UNIVERSITÉ JOSEPH FOURIER - GRENOBLE I
U.F.R. D'INFORMATIQUE
ET DE MATHÉMATIQUES APPLIQUÉES

THÈSE
pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Discipline : Mathématiques Appliquées

Présentée et soutenue publiquement

par

NATHALIE PEYRARD

Le 30 octobre 2001

APPROXIMATIONS DE TYPE CHAMP MOYEN
DES MODÈLES DE CHAMP DE MARKOV
POUR LA SEGMENTATION DE DONNÉES SPATIALES

Jury

Anatoli Iouditski	Président
Josiane Zerubia	Rapporteur
Bernard Chalmond	Rapporteur
Gilles Celeux	Directeur de thèse
Florence Forbes	Examinateur

UNIVERSITÉ JOSEPH FOURIER - GRENOBLE I
U.F.R. D'INFORMATIQUE
ET DE MATHÉMATIQUES APPLIQUÉES

THÈSE
pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Discipline : Mathématiques Appliquées

Présentée et soutenue publiquement

par

NATHALIE PEYRARD

Le 30 octobre 2001

APPROXIMATIONS DE TYPE CHAMP MOYEN
DES MODÈLES DE CHAMP DE MARKOV
POUR LA SEGMENTATION DE DONNÉES SPATIALES

Jury

Anatoli Iouditski	Président
Josiane Zerubia	Rapporteur
Bernard Chalmond	Rapporteur
Gilles Celeux	Directeur de thèse
Florence Forbes	Examinateur

Remerciements

A mon directeur de thèse, Gilles Celeux, j'adresse mes sincères remerciements pour avoir accepté de diriger cette thèse, pour l'excellence des conseils et des orientations qu'il m'a apportés tout au long de ces trois années.

Je remercie également ma co-directrice, Florence Forbes, pour son suivi attentif, ses indications rigoureuses et pour m'avoir toujours offert son temps sans compter.

A Adrian Raftery, j'exprime ma profonde reconnaissance pour m'avoir invitée au sein de son équipe à deux reprises durant ma thèse. J'ai beaucoup appris grâce à cette enrichissante collaboration.

Je tiens à remercier Josiane Zerubia et Bernard Chalmond pour l'intérêt qu'ils ont porté à ma thèse en acceptant d'en être les rapporteurs. Je remercie également Anatoli Iouditski qui m'a fait l'honneur de présider mon jury de thèse.

Enfin, je n'oublie pas les membres du projet IS2 et plus généralement mes collègues et amis de l'unité Rhône-Alpes de l'Inria : grâce à eux ma vie de thésarde grenobloise a été très agréable et joyeuse.

Enfin, mes pensées vont à mes parents qui m'ont stimulée et encouragée tout au long de mes années d'étude, ainsi quà Manu pour sa patience et son soutien pendant les derniers mois de la thèse.

Table des matières

Notations et conventions

Les variables aléatoires sont représentées par des lettres majuscules et les réalisations de ces variables aléatoires par des lettres minuscules.

S, ensemble de sites (par exemple pixels d'une image) ;

n, nombre d'éléments de S ;

$N(i)$, ensemble des indices des sites voisins du site i ;

N_v, nombre de voisins d'un pixel associé à un système de voisinage régulier ;

$i \sim j$, paire de sites voisins ;

$\mathbf{Z} = \{Z_1, \ldots, Z_n\}$, variables cachées ;

$\mathbf{Y} = \{Y_1, \ldots, Y_n\}$, variables observées ;

$\mathbf{z}$ (*resp.* $\mathbf{y}$) réalisation de la variable $\mathbf{Z}$ (*resp.* $\mathbf{Y}$) ;

$V = \{e_1, \ldots, e_K\}$, base canonique de $I\!R^K$;

$\mathcal{V} = \{\mathbf{u} \in [0,1]^K \ \sum_{k=1}^{K} u_k = 1\}$;

$P_G(\mathbf{z})$, $P_G(\mathbf{z} \mid \beta)$, distribution de Gibbs (ou markovienne) ;

$H(\mathbf{z} \mid \beta)$, énergie associée à la distribution de Gibbs $P_G(\mathbf{z} \mid \beta)$;

K, nombre de classes d'un modèle de champ de Markov discret ;

$\mathbf{p}_G = \{P_G(Z_i = e_k \mid \mathbf{y},\theta), i \in S, k \in \{1, \ldots, K\}\}$, vecteur des probabilités markoviennes locales conditionnelles dans le cas d'un champ de Markov discret ;

$g(\mathbf{y} \mid \mathbf{z},\theta)$, loi conditionnelle des données observées ;

$\Psi = (\theta,\beta)$, paramètre du modèle de champ de Markov caché ;

$P_G(\mathbf{z},\mathbf{y} \mid \Psi)$, distribution des données complètes dans le modèle de champ de Markov caché ;

$H(\mathbf{z} \mid \mathbf{y},\Psi)$, énergie associée au champ de Markov conditionnel ;

$L(\Psi,\mathbf{y})$ et $LL(\Psi,\mathbf{y})$, vraisemblance et log-vraisemblance des paramètres du modèle de champ de Markov caché ;

$L_c(\Psi,\mathbf{z},\mathbf{y})$, vraisemblance complète des paramètres du modèle de champ de Markov caché ;

$\tilde{\mathbf{z}}$, champ constant dans l'approximation de type champ moyen de $P_G(\mathbf{z} \mid \beta)$;

$P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta)$, approximation de type champ moyen de $P_G(\mathbf{z} \mid \beta)$;

$\mathrm{PV}(\mathbf{z} \mid \beta)$, pseudo-vraisemblance associée à $P_G(\mathbf{z} \mid \beta)$;

$g_G(\mathbf{y} \mid \Psi)$, loi des observations lorsque la loi des variables cachées est $P_G(\mathbf{z} \mid \beta)$;

$g_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi)$, loi des observations lorsque la loi des variables cachées est $P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta)$;

$L_{\tilde{\mathbf{z}}}(\Psi,\mathbf{y})$, vraisemblance des paramètres pour le modèle défini par $P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta)$ et $g(\mathbf{y} \mid \mathbf{z},\theta)$;

$\Psi^{\mathrm{MV}} = (\theta^{\mathrm{MV}},\beta^{\mathrm{MV}})$, estimateur du maximum de vraisemblance de Ψ ;

β^{PV}, estimateur du maximum de pseudo-vraisemblance de β ;

$I_A(x)$, fonction indicatrice de l'ensemble A, égale à 1 si $x \in A$ et nulle sinon ;

$\mathbb{E}[X]$, espérance d'une variable aléatoire X donnée ;

$\mathbb{E}_P[X]$, espérance d'une variable aléatoire X donnée selon la loi de probabilité P ;

$KL(P_1,P_2)$, divergence de Kullback-Leibler entre deux distributions P_1 et P_2 ;

$\mathbf{S}(P) = -\sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z})$, entropie de la distribution P ;

$d_{ik}(\mathbf{z}) = I_{e_k}(z_i)$, pour tout $\mathbf{z} \in V^n$, et pour tout $i \in S$, $k \in \{1,\ldots,K\}$;

$\mathbf{d} = \{d_{ik},\ i \in S,\ k \in \{1,\ldots,K\}\}$;

$\mathbf{f}^{(q)} = \{f_{ik}^{(q)},\ i \in S,\ k \in \{1,\ldots,K\}\}$, vecteur des fréquences markoviennes conditionnelles empiriques dans le cas d'un champ de Markov discret ;

$\mathbf{u}^t$, transposé du vecteur $\mathbf{u}$.

