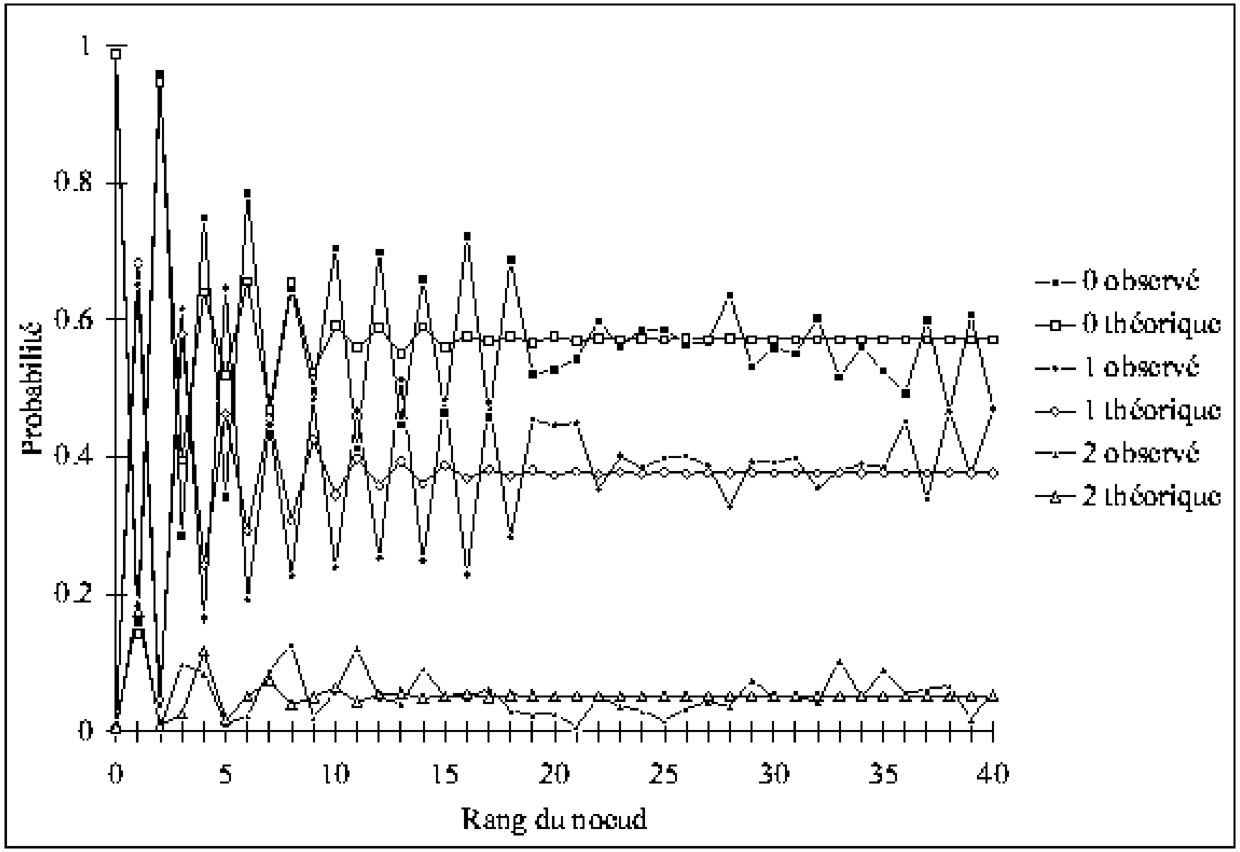
	paramètres	$2\log L$	BIC
chaîne de Markov	31	-8842.8	-9118.1
chaîne de Markov cachée	32	-8655.9	-8940

La différence de BIC (> 10) entre la chaîne de Markov et la chaîne de Markov cachée permet de sélectionner ce dernier modèle.

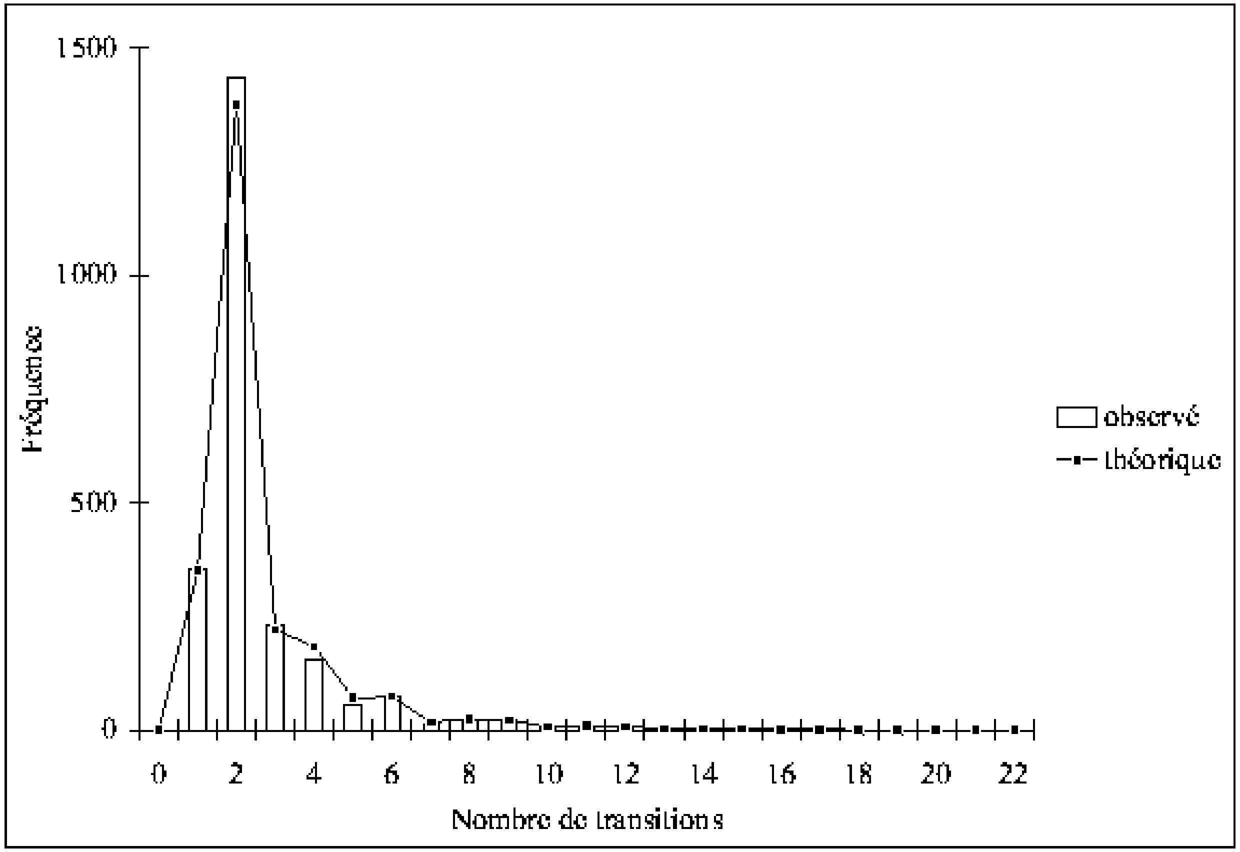
Chaîne de Markov d'ordre 3, Ajustement des lois marginales fonction du paramètre d'index



Chaîne de Markov cachée d'ordre 3, Ajustement des lois marginales fonction du paramètre d'index



Chaîne de Markov cachée d'ordre 3, Ajustement de la loi du temps de retour pour 1 rameau



Interprétation

Matrice des probabilités d'observation de la chaîne de Markov cachée d'ordre 3

$$\widehat{B} = \begin{pmatrix} \mathbf{1} & 0 & 0 \\ 0.15 & \mathbf{0.83} & 0.02 \\ 0.05 & 0.16 & \mathbf{0.79} \end{pmatrix}$$

Le bruit traduit dans la matrice $\widehat{B}$ peut être interprété comme le résultat de l'avortement précoce de certains rameaux qui n'ont donc pas atteint une taille suffisante pour être observables a posteriori.

L'utilisation d'une chaîne de Markov cachée permet de séparer :

- l'information structurelle, traduite dans les probabilités de transition,
- l'information relative au “bruit” traduite dans les probabilités d'observation.