Présentation générale

Contexte

La *segmentation d'images* est le traitement qui consiste à regrouper les pixels d'une image en classes homogènes par rapport aux mesures observées en chacun de ces pixels. La mesure en un pixel, qui n'est pas forcément unidimensionnelle, permet de caractériser le pixel. L'objectif est de résumer l'information contenue dans l'ensemble des données observées, de sorte que le nombre de classes présentes dans l'image segmentée soit très inférieur au nombre des valeurs distinctes présentes dans les observations. L'image segmentée n'a généralement pas de sens physique mais elle permet une présentation compacte et plus facilement exploitable des données. Cette technique a des applications dans de nombreux domaines, comme par exemple la médecine ou l'imagerie satellitaire.
De nombreuses méthodes ont été développées pour la segmentation d'images. Deux familles importantes sont les méthodes par *recherche de frontières* et les méthodes par *recherche de régions* (*cf.* Cocquerez et Philipp 1995). Les méthodes par *recherche de frontières* sont des méthodes de détection de contours correspondant à une variation d'intensité ou à une discontinuité entre deux ensembles connexes de pixels. Ensuite, les contours obtenus doivent être fermés de façon à obtenir une partition de l'ensemble des pixels de l'image. Ainsi, la méthode du *gradient morphologique* (Serra 1982) consiste à repérer des sauts dans la norme du gradient par dilatation et érosion de l'image en niveaux de gris.
L'approche pour la segmentation par *recherche de régions* consiste à regrouper les pixels ayant des propriétés communes et à leur attribuer une même étiquette. Ce travail de thèse s'inscrit dans les développement actuels en recherche sur les méthodes de segmentation par *recherche de régions* et plus spécifiquement les méthodes s'appuyant sur une modélisation probabiliste. On considère alors que l'image observée et l'image segmentée sont des réalisations de variables aléatoires liées et le problème de segmentation est vu comme un problème de données manquantes, les données à restaurer étant les étiquettes assignées à chaque pixel dans l'image segmentée. La richesse des modèles probabilistes envisageables permet de saisir les caractéristiques globales et locales des images. Parmi ceux-ci, nous nous sommes intéressés aux modèles de champs de Markov cachés discrets qui permettent, grâce à leur structure, de prendre en compte les dépendances spatiales entre les différents pixels de l'image et d'incorporer une information a priori sur la segmentation cherchée. Les champs de Markov, ou de manière équivalente les champs de Gibbs, bien connus en physique statistique (Chandler

1987), ont trouvé un essor considérable lorsque leur utilité dans le domaine de l'image a été comprise (voir les ouvrages généraux Guyon (1993), Winkler (1995), Li (1995) et Chalmond (2000)).

Segmentation markovienne non supervisée

Les travaux réalisés pendant cette thèse concernent la *segmentation markovienne non supervisée*, c'est-à-dire dans un contexte où les paramètres du modèle de champ de Markov caché ne sont pas connus, et la sélection de modèle markovien. Nous avons classé les différentes techniques généralement utilisées pour traiter le premier problème en deux approches : l'approche *image* et l'approche *mélange*. Dans les deux cas, du fait de la complexité des modèles de champs de Markov, il est nécessaire en pratique de faire des approximations. Les méthodes de l'approche *image* sont des méthodes qui alternent segmentation et estimation afin de ce ramener au cas supervisé et avec données complètes. La segmentation est idéalement obtenue par recherche du mode conditionnel de la distribution du champ de Markov pour la valeur courante des paramètres du modèle. Les paramètres sont estimés par maximisation de la vraisemblance pour les données complètes (observations et segmentation) pour la segmentation courante. Une méthode de référence simple et rapide s'inscrivant dans cette approche est l'algorithme *Iterated Conditional Modes* (ICM), qui utilise l'approximation de la distribution d'un champ de Markov par la pseudo-vraisemblance (Besag 1986).
Dans l'approche *mélange*, le modèle de champ de Markov caché est interprété comme un modèle de mélange fini avec prise en compte d'une structure de dépendance spatiale allant au-delà de l'hypothèse naïve d'indépendance. Les techniques d'estimation sont alors les mêmes que dans le cas d'un modèle de mélange fini indépendant et visent à atteindre le maximum de vraisemblance des paramètres pour les données observées. Ces méthodes travaillent avec les probabilités markoviennes conditionnelles et non pas avec des restaurations. La mise en œuvre dans le cas dépendant n'est pas aussi simple et des approximations sont nécessaires. S'inscrivant dans l'approche *mélange*, l'algorithme EM (Expectation-Maximization) (Dempster, Laird, et Rubin 1977) est un algorithme itératif pour le calcul de l'estimateur du maximum de vraisemblance des paramètres du modèle dans le cas de données manquantes. Les techniques existantes pour traiter la complexité de l'algorithme EM dans le cas des modèles de champs de Markov cachés s'appuient sur des simulations de Monte-Carlo (*e.g.* échantillonneur de Gibbs, Geman et Geman 1984). C'est par exemple le cas de l'algorithme stochastique Monte-Carlo EM (MCEM, Wei et Tanner 1990) ou de l'algorithme EM gibbsien (Chalmond 1989), qui utilisent à chaque itération des réalisations du champ de Markov afin d'approcher l'espérance conditionnellement aux données observées de la distribution des données complètes. Les versions de EM basées sur des simulations de Monte-Carlo conduisent à de bons résultats, cependant elles sont très lourdes en temps de calcul.
D'autre part, en pratique le nombre de classes présentes dans l'image segmentée n'est pas connu. Plus généralement, il est nécessaire de sélectionner le modèle markovien (*e.g.* structure du voisinage). Il s'agit d'un problème ouvert puisque ici encore la complexité des modèles de

champs de Markov cachés ne permet pas l'application directe des méthodes classiquement utilisées.

Les résultats obtenus durant cette thèse concernent la segmentation markovienne non supervisée abordée du point de vue de l'approche *mélange*. Les travaux réalisés portent sur les trois points suivants, que nous détaillons ensuite :

- le développement de méthodes peu coûteuses pour la segmentation markovienne non supervisée, inspirées des méthodes de l'approche *mélange* et basées sur l'approximation en champ moyen d'un champ de Markov ;
- la démonstration de résultats théoriques sur la convergence de l'algorithme MCEM dans le cadre des modèles de champs de Markov cachés discrets ;
- le développement de nouveaux critères pour le choix du nombre de classes dans un modèle de champ de Markov caché discret.

Principe du champ moyen et EM

La première contribution de cette thèse est le développement de méthodes simples pour la mise en œuvre de l'algorithme EM pour champs de Markov cachés. Celles-ci offrent une solution moins coûteuse que l'utilisation de simulations de Monte-Carlo.
Pour cela, nous nous appuyons sur le principe du *champ moyen* issu de la mécanique statistique (Chandler 1987). L'approximation en champ moyen est une méthode simple pour approcher l'espérance d'un champ de Markov. L'idée consiste, lorsque l'on considère un pixel, à négliger les fluctuations des pixels dans son voisinage en les fixant à leur valeur moyenne. Cela conduit à un système de variables indépendantes plus facile à manipuler. Ce principe a été utilisé par Zhang (1992) pour rendre possible la mise en œuvre de l'algorithme EM pour le modèle de champ de Markov caché et donne des résultats satisfaisants. Nous nous sommes attachés dans un premier temps à étudier l'algorithme correspondant, que nous avons nommé *algorithme en champ moyen*, présenté jusqu'à présent d'un point de vue essentiellement heuristique par Zhang (1992). Cela afin de mieux comprendre la nature des approximations faites et la structure de la procédure. Cette étude a permis de mettre en évidence une justification probabiliste des approximations utilisées et nous a conduits à proposer une famille d'algorithmes de type EM pour l'estimation des paramètres d'un modèle de champ de Markov caché, que nous appelons algorithmes EM *de type champ moyen*. Plus précisemment, nous avons considéré les variations possibles autour du principe du champ moyen en proposant une famille générale de modèles *de type champ moyen* pour approximer un champ de Markov. L'idée conductrice est de garder les caractéristiques attrayantes de l'approximation en champ moyen en se ramenant à un système de variables indépendantes, obtenu en fixant le voisinage de chaque pixel à une constante arbitraire, et d'explorer d'autres choix pour cette constante que celui de la valeur moyenne. En particulier, nous avons étudié deux choix assez naturels, le mode et une simulation du champ de Markov, conduisant respectivement aux approximation *en champ modal* et *en champ simulé* d'une distribution de