“Débruitage” des séquences observées

```
0 1 0 1 0 1 0 2 0 0 2 0 1 0 0 2 0 0 0 0 1 0 0 2 0 0 1 0 1
                                    1

0 0 0 1 0 0 0 0 2 0 0 1 0 0 2 0 0 2 0 0 1 0 1 0 1 0 0 1 0 1
  1       1           2

0 1 0 2 0 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 1 0 1 0 1
      1

0 2 0 0 1 0 0 2 0 1 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 1 0 1 2 0 0 0 0 0 1
        2                   1                 1                         2
```

Modèles statistiques d'analyse d'images

Image : champ aléatoire bruité sur l'ensemble des sites (pixels). L'image observée est la réalisation d'un champ $Y = \{Y_s, i \in S\}$, $Y_i \in R^d$. L'image à restaurer définit un champ $Z = \{Z_i, i \in S\}$, $Z_i \in V = \{v_1, \ldots, v_k\}$.
L'hypothèse est que $Y = R(Z, B)$, où $B = \{B_i, i \in S\}$ est un bruit.
On note $Z = \{Z_{ij}, i \in S\}$, où $Z_{ij} = 1$ si le pixel $i \in v_j$ et 0 sinon. La réalisation z de Z est inconnue.
Segmentation d'image : Le problème est d'estimer z au vu de y.
Deux stratégies sont en concurrence.

- Le maximum a posteriori (MAP) qui consiste à maximiser $P(Z = z|y)$. Cette stratégie correspond à un coût 0-1 et produit une bonne restitution de la scène, mais est très difficile à réaliser.
- La maximisation des probabilités marginales (MPM) qui consiste à maximiser pour tout i $P(Z_i = z_i|y)$. Cette stratégie conduit à minimiser le nombre de pixels mal classés.

Champ de Markov

$Z = \{Z_i,\ i \in S\}$ sur S sites, $Z_i \in V$ fini, $N(i)$ voisins de i

$$\begin{aligned} \forall z, \quad P_G(z_i \mid z_{S\setminus\{i\}}) &= P_G(z_i \mid z_j,\ j \in N(i)) \\ \forall z, \quad P_G(z) &> 0. \end{aligned}$$

Théorème de Hammersley-Clifford (La connaissance des loi conditionnelles permet de retrouver la loi conjointe)

$$P_G(z) = \frac{\exp(-H(z))}{W} \text{ (distribution de Gibbs)}$$

où $H(z)$, fonction d'énergie, est une somme de fonctions sur les pixels et les combinaisons de pixels voisins (les cliques) si ces lois conditionnelles sont strictement positives pour tout z ($\rightarrow$ voisinages symétriques).
W constante de normalisation (fonction de partition) incalculable : elle s'écrit comme somme sur toutes les images autorisées par le modèle de $\exp(-H(\mathbf{z}))$.
Mais les lois conditionnelles ne dépendent pas de W $\rightarrow$ d'où l'approximation de $P_G(z)$ par l'échantillonnage de Gibbs.

Champ de Markov caché

Observations, mesures, $Y = \{Y_i, i \in S\}$

données cachées, affectation à des classes, Z champ de Markov caché (CMC), $P_G(z|\beta)$

$$Y|Z = z \quad \rightarrow f(y \mid z, \theta) = \prod_{i \in S} f_i(y_i \mid z_i, \theta)$$

Vraisemblance complète

$$\begin{aligned} P_G(y, z \mid \theta, \beta) &= f(y \mid z, \theta)\, P_G(z \mid \beta) \\ &= W(\beta)^{-1} \exp\{-H(z \mid \beta) + \sum_{i \in S} \log f_i(y_i \mid z_i,\ \theta)\} \end{aligned}$$

$\Rightarrow$ $Z|Y = y$ est également un champ de Markov

$\Rightarrow$ Utilisation des principes d'approximation des champs de Markov

Le modèle de Potts

$Z = \{Z_i, i \in S\}$ est un modèle de Potts à k couleurs si :
$z_i \in \{0,1\}^k$ avec $z_{i,\ell} = 1$ ssi $i \in$ classe ℓ

$$P_G(z|\beta) = W(\beta)^{-1} \exp\left(\beta \sum_{i \sim j} z_i^t z_j\right)$$

$Y|Z = z$ admet la loi de densité $f(y_i|z_i, \theta) = \mathcal{N}(\ell, \sigma_\ell^2)$, si $i \in$ classe ℓ

- Le paramètre β est porteur de la dépendance spatiale.
- Le modèle de Potts dépend du système de voisinage choisi, classiquement 4 voisins ou plus communément 8 voisins.
- On peut considérer aussi un modèle de Potts avec un terme sur les singletons

$$P_G(z|\beta) = W(\beta)^{-1} \exp\left(- \sum_{i \in S} z_i^t a_i + \beta \sum_{i \sim j} z_i^t z_j\right)$$

EM pour champs de Markov cachés ? (Estimation de $\Psi = (\theta, \beta)$)

Champ de Markov caché Z
Champ de Markov conditionnel $Z|Y = y$

(E)

$$\begin{aligned}
Q(\Psi \mid \Psi^{(q)}) &= E_{\Psi^{(q)}}[\log P_G(y, Z \mid \Psi) \mid Y = y] \\
&= E_{\Psi^{(q)}}[\log P_G(y \mid Z, \theta) \mid Y = y] + E_{\Psi^{(q)}}[\log P_G(Z \mid \beta) \mid Y = y] \\
&= \underbrace{\sum_{i \in S} \sum_{z_i} \log f_i(y_i \mid z_i, \theta)\ P_G(z_i \mid y, \Psi^{(q)})}_{Q(\theta \mid \Psi^{(q)})} \underbrace{- \log W(\beta) - \sum_c \sum_{z_c} V_c(z_c|\beta)\ P_G(z_c \mid y, \Psi^{(q)})}_{Q(\beta \mid \Psi^{(q)})}
\end{aligned}$$

(M) $$\Psi^{(q+1)} = (\theta^{(q+1)}, \beta^{(q+1)}) = \arg\max_{\Psi} Q(\Psi \mid \Psi^{(q)})$$

Difficultés :

- Calcul de $P_G(z_i \mid y, \Psi^{(q)})$, $P_G(z_c \mid y, \Psi^{(q)})$ et de $W(\beta)$ dans l'étape **E**.
- Estimer β dans l'étape **M**.

La pseudo-vraisemblance

Une approximation classique (Besag 1975) pour traiter le problème d'estimation de la fonction de partition $W(\beta)$ est d'approximer $P_G(z|\beta)$ par la pseudo-vraisemblance

$$\mathcal{PL}(z) = \prod_{i \in S} P_G(z_i \mid z_{N(i)}). \tag{1}$$

Chaque terme dans ce produit est facile à calculer.
Mais $\mathcal{PL}(z)$ n'est une distribution de probabilité que dans le cas où les variables sont indépendantes. De plus, cette approximation ne conduit pas à faciliter le calcul des $P_G(z_i \mid y, \Psi)$.
Cette approximation est utilisée par de nombreux algorithmes d'estimation des paramètres d'un champ de Markov cachée, notamment l'algorithme ICM (Besag 1986) et l'algorithme EM Gibbsien (Chalmond 1989).

L'algorithme ICM (Iterated Conditional Mode)

On a

$$P_G(Z_i = \mathbf{z}_s | y, z_{S/\{i\},\Psi}) \propto \prod_{\ell=1}^{k} (f(y_i, \theta_\ell))^{z_{i\ell}} P_G(z_i | z_j, j \in N(i), \beta).$$

- Étape de classification $z^{(q)}$ obtenu par maximisation de proche en proche en z_i de $P_G(Z_i = \mathbf{z}_s | y, z_{S/\{i\}}, \Psi)$.
- Étape de maximisation Les θ_ℓ sont obtenus par maximisation de

$$\sum_{i/z_{i\ell}=1} z_{i\ell} \log f(y_i, \theta_\ell)$$

  et β est obtenu par maximisation de la pseudo-vraisemblance

$$\sum_{i \in S} \log P_G(z_i | z_{N(i)}, \beta).$$

Caractéristiques de ICM

- méthode de référence, simplicité, convergence rapide, bons résultats avec des voisinages du second ordre (8 voisins).
- forte dépendance des conditions initiales, sous-optimalité (biais pouvant être important).

L'algorithme EM Gibbsien

- Étape MCMC approximation de Monte-Carlo de $P_G(z|y,\Psi)$ après passages répétés d'un échantillonneur de Gibbs

$$P_G(z|y,\Psi) \approx \frac{1}{M}\sum_{m=1}^{M} P_G(z^m|y,\Psi).$$

- Étape de maximisation θ maximise $Q(\theta|\Psi^{(q)})$ et β maximise $Q(\beta|\Psi^{(q)})$, $P_G(z|\beta)$ étant remplacé par son approximation par la pseudo-vraisemblance.

Caractéristiques de EM Gibbsien

- Lenteur de chez Lenteur.
- Bonnes estimations attendues, mais aucun résultat d'optimalité.

L'approximation en champ moyen

Idée : se ramener à un système de variables indépendantes en négligeant les fluctuations des voisins d'un site $i \to$
Faire comme si ces voisins prenaient leur valeur moyenne.