Gibbs. La structure des algorithmes EM de type champ moyen s'interprète alors comme un algorithme EM pour modèle de mélange fini indépendant au cours duquel le modèle change de manière adaptative à chaque itération en fonction de l'approximation courante du voisinage. Ces algorithmes présentent l'avantage de prendre en compte l'information spatiale existant dans le modèle markovien initial, tout en gardant la simplicité du cas indépendant. Nous distinguons deux procédures parmi les algorithmes EM de type champ moyen : l'algorithme *en champ modal* et l'algorithme *en champ simulé*. Le premier, proche de l'algorithme ICM, est basé sur l'approximation en champ modal d'un champ de Markov, alors que le second est lié à l'approximation en champ simulé et utilise l'échantillonneur de Gibbs.
Afin de valider ces développements méthodologiques, nous avons évalué sur des images simulées et réelles, et pour un modèle de champ de Markov simple, le modèle de Potts, les performances des algorithmes en champ moyen, en champ modal et en champ simulé. Nous les avons comparés entre eux et aussi avec l'algorithme ICM – algorithme de référence mais connu pour induire un biais sur les estimations des paramètres – et avec l'algorithme EM pour un modèle de mélange fini indépendant. Cette étude, outre son intérêt pour la validation des nouveaux algorithmes proposés, vient compléter les rares expériences comparatives réalisées sur l'algorithme en champ moyen. La comparaison avec l'algorithme EM pour un modèle de mélange fini indépendant confirme l'importance de traiter l'information spatiale présente dans une image et montre que les approximations de type champ moyen réalisées préservent cette information. En particulier, les expériences réalisées mettent en avant les bonnes performances de l'algorithme en champ simulé, que ce soit en termes d'estimation dans le cas de données simulées ou en termes de segmentation. Notamment, on ne retrouve pas le problème de biais inhérent à ICM. D'autre part, nous avons constaté sur quelques exemples un effet de surlissage, via le paramètre spatial du modèle de Potts, et un problème de convergence lente pour l'algorithme en champ moyen, que l'on n'observe pas avec l'algorithme en champ simulé.

Convergence de l'algorithme MCEM

Nous pensons que les bonnes performances de l'algorithme en champ simulé sont à attribuer à l'étape de restauration stochastique du voisinage à chaque itération. Ceci laisse espérer de bonnes propriétés de convergence de l'algorithme. Aussi l'ergodicité de la chaîne des estimateurs est-elle établie facilement. Cependant, du fait des approximations réalisées, la nature de la loi stationnaire reste inconnue. L'algorithme en champ simulé, tout en étant moins coûteux que l'algorithme MCEM, reste néanmoins très proche de cette procédure. Le deuxième axe de recherche de ces travaux de thèse a donc consisté à étudier les propriétés de convergence de l'algorithme MCEM dans le cas des champs de Markov cachés, puis à voir si elles peuvent s'étendre à l'algorithme en champ simulé. Dans ce contexte peu de résultats théoriques ont été établis sur MCEM à cause de la complexité du modèle. Nous contribuons à la compréhension de cette procédure en montrant que, dans le cas où seuls les paramètres de la densité conditionnelle des observations sont inconnus, sous certaines conditions raisonnables de régularité, l'estimateur des paramètres obtenu par MCEM converge presque sûrement vers l'estimateur du maximum de vraisemblance. Ce nouveau résultat vient compléter les résultats disponibles jusqu'à présent sur MCEM pour champ de Markov caché, qui laissaient espérer au mieux la convergence en probabilité (Comer et Delp 2000). Nous montrons que ces résultats s'étendent à une version généralisée de l'algorithme en champ simulé pour laquelle, à chaque itération, l'approximation des probabilités markoviennes conditionnelles se base non plus sur une mais sur T simulations. L'algorithme en champ simulé généralisé a alors une structure très semblable à celle de l'algorithme MCEM pour champ de Markov caché et nous montrons que les deux procédures ont les mêmes propriétés théoriques de convergence.

Choix du nombre de classes

Enfin, la troisième contribution de cette thèse a consisté à étendre le principe du champ moyen au problème de la sélection d'un modèle de champ de Markov caché et plus particulièrement au cas du choix du nombre de classes dans le modèle. En pratique, le nombre de classes présentes dans la segmentation cherchée n'est pas forcément connu. La résolution de ce problème peut se faire à partir d'un avis d'expert, mais des méthodes automatiques sont souhaitables. L'approche bayésienne pour la résolution de ce problème consiste à rechercher le modèle de champ de Markov caché qui rend les données observées les plus probables. Cette approche fournit les critères BIC (Schwarz 1978) et ICL (Biernacki, Celeux, et Govaert 2000). Cependant, ici encore les difficultés inhérentes au modèle se retrouvent et il n'est pas possible de calculer ces critères directement. Nous avons étudié les possibilités de l'outil champ moyen pour les approcher. Cela nous a conduit à deux techniques d'approximation de ces critères : soit par approximation de toute la distribution markovienne, soit par approximation uniquement de la constante de normalisation. Dans les deux cas, les approximations des critères BIC et ICL obtenues sont simples à calculer et on constate en pratique l'importance de la prise en compte de l'information spatiale pour le choix du nombre de classes dans une image. Il apparaît que les critères approchés par la première méthode présentent les meilleures performances dans le cas d'images simulées, alors que dans le cas d'images réelles les critères obtenus avec la deuxième méthode d'approximation conduisent à des résultats plus pertinents.

Ces derniers travaux confirment ainsi l'intérêt de l'utilisation du principe du champ moyen pour un autre problème que celui de l'estimation des paramètres d'un modèle de champ de Markov caché. Grâce à cet outil, nous avons développé une méthodologie complète traitant les trois problèmes, par ailleurs souvent imbriqués, de la sélection, de l'estimation et de la segmentation.

Organisation du mémoire

Le premier chapitre de ce mémoire est consacré à la description des modèles de champ de Markov pour l'image. Les principales méthodes d'estimation des paramètres et de segmentation markovienne non supervisées de l'approche *image* et de l'approche *mélange* sont présentées dans le chapitre 2. Le principe du champ moyen est décrit dans le chapitre 3 et le chapitre 4 est dédié aux algorithmes de type champ moyen, avec une partie méthodologique et une partie expérimentale. Les résultats théoriques établis sur l'algorithme MCEM pour champs de Markov cachés sont présentés dans le chapitre 5. Le dernier chapitre traite de l'utilisation du principe du champ moyen pour la sélection du nombre de classes dans un modèle de champ de Markov caché. Enfin, ce mémoire s'achève par une conclusion sur les travaux présentés.

Chapitre 1

Modèles markoviens pour l'image

L'introduction de méthodes statistiques pour l'analyse d'images commence avec la segmentation d'images satellitaires, dont l'un des précurseurs est Switer (1969). Les modèles stochastiques tels que les champs aléatoires proposés par Abend, Harley, et Kanal (1965) existent depuis plus de trente ans et les champs de Markov, introduits par Lévy (1948), sont également décrits dans les travaux de Dobrushin (1968). Cependant, leur intérêt pour la modélisation d'images n'est apparu que plus tard, (*cf.* Jain 1981). L'utilisation de modèles stochastiques permet de tenir compte de contraintes dues à la nature de l'image et d'exprimer des connaissances a priori. Ils visent à saisir par quelques paramètres la nature intrinsèque de l'image. Les méthodes de segmentation ou de restauration basées sur des modèles stochastiques présentent l'avantage d'être automatiques et non subjectives, à l'opposé des méthodes basées sur l'avis d'experts (*e.g.* segmentation par l'œil humain).