$\Rightarrow$ Équation de cohérence (au point fixe) :
Les moyennes calculées par l'approximation doivent être égales aux moyennes utilisées pour la définir. Pour le modèle de Potts, cela donne

$$\forall i \in S, \quad \bar{z}_i = \frac{\sum_{z_i} z_i \exp\{\beta z_i^t \sum_{j \in N(i)} \bar{z}_j\}}{\sum_{z_i} \exp\{\beta z_i^t \sum_{j \in N(i)} \bar{z}_j\}}.$$

Algorithme général de type champ moyen

Itération q:

(1) Générer $\tilde{z}^{(q)} \leftarrow y$ et $\Psi^{(q-1)}$

$$P_G(z \mid \beta) \leftarrow P_{\tilde{z}^{(q)}}(z \mid \beta) \equiv \prod_{i \in S} P_G(z_i \mid \tilde{z}^{(q)}_{N(i)}, \beta)$$

(2) EM sur ce modèle :

$$P_G(y, z \mid \Psi) \leftarrow \prod_{i \in S} \{f_i(y_i \mid z_i, \theta)\ P_G(z_i \mid \tilde{z}^{(q)}_{N(i)}, \beta)\}$$

Remarques

- Algorithmes de type champ moyen: $\tilde{z}^{(q)}$ vu comme une configuration de voisins,
- choix de $\tilde{z}^{(q)}$ basé sur $P_G(z \mid y, \beta)$ et non sur $P_G(z|\beta)$ pour tenir compte des données y.
- mise à jour de $\tilde{z}^{(q)}$ séquentielle et non synchrone.
- un seul cycle (tous les pixels) : $\sim$ champ moyen de Zhang, ICM
- une seule itération de EM : $\sim$ Zhang (1992).

Choix des voisins

- algorithme champ moyen : estimateur champ moyen de $P_G(z \mid y, \Psi^{(q-1)})$ (équation de point fixe)
- algorithme champ modal : estimateur champ modal de $P_G(z \mid y, \Psi^{(q-1)})$ ($\sim$ ICM)
- algorithme champ simulé : réalisation de $P_G(z \mid y, \Psi^{(q-1)})$ (échantillonneur de Gibbs)

Les voisins sont fixés à des valeurs constantes $\tilde{z}$ et l'approximation est de la forme

$$P_G(z) \quad \simeq \quad P_{\tilde{z}}(z) = \prod_{i \in S} P_G(z_i \mid \tilde{z}_{N(i)})$$

Elle est différente de l'approximation par pseudo-vraisemblance $\prod_{i \in S} P_G(z_i \mid z_{N(i)})$

Algorithme NEM : Neighborhood EM

Point de vue de la classification pénalisée (Ambroise, Govaert 1996)

- Critère à optimiser sur $\mathbf{c}$ matrice de classification floue

$$W(\mathbf{c}, \theta) = FCL(\mathbf{c}, \theta) + \beta \sum_{i \sim i'} \mathbf{c}_i \cdot \mathbf{c}_{i'}^t$$

  Pénalisation du critère d'Hathaway

$$FCL(\mathbf{c}, \theta) = \sum_{i=1}^{n} \sum_{\ell=1}^{q} c_{i\ell} \log \pi_\ell p_\ell(\mathbf{x}_i; \mathbf{a}_\ell) - \sum_{i=1}^{n} \sum_{\ell=1}^{q} c_{i\ell} \log c_{i\ell}$$

  avec un terme de régularité spatiale.

- Optimisation alternée
  - Etape E : $\mathbf{c}^{(n+1)} = \arg\max_{\mathbf{c}} W(\mathbf{c}, \Phi^{(n)})$:
    méthode itérative de point fixe
  - Etape M : identique à l'étape M de EM

- Équivalent approximation en champ moyen

- Si $\mathbf{c}$ représente une partition des sites, la procédure d'optimisation alternée est l'algorithme ICM.

Applications

Modèle de Potts à k couleurs :

$z_i \in \{0,1\}^k$ avec $z_{i,\ell} = 1$ ssi $i \in$ class ℓ

$$P_G(z|\beta) = W(\beta)^{-1} \exp\,(\beta \sum_{i\sim j} z_i^t z_j)$$

$f(y_i|z_i,\theta) = \mathcal{N}(\ell, \sigma_\ell^2)$, si $i \in$ classe ℓ

Algorithmes utilisés :

- champs moyen, modal, simulé
- ICM
- EM pour les mélanges (segmentation aveugle)

Simulations de modèles de Potts, 2 couleurs, avec bruit gaussien

- images 150×150
- voisinage du 1er ordre (4)
- $\theta = \{(m_\ell, \sigma_\ell), \ell = 1, 2\}$ avec $m_\ell = \ell$ et $\sigma_\ell = 1$, et $\beta = 0.2, 0.6$

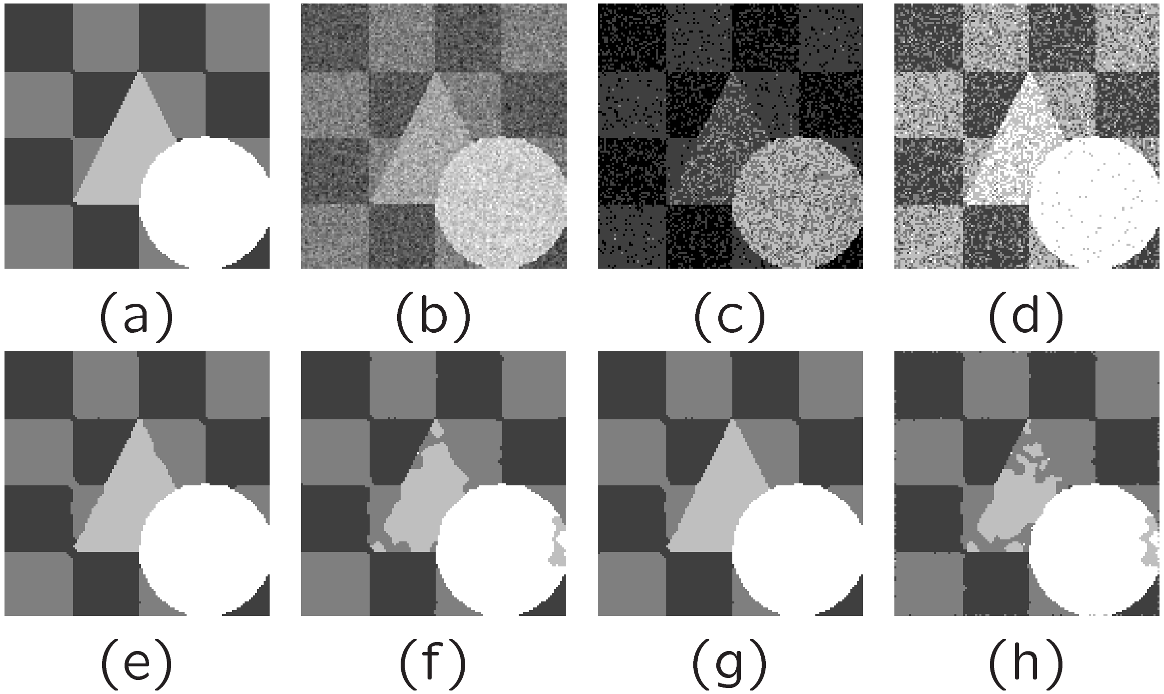
Estimations des paramètres du modèle de Potts

	β	μ_1	μ_2	σ
true values	0.2	1	2	1
EM	-	1.03	2.05	1.00
Mean Field	1.02	1.26	1.71	1.10
Mode Field	0.52	0.87	2.16	0.92
Simulated Field	0.14	0.99	2.02	1.00
ICM	0.10	0.63	2.42	0.67
true values	0.6	1	2	1
EM	-	0.98	1.97	1.00
Mean Field	1.40	1.25	1.77	1.09
Mode Field	0.85	0.90	2.08	0.95
Simulated Field	0.54	0.97	2.02	1.00
ICM	0.28	0.60	2.37	0.68

- Bon comportement du champ simulé.
- Le biais induit par ICM est ici important.

Image 4 couleurs degradée par un bruit gaussien

- images 128×128
- voisinage du 2ème ordre (8)
- $\theta = \{(m_\ell, \sigma_\ell), \ell = 1, \ldots, 4\}$ avec $m_\ell = \ell$ et $\sigma_\ell = 0.5$ pour $\ell = 1, \ldots, 4$.



(a) original image, (b) simulated image obtained by superposition of Gaussian noise with standard deviation $\sigma = 0.5$, (c) initial segmentation using simple thresholding, (d) EM segmentation, (e) mean field segmentation, (f) mode field segmentation, (g) simulated field segmentation, (h) ICM segmentation.

Interrogations sur le rôle de β

- Importance ?
- Lien entre β et k ?
- β paramètre ou température ?
- Modèle de Potts sans terme sur les singletons
  $\rightarrow$ Les couleurs sont considérées à parts égales...
  $\rightarrow$ Que se passe-t-il si on introduit effectivement des termes du premier ordre ?