Les modèles de champs de Markov introduisent la prise en compte de la notion de proximité spatiale dans l'image et permettent de pondérer la portée de l'influence de ces dépendances spatiales. Lorsque les observations étudiées sont organisées géographiquement, l'état d'un site est porteur d'information sur l'état de ses voisins (*e.g.* tissus cellulaires, relevés pluviométriques sur une région d'étude). De plus, les techniques utilisées pour relever les données (scanner, satellite) créent des superpositions entre les mesures faites sur des pixels voisins, et ajoute une autre source de dépendance spatiale qui rend d'autant plus attrayante et nécessaire l'utilisation de modèles markoviens. Ces modèles permettent de décrire une image par ses carastéristiques locales, plutôt que globales. Les travaux réalisés par Besag (1974) ou Besag (1986) sur l'estimation des paramètres par maximum de pseudo-vraisemblance ont contribué à l'essor des champs de Markov pour l'image. Largement reprise, cette technique d'estimation d'un champ de Markov a conduit à des algorithmes simples et viables d'estimation-restauration alternées. Dans la même période, Geman et Geman (1984), en exploitant l'équivalence entre champ de Markov et distribution de Gibbs, ont amélioré la compréhension de ces modèles. Ils ont développé des algorithmes efficaces de simulation (*échantillonneur de Gibbs*) et de restauration (*recuit simulé*) basés sur l'interprétation gibbsienne des champs de Markov. Pour le lecteur intéressé, des descriptions plus détaillées de l'origine et de l'évolution de l'utilisation des champs de Markov en analyse d'images peuvent être trouvées dans Jain (1981) et dans McLachlan (1992), chapitre 13.

Les enjeux principaux en analyse d'images basée sur un modèle stochastique, et plus encore dans l'analyse markovienne d'images, sont la sélection de modèle, l'estimation des paramètres, la segmentation, thèmes qui seront abordés dans les chapitres suivants. Ce chapitre a pour but de donner une présentation simple et générale des champs de Markov pour la modélisation d'images, ainsi que des méthodes de simulation de ces modèles couramment utilisées.

1.1 Champ de Markov et distribution de Gibbs

Considérons un ensemble S de sites spatialement organisés, indicés par $\{1,\ldots,n\}$ sur lequel un système de voisinage est défini. Puisque nous nous intéressons à l'image, les éléments de S seront interprétés comme les pixels d'une grille régulière en dimension 2. La relation de voisinage entre deux sites i et j est représentée par $i \sim j$, et on note $N(i)$ l'ensemble des sites voisins du site i. On appelle alors clique d'ordre r, tout sous-ensemble c de sites de S à r éléments tel que pour tout i et tout j dans c, alors i et j sont voisins, et on note $\mathcal{C}$ l'ensemble des cliques relatives à un système de voisinage. Deux systèmes de voisinage souvent utilisés lorsque S est une grille régulière sont illustrés sur la figure 1.1 : il s'agit des voisinages appelés voisinages d'ordre 1 et d'ordre 2, pour lesquels on considère respectivement les 4 et les 8 plus proches voisins d'un site. Les cliques associées au voisinage d'ordre 1 sont des cliques d'ordre 1 ou 2 (*cf.* figure 1.1 (c)). Dans le voisinage d'ordre 2 interviennent en plus des cliques d'ordre 3 (*cf.* figure 1.1 (d)). Considérons un système de variables aléatoires $\mathbf{Z} = \{Z_1,\ldots,Z_n\}$ définies sur S. De manière générale, les notations $\mathbf{Z}_A$ et $\mathbf{z}_A$ désignent respectivement l'ensemble des variables $\{Z_i, i \in A\}$ et un réalisation de $\mathbf{Z}_A$, pour un sous-ensemble A quelconque de S. L'ensemble $S\backslash\{i\}$ est l'ensemble des sites de S, privé du site i.

Définition 1.1 *Un champ aléatoire* $\mathbf{Z}$ *est un champ de Markov sur* S *si la distribution jointe* $P(\mathbf{z})$ *vérifie les deux propriétés suivantes :*

Propriété markovienne :

$$\forall\, \mathbf{z}, \quad P(z_i \mid \mathbf{Z}_{S\backslash\{i\}} = \mathbf{z}_{S\backslash\{i\}}) = P(z_i \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}). \tag{1.1}$$

Positivité :

$$\forall\, \mathbf{z}, \quad P(\mathbf{z}) > 0. \tag{1.2}$$

La propriété markovienne (1.1) traduit le fait que l'influence des autres sites de S sur un site i se réduit à l'influence du voisinage de i. Il s'agit d'une extension en dimension 2 de la notion de chaîne de Markov. Les caractéristiques locales définies par (1.1) sont en général faciles à calculer si le nombre de voisins d'un site reste petit. Se donner un ensemble de caractéristiques locales ne permet pas toujours de définir de manière cohérente la distribution jointe $P(\mathbf{z})$. Elles doivent pour cela vérifier certaines contraintes (voir Chalmond 2000). Dès lors que la propriété de positivité (1.2) est vérifiée, la distribution jointe $P(\mathbf{z})$ du champ de Markov est déterminée de manière unique à partir des caractétistiques locales. Cependant, la formulation (1.1) ne permet pas d'avoir une expression simple de cette distribution jointe. On utilise alors une autre caractérisation : le théorème d'Hammersley-Clifford (voir Besag 1974) montre que si $\mathbf{Z}$ est un champ de Markov, alors $P(\mathbf{z})$ est une distribution de Gibbs et réciproquement.

Définition 1.2 *Un champ aléatoire $\mathbf{Z}$ est régi par une distribution de Gibbs si la distribution de probabilité jointe est de la forme*

$$P_G(\mathbf{z}) \;=\; W^{-1}\exp(-H(\mathbf{z})), \tag{1.3}$$

où la fonction énergie H, définie à une constante près, se décompose en une somme de fonctions potentiels V_c associées aux cliques $c \in \mathcal{C}$,

$$H(\mathbf{z}) \;=\; \sum_{c\in\mathcal{C}} V_c(\mathbf{z}_c), \tag{1.4}$$

et $W \;=\; \sum_{\mathbf{z}} \exp(-H(\mathbf{z}))$ est la constante de normalisation, appelée aussi fonction de partition.

À l'origine, les distributions de Gibbs étaient utilisées en physique statistique pour modéliser les systèmes physiques à l'échelle moléculaire. Elles représentent les systèmes les plus désordonnés au sens où, parmi les distributions de probabilité P telles que l'espérance de l'énergie $\mathbb{E}_P[H(\mathbf{Z})]$ est fixée, la distribution (1.3) est celle qui maximise l'entropie

$$\mathbf{S}(P) \;=\; -\sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z}).$$

La notation P_G sera reprise par la suite pour désigner toute distribution de probabilité vérifiant la définition 1.2. L'équivalence champ de Markov/distribution de Gibbs permet d'avoir une formule explicite de la distribution jointe et donc une description globale du modèle. Grâce aux potentiels il est possible de spécifier les dépendances spatiales locales du champ de Markov. Selon les modèles, les potentiels sont plus ou moins complexes, ce qui influe sur la richesse des caractérisations locales possibles, mais aussi sur la facilité à manipuler le modèle. Les probabilités conditionnelles (1.1) se retrouvent directement à partir de (1.3) :

$$\forall\, \mathbf{z}, \quad P(z_i \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}) = \frac{\exp(-\sum_{c\ni i} V_c(\mathbf{z}_c))}{\sum_{z_i'} \exp(-\sum_{c\ni i} V_c(\mathbf{z}_c'))}\,, \tag{1.5}$$

avec $\mathbf{z}_c' = \{z_i'\} \cup \{z_j,\, j \in c,\, j \neq i\}$. Si l'on a dit que celles-ci sont facilement calculables de manière analytique, ce n'est pas le cas de la distribution de Gibbs (1.3), même si l'on dispose

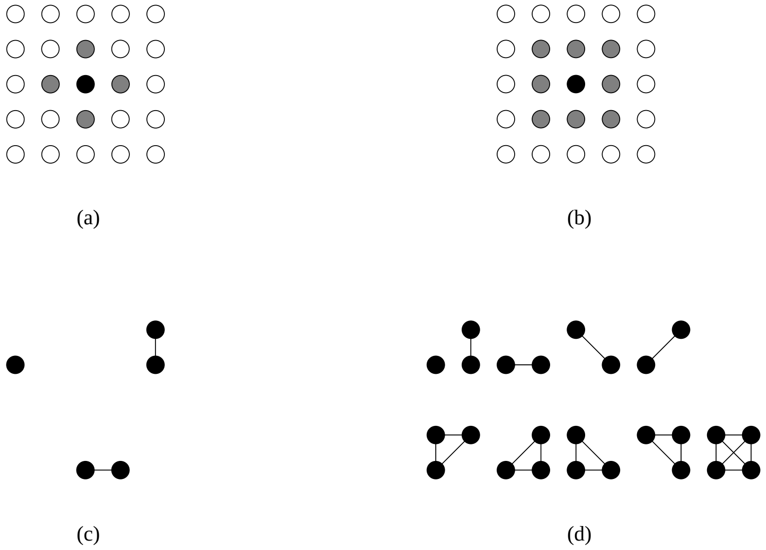


FIG. 1.1 – *représentation sur une grille régulière des systèmes de voisinage d'ordre 1 (à gauche) et d'ordre 2 (à droite). En haut, les pixels gris forment le voisinage du pixel noir. En bas, les cliques correspondantes sont représentées.*

d'une formule explicite. En effet, cela nécessite d'évaluer la fonction de partition W,

$$\begin{aligned} \text{cas continu :} \quad W &= \int_{\mathbf{z}} \exp(-H(\mathbf{z})), \\ \text{cas discret :} \quad W &= \sum_{\mathbf{z}} \exp(-H(\mathbf{z})). \end{aligned}$$

Lorsque la taille de S est trop importante, le nombre de configurations possibles pour $\mathbf{z}$ explose, ce qui rend le calcul de W impossible de manière directe. Dans le cas d'une image binaire de taille 256×256, W est une somme de 2^{256^2} termes !

1.2 Exemples

Nous décrivons dans ce paragraphe quelques exemples de structures possibles pour un champ de Markov. Une remarque souvent formulée à propos de ces modèles est qu'une réalisation typique (figure 1.2) n'a pas l'allure des images que l'on est amené à analyser en pratique. Une réponse possible (Besag, York, et Mollié 1991) est qu'à travers un modèle de champ de Markov on cherche à décrire les caractéristiques locales de l'image, plutôt que l'image globale. Il s'agit de traduire une connaissance ou des contraintes a priori. C'est la combinaison de cette information et de celle contenue dans les observations $\mathbf{y}$, que l'on retrouve dans la distribution conditionnelle du champ de Markov, qui conduit à une segmentation qui ressemble à une image "réelle". Les modèles présentés ci-dessous correspondent ainsi à différents choix d'appréhension des caractéristiques locales et de la complexité du modèle.

Champ de Markov auto-normal
Un modèle simple pour décrire un champ de markov $\mathbf{Z}$ où chaque variable Z_i est à valeurs continues est le *modèle auto-normal* de Besag (1974). Dans le cas d'une espérance nulle, la distribution conditionnelle de Z_i sachant les valeurs du champ sur le voisinage est donnée par

$$\forall\, \mathbf{z}, \quad P(z_i \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}) \;\; \propto \;\; \exp\{-\frac{1}{2\lambda_i}(z_i - \sum_{j \in N(i)} \beta_{ij} z_j)^2\}.$$

Les paramètres vérifient les propriétés suivantes: $\beta_{ij} = 0$ sauf si $i \sim j$, et $\beta_{ij}\lambda_j = \beta_{ji}\lambda_i$. On obtient pour cet exemple une expression simple de la distribution jointe,

$$P(\mathbf{z}) \;\; \propto \;\; \exp\{-\frac{1}{2}\mathbf{z}^t Q \mathbf{z}\},$$

où Q est une matrice $n \times n$ symétrique. Elle est de plus définie positive sous certaines conditions sur les paramètres λ_i et β_{ij} (voir Besag 1974).
Ce modèle fait partie de la famille des champs de Markov gaussiens, notamment utilisés pour la modélisation de textures (Chellapa, Chatterjee, et Bagdazian 1985).

Champs de Markov discrets
Dans d'autres applications, la variable Z_i représente une étiquette attribuée au site i, parmi K états possibles. On parle alors de champ de Markov discret. Les étiquettes peuvent être ordonnées (niveau de gris, couleurs) ou non (catégories, classes). Largement développés en analyse d'images, ces modèles sont en général définis à partir de leur fonction énergie H. Dans le cas de potentiels non nuls uniquement sur les cliques d'ordre inférieur à 3, celle-ci peut s'écrire

$$H(\mathbf{z} \mid \beta) \;\; = \;\; \sum_{i=1}^{n} \left[V_i(z_i) - \frac{\beta}{2} \sum_{j \in N(i)} V_{ij}(z_i,z_j) \right], \tag{1.6}$$

avec $V_{ij}(z_i,z_j) = V_{ji}(z_j,z_i)$ et $\beta \in I\!R$. Sous cette écriture, le potentiel associé à la clique $\{i,j\}$ est $\beta V_{ij}(z_i,z_j)$. Les potentiels sur les singletons pondèrent l'importance relative des différentes couleurs. Les potentiels sur les paires de sites voisins sont les termes de corrélation spatiale. L'influence spatiale est réglée par le paramètre β. Lorsque les différents états possibles représentent des classes, plutôt que de les coder de manière arbitraire par un numéro entre 1 et K, il est souvent plus pratique d'adopter une notation vectorielle. Une réalisation

de la variable Z_i est alors un vecteur à K composantes, toutes égales à zéro, sauf une, qui correspond à l'étiquette du site i et qui est égale à 1. On notera $e_1,e_2,\dots,e_K$ les K vecteurs canoniques correspondant aux K états possibles sous cette convention. C'est cette notation qui sera utilisée par la suite pour représenter un champ de Markov discret.
Un exemple de champ de Markov à K étiquettes est le *modèle de Potts sans champ externe* (que nous appellerons aussi *modèle de Potts simple*), pour lequel les potentiels sont définis ainsi,

$$\begin{array}{rcl} \forall i \in S,\, \forall z_i, \quad V_i(z_i) & = & 0, \\ \forall i,j \in S,\, \forall z_i,z_j, \quad V_{ij}(z_i,z_j) & = & V_{ji}(z_j,z_i) = I_{\{Z_i=Z_j\}}(z_i,z_j), \end{array}$$

où $I_A(x)$ est la fonction indicatrice de l'ensemble A, égale à 1 si $x \in A$ et nulle sinon. Lorsque β est positif, ce modèle favorise des configurations où deux pixels voisins sont dans le même état. En l'absence d'interactions spatiales (*i.e.* $\beta = 0$), supposer que les coefficients $V_i(z_i)$ sont tous nuls signifie que les états sont considérés équiprobables. Le modèle s'étend au cas de potentiels sur les singletons non nuls, et l'on parle alors de *modèle de Potts avec champ externe*. Si l'on considère la transformation ci-dessous,

$$Z_i = \begin{pmatrix} 1 \\ -1 \end{pmatrix} X_i + \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

un champ aléatoire $\mathbf{Z}$ dont la distribution est celle du modèle de Potts avec champ externe à deux états est équivalent au champ $\mathbf{X}$ dont les états sont codés sur $\{-1,1\}$ et dont les potentiels sont définis ainsi,

$$\begin{array}{rcl} \forall i \in S,\, \forall x_i, \quad V_i(x_i) & = & \alpha_i\, x_i, \\ \forall i,j \in S,\, \forall x_i,x_j, \quad V_{ij}(x_i,x_j) & = & V_{ji}(x_j,x_i) = I_{\{X_i=X_j\}}(x_i,x_j). \end{array} \tag{1.7}$$

Le champ markovien $\mathbf{X}$ est issu de la mécanique statistique et porte le nom de *modèle d'Ising*. Il était à l'origine utilisé pour décrire un système magnétique sur une grille régulière (voir par exemple Chandler 1987). Les deux états possibles de la variable Z_i correspondent aux deux orientations possibles du spin au site i.
Le modèle de Potts simple présente une rupture de symétrie des états. En effet, pour des valeurs du paramètre spatial inférieures à un certain seuil les réalisations obtenues sont fortement bruitées alors que pour des valeurs supérieures à ce seuil un des états possibles devient prépondérant. La figure 1.2 illustre ce phénomène avec des exemples de réalisation d'un modèle de Potts simple à 2 états, pour plusieurs valeurs du paramètre β et différents types de voisinage. Lorsque β augmente, on observe rapidement la disparition d'une couleur au profit de l'autre. Ce phénomène est d'autant plus rapide que le système de voisinage est étendu. Cette rupture de symétrie entre les différents états correspond à une transition de phase du modèle de Potts simple (voir par exemple Chandler 1987). Le phénomène disparaît en présence d'un champ externe non nul.

Geman et Geman (1984) ou Zhang (1992) utilisent un voisinage d'ordre 2 et le modèle plus complexe suivant pour représenter les étiquettes des pixels dans un problème de segmentation d'images

$$\begin{aligned} \forall i, \forall z_i, \quad V_i(z_i) &= \alpha_k, \text{ si le pixel } i \text{ appartient à la classe } k, \\ \forall i,j, \forall z_i,z_j, \quad V_{ij}(z_i,z_j) &= V_{ji}(z_j,z_i) = \gamma_m(2I_{\{Z_i \neq Z_j\}}(z_i,z_j) - 1). \end{aligned}$$

Il s'agit d'un modèle homogène anisotrope. Homogène car les potentiels ne dépendent que de la valeur du champ sur les pixels, mais pas de la position des pixels. Anisotrope car l'influence d'un voisin j d'un pixel i dépend des 4 orientations possibles d'une clique d'ordre 2 (Nord, Sud, Est, Ouest) à travers les paramètres γ_m, $m = 1,2,3$ ou 4.

Des modèles plus complexes encore ont été proposés pour décrire des interactions de plus longue portée. Avec plus de degrés de liberté, ils permettent de représenter des types d'images plus variés. En contrepartie, les analyses classiques deviennent plus difficiles à mettre en œuvre à cause de la structure du voisinage et du grand nombre de paramètres des modèles. En pratique les temps de calculs sont augmentés. Par exemple le *modèle chien* (Descombes, Morris, et Zerubia 1996) est un modèle qui utilise des cliques de 3×3 pixels, ce qui implique un voisinage $N(i)$ de 5×5 pixels. Les configurations possibles des cliques sont regroupées en classes en fonction de leur géométrie, et un potentiel est alloué à chaque classe. Ce modèle est plus souple que le modèle de Potts car il préserve les structures fines éventuellement présentes dans les données.

Processus de ligne

Geman et Geman (1984) proposent d'associer au système d'étiquettes $\mathbf{Z} = \{Z_i,\ i \in S\}$, un processus binaire de lignes $\mathbf{L}$. Ce processus se localise sur la grille $\bar{S}$ duale de S, dont les pixels se situent entre chaque paire horizontale ou verticale de pixels adjacents de S. Concrètement, en chaque pixel de $\bar{S}$ se trouve un marqueur binaire qui exprime la présence ou l'absence d'une ligne. L'ensemble $(\mathbf{Z},\mathbf{L})$ forme encore un champ de Markov, mais le voisinage est complexe : un pixel de S ou de $\bar{S}$ possède des pixels de S et de $\bar{S}$ dans son voisinage. L'avantage de ce modèle est d'apporter une information qui permet de préserver les contours dans l'image. Il est repris par Chalmond (1988) pour la restauration d'images à résonance magnétique nucléaire.

Champ de Markov causal

Les modèles décrits jusqu'à présent font partie de la famille des champs de Markov symétriques, ou encore non causaux. Pour ces modèles le système de voisinage induit une symétrie entre les sites, puisque si $j \in N(i)$, alors $i \in N(j)$. La difficulté des modèles non causaux réside dans l'évaluation de la probabilité $P_G(z_i,\mathbf{z}_{N(i)})$, même pour un système de voisinage simple. Les champs de Markov causaux, encore appelés *modèles mesh* (Abend, Harley, et Kanal 1965), sont construits de sorte à contourner cette difficulté. Dans ce modèle, la distribution du champ au pixel i ne dépend que de la valeur d'un petit nombre des prédécesseurs de i, les prédécesseurs étant définis suivant l'ordre lexicographique. Dans ce cas la relation de voisinage n'est plus symétrique. La probabilité $P_G(z_i,\mathbf{z}_{N(i)})$ se décompose en un produit

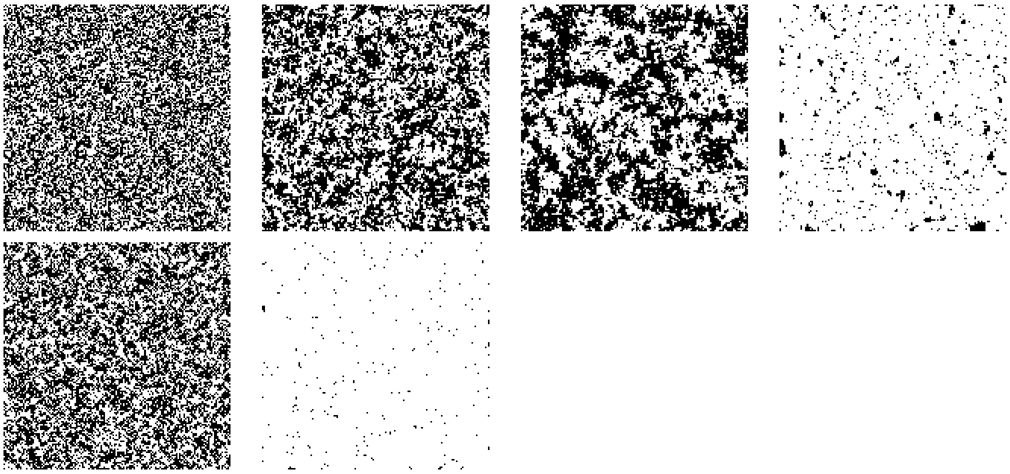


FIG. 1.2 – *simulation d'un modèle de Potts simple à 2 états. En haut et de gauche à droite, voisinage d'ordre 1 et* $\beta = 0.2,\ 0.6,\ 0.78,\ 0.98$. *En bas et de gauche à droite, voisinage d'ordre 2 et* $\beta = 0.2,\ 0.6$.

de probabilités conditionnelles facilement calculables, comme dans le cas d'une chaîne de Markov. Les modèles causaux, plus faciles à manipuler, ont peu de signification en analyse d'images, du fait de cette asymétrie non naturelle dans le contexte de données spatiales, et sont donc peu utilisés.

Dans ce mémoire, seuls les champs de Markov symétriques seront considérés par la suite. Notre étude se limite également au cas des champs de Markov discrets et non hiérarchiques, décrits dans le paragraphe suivant (voir Graffigne, Heitz, Perez, Preteux, Sigelle, et Zerubia (1995) pour une synthèse sur les modèles de champs de Markov hiérarchiques utilisés en analyse d'images).

1.3 Champ de Markov caché discret

Nous détaillons maintenant les champs de Markov cachés discrets, objets de ce travail de thèse. Les domaines d'utilisation de ces modèles sont en extension, car de plus en plus de secteurs font appel à des techniques de restauration, de segmentation, ou encore de reconnaissance de formes. Le contexte est celui des problèmes à données incomplètes : en chaque site i de l'ensemble S, deux données de types différents s'expriment, l'une observée, y_i, l'autre manquante (ou cachée), z_i. En segmentation d'images ou reconnaissance de formes, le champ $\mathbf{z} = \{z_i,\ i \in S\}$ représente une classification des pixels de S en fonction des observations $\mathbf{y} = \{y_i,\ i \in S\}$. En restauration d'images il représente l'image originale dont $\mathbf{y}$ est une observation dégradée. L'objectif global est de retrouver des informations sur $\mathbf{z}$ à partir des données observées $\mathbf{y}$.

Lorsque les composantes du champ $\mathbf{Z}$ sont à valeurs discrètes, le modèle de champ de Markov

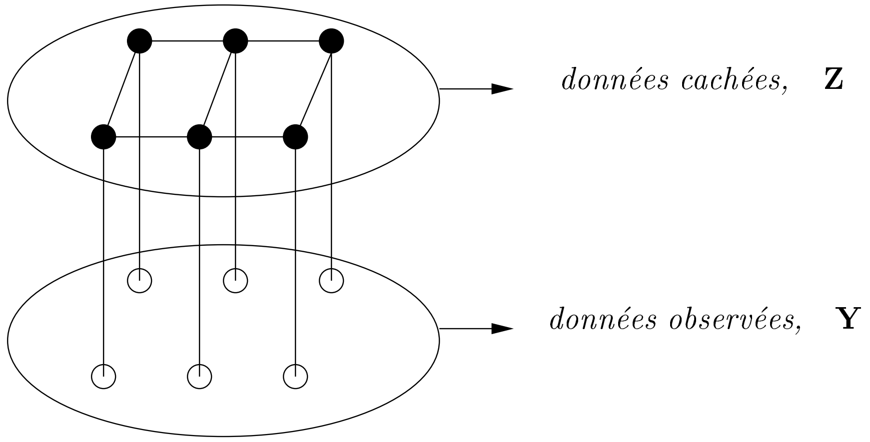


FIG. 1.3 – *représentation sur une grille régulière des dépendances entre l'ensemble des données observées* $\mathbf{Y}$ *et des données manquantes* $\mathbf{Z}$ *pour un modèle de champ de Markov caché et dans le cas d'un voisinage d'ordre 1.*

caché utilisé pour représenter les données complètes $(\mathbf{Z},\mathbf{Y})$ est le suivant : les données cachées $\mathbf{Z}$ sont décrites par un modèle de champ de Markov discret, de distribution de probabilité P_G définie par une énergie H comme dans (1.3) et dépendant d'un paramètre β. Chaque variable Z_i peut prendre un nombre fini d'états parmi K.
Les observations $\mathbf{Y} = \{Y_i,\, i \in S\}$ sont supposées indépendantes conditionnellement à $\mathbf{Z}$, de sorte que la densité conditionnelle $g(\mathbf{y} \mid \mathbf{z})$ se décompose ainsi,

$$g(\mathbf{y} \mid \mathbf{z}) \;=\; \prod_{i \in S} g(y_i \mid z_i,\, \theta), \tag{1.8}$$

où la forme des densités conditionnelles $g(y_i \mid Z_i = e_k,\, \theta)$ est connue, et où θ est un paramètre dans $\Theta \subset I\!R^d$. Le paramètre du modèle pour les variables complètes est noté $\Psi = (\theta,\beta)$. On note $P_G(\mathbf{z},\mathbf{y} \mid \Psi)$ la distribution des données complètes et $g_G(\mathbf{y} \mid \Psi)$ la densité des données observées lorsque la loi des données cachées est $P_G(\mathbf{z} \mid \beta)$. Les dépendances entre les données observées $\mathbf{Y}$ et les données manquantes $\mathbf{Z}$ pour le modèle de champ de Markov caché sont représentées sur la figure 1.3.

En pratique, ces densités sont souvent choisies parmi des densités gaussiennes (Besag, York, et Mollié 1991, Qian et Titterington 1991), de Laplace (Dang 1998) ou de Poisson (Besag, York, et Mollié 1991). En supposant que $g(y_i \mid z_i,\, \theta)$ est strictement positive pour tout z_i et tout y_i, on peut réécrire

$$g(\mathbf{y} \mid \mathbf{z}) \;=\; \exp\left(\sum_{i \in S} \log g(y_i \mid z_i,\, \theta)\right).$$

Des modèles plus complexes, prenant en compte un possible recouvrement dans les données observées ou intégrant des dépendances entre pixels dans la distribution conditionnelle

$g(\mathbf{y} \mid \mathbf{z},\theta)$, ont aussi été considérés (Besag 1986).
D'après (1.8), le champ conditionnel $\mathbf{Z}$ sachant $\mathbf{Y} = \mathbf{y}$ est également un champ de Markov, de fonction énergie associée,

$$H(\mathbf{z} \mid \mathbf{y},\Psi) \;=\; H(\mathbf{z} \mid \beta) - \sum_{i \in S} \log g(y_i \mid z_i,\, \theta). \tag{1.9}$$

Le terme $\sum_{i \in S} \log g(y_i \mid z_i,\, \theta)$, correspondant à la somme des potentiels sur les singletons, est souvent appelé *terme d'attache aux données*. Les distributions de probabilité des champs $\mathbf{Z}$ et $\mathbf{Z}$ sachant $\mathbf{Y} = \mathbf{y}$ seront appelées respectivement distribution marginale et conditionnelle dans ce mémoire de thèse. La difficulté inhérente à l'utilisation du modèle de champ de Markov caché réside dans le calcul de quantités impliquant ces deux distributions. Comme déjà souligné précédemment dans ce chapitre, les fonctions de partition marginale et conditionnelle ne sont pas accessibles sous une forme analytique simple, et des approximations sont nécessaires lorsqu'il s'agit d'estimer les paramètres ou de déterminer le mode de la distribution de Gibbs conditionnelle.

Dans le cas de la segmentation markovienne d'images, la variable Z_i représente une étiquette associée au site i. On parle parfois aussi de couleur ou de classe. Il s'agit d'un codage, et les valeurs possibles des étiquettes ne sont pas forcément ordonnées. L'image observée $\mathbf{y}$ constitue un ensemble de mesures relevées en chaque site. L'objectif est de trouver une classification des sites qui soit homogène du point de vue des observations $\mathbf{y}$, mais aussi du point de vue de la structure spatiale entre les sites. Si l'information spatiale n'est pas prise en compte, les classes obtenues sont morcelées. Il peut arriver qu'une classe de taille très petite soit enclavée dans une classe plus grosse. Ces situations sont peu représentatives de l'idée a priori d'une segmentation de données spatiales.
Notons que les observations ne sont pas forcement de la même nature que le champ caché. En particulier, les mesures réalisées en chaque pixel peuvent être multidimensionelles et à valeurs continues. Dans ce cas, la segmentation des données observées fournit une image unidimensionnelle plus facilement analysable. C'est par exemple le cas du traitement des données d'Imagerie à Résonance Magnétique (IRM). À partir de la mesure de l'intensité de la réaction au produit injecté en chaque pixel définissant la zone d'intérêt et pour différents temps, l'objectif pour le médecin est de repérer la présence d'une éventuelle zone à risque et de délimiter son contour. Les données observées se présentent sous la forme de huit images de la même zone correspondant à huit paramètres caractérisant l'évolution de la réaction. L'image segmentée fournit alors une information synthétisée, plus facilement interprétable que les données brutes (Forbes, Fraley, Georgian-Smith, Goldhaber, Peyrard, et Raftery 2001). La figure 1.4 montre un exemple de segmentation markovienne de données d'IRM pour la détection de tumeurs du sein.
Un autre contexte d'utilisation des modèles de champs de Markov cachés discrets est celui de la restauration d'images. La vraie image $\mathbf{z}$ n'est observée qu'à travers une version bruitée, $\mathbf{y}$. Dans le cas discret qui nous intéresse, l'image cachée prend un nombre fini de valeurs ordonnées (*e.g.* niveaux de gris). Il s'agit par exemple de retrouver la vraie valeur de chaque

pixel d'une zone du globe discrétisée sur une grille S, lorsque $\mathbf{y}$ représente l'image correspondante observée par satellite. On trouve d'autres applications, notamment dans l'industrie du bois pour des questions d'analyse de la texture et d'extraction de forme, ou pour la reconstruction de surface et la détermination de contours en vision par ordinateur.
Dans le cas des champs de Markov cachés discrets, les notions de segmentation et de restauration ne dépendent que de l'interprétation de la variable $\mathbf{Z}$ pour l'application considérée, mais les techniques d'analyse restent les mêmes. Par la suite, nous adopterons le terme unique de segmentation pour désigner les deux notions et les étiquettes z_i seront interprétées comme des classes.

1.4 Simuler un champ de Markov

Dans le cas de la segmentation d'images non supervisée (*i.e.* lorsque les paramètres du modèle de champ de Markov caché sont inconnus), certaines méthodes d'estimation des paramètres demandent de générer des réalisations du champ $\mathbf{Z}$ pour se ramener au contexte classique où toutes les données sont observées. La complexité des champs de Markov conduit également à utiliser des méthodes de type Monte-Carlo (Robert et Casella 1999) pour approximer des espérances par rapport à la distribution de Gibbs car le calcul n'est pas possible directement. Les techniques classiques de simulation (*e.g. méthode de rejet*) ne peuvent pas être envisagées pour ces modèles car elles nécessitent le calcul de la fonction de partition. Il existe des méthodes plus adaptées au modèle de champ de Markov qui permettent de simuler de manière efficace et simple une distribution de Gibbs. Nous en décrivons deux, qui sont celles qui ont été utilisées pour réaliser les simulations présentées dans cette thèse et qui comptent parmi les plus classiques.

Échantillonneur de Gibbs
Un exemple de méthode de type Chaîne de Markov de Monte-Carlo (MCMC) pour la simulation de la distribution de probabilité d'un champ de Markov est l'*échantillonneur de Gibbs*, proposé par Geman et Geman (1984). Il s'agit d'un cas particulier de l'algorithme de Metropolis-Hasting (Metropolis, Rosenbluth, Rosenbluth, Teller, et Teller 1953) à une seule composante. L'algorithme consiste à générer une chaîne de Markov $\{\mathbf{z}^{(q)}\}_{q \in I\!N}$ pour laquelle les probabilités de transition sont données par les caractéristiques locales de la distribution d'intérêt $P_G(\mathbf{z})$. À chaque pas, un seul site est visité, de telle sorte que $\mathbf{z}^{(q)}$ et $\mathbf{z}^{(q+1)}$ ne peuvent différer que par une composante. Si l'on note $\{i_1, i_2, \ldots\}$ la suite des indices des sites visités, la probabilité de transition de la chaîne à l'étape q est donnée par

$$P(\mathbf{Z}^{(q)} = \mathbf{z} \mid \mathbf{Z}^{(q-1)} = \mathbf{z}', \mathbf{Y} = \mathbf{y}) \\ = \begin{cases} 0 & \text{si } \exists\, i \neq i_q \,/\, z_i \neq z'_i, \\ P_G(Z_{i_q} = z_{i_q} \mid \mathbf{Z}_{N(i_q)} = \mathbf{z}'_{N(i_q)}) \text{ sinon.} \end{cases}$$

L'ordre de visite des sites le plus simple est l'ordre croissant des indices. Dans ce cas, la chaîne $\{\mathbf{z}^{(q)}\}_{q \in I\!N}$ est irréductible, apériodique et elle admet $P_G(\mathbf{z})$ comme unique mesure

FIG. 1.4 – *segmentation markovienne de données multidimensionnelles issues d'un examen IRM : en haut, exemples de 2 dimensions des données observées (parmi 8). En bas, segmentation avec un modèle de Potts simple et une loi conditionnelle gaussienne multidimentionnelle sur les observations.*

stationnaire, quelle que soit l'initialisation (voir Geman et Geman 1984, Théoreme A) Il suffit ainsi que chaque pixel soit mis à jour de manière séquentielle, peut importe l'ordre choisi pour cela. Cependant, l'algorithme peut mettre longtemps à converger, du fait qu'un seul pixel est modifié à chaque itération. La version synchrone (mise à jour en parallèle) de l'échantillonneur de Gibbs n'est pas valide. Il existe encore une distribution invariante, mais elle n'est pas identifiée.
Le calcul des probabilités de transition est simple car cela ne nécessite pas de recours à la fonction de partition, ce qui rend facile la mise en œuvre de l'algorithme.

Algorithme de Swendsen-Wang
Développé pour le cas du modèle de Potts simple, l'*algorithme de Swendsen-Wang* (Swendsen et Wang 1987) utilise des variables auxiliaires $\mathbf{U} = \{U_{ij}, i \sim j \in S\}$, appelées variables de lien, pour simuler suivant la distribution P_G. L'intérêt d'introduire ces variables est que les distributions conditionnelles $P(\mathbf{u} \mid \mathbf{z})$ et $P(\mathbf{z} \mid \mathbf{u})$ sont plus faciles à simuler que la distribution de Gibbs. Elles traduisent l'existence ou non d'un lien entre deux pixels voisins de même état. La distribution conditionnelle $P(\mathbf{u} \mid \mathbf{z})$ est définie de telle sorte que les variables U_{ij} soient indépendantes et uniformément distribuées,

$$P(u_{ij} \mid \mathbf{z}) \quad \propto \quad \exp(-\beta I_{\{Z_i=Z_j\}}(z_i,z_j)) \times I_{\{0 \leq U_{ij} \leq \exp(\beta I_{\{Z_i=Z_j\}}(z_i,z_j))\}}(u_{ij}). \qquad (1.10)$$

Si deux pixels voisins ne sont pas dans le même état, la variable lien U_{ij} est inférieure à 1. De manière intuitive, on dira dans ce cas qu'il n'y a pas de lien créé entre les pixels i et j. Un lien ne peut dont exister qu'entre deux pixels voisins d'état identique, et avec une probabilité $1 - \exp(-\beta)$. La distribution (1.10) et la distribution de Gibbs $P_G(\mathbf{z})$ définissent une distribution jointe $P(\mathbf{u},\mathbf{z})$. On en déduit

$$P(\mathbf{z} \mid \mathbf{u}) \quad \propto \quad \prod_{i \sim j} I_{\{0 \leq U_{ij} \leq \exp(\beta I_{\{Z_i=Z_j\}}(z_i,z_j))\}}(u_{ij}),$$

de sorte que si la variable de lien U_{ij} prend une valeur supérieure à 1, les pixels i et$j$ sont contraints à avoir la même étiquette.
Au cours de l'algorithme $\mathbf{u}$ et $\mathbf{z}$ sont mis à jour alternativement, suivant les distributions conditionnelles $P(\mathbf{u} \mid \mathbf{z})$ et $P(\mathbf{z} \mid \mathbf{u})$ respectivement. L'étape de génération de $\mathbf{z}$ revient donc à repérer les composantes connexes déterminées par la présence ou non de lien entre les pixels, puis à affecter les sites d'une composante à un même état, choisi uniformément parmi les K états possibles. La mise à jour de $\mathbf{u}$ a tendance à faire grossir les classes existantes. L'algorithme converge rapidement vers la distribution à l'équilibre P_G et ensuite explore relativement librement l'espace d'états de cette distribution. Si le paramètre spatial β est grand, l'échantillonneur de Gibbs est plus lent à se déplacer que l'algorithme de Swendsen-Wang. Cette dernière méthode, bien que plus coûteuse, est alors plus performante grâce à sa grande mobilité. À l'inverse, lorsque β est petit l'algorithme de Swendsen-Wang a tendance à fournir des réalisations formées d'un grand nombre de petites composantes, car la probabilité de créer un lien devient faible. Dans ce cas, la méthode de l'échantillonneur de Gibbs est

plus satisfaisante.
L'algorithme de Swendsen-Wang peut être facilement étendu à d'autres modèles de champs de Markov (*cf.* Higdon 1993, Morris 1999). Cependant il est moins performant pour des modèles avec interactions de longue portée et dès que la symétrie est perdue, comme c'est le cas par exemple dans un modèle de champ de Markov avec terme d'attache aux données.

Chapitre 2

Segmentation markovienne non supervisée

Nous présentons dans ce chapitre les méthodes classiques pour la segmentation markovienne d'images (voir les ouvrages généraux Cocquerez et Philipp (1995), Guyon (1993), Winkler (1995), Li (1995) et Chalmond (2000)). Nous nous plaçons dans un cadre non supervisé, c'est à dire que les paramètres du modèle de champ de Markov caché doivent être estimés. Nous avons regroupé les différentes méthodes dans deux familles : l'approche *image* et l'approche *mélange*. Les méthodes de segmentation non supervisée placées dans la première approche sont des adaptations des méthodes de segmentation supervisées pour l'image : elles alternent segmentation et estimation et à chaque étape, la segmentation courante est utilisée pour estimer les paramètres du modèle. Il s'agit de méthodes de référence en segmentation markovienne d'images. À ces méthodes nous avons opposé les méthodes de l'approche *mélange* qui traitent le problème d'estimation des paramètres d'un modèle de champ de Markov caché comme un problème de données manquantes. Les travaux sur la segmentation d'images présentés dans ce mémoire de thèse s'inscrivent dans cette seconde approche. Les techniques concernées ne sont pas spécifiques au contexte de l'image : ce sont des extensions au cas spatial de méthodes d'estimation et classification de mélanges. Ainsi, les algorithmes d'estimation de cette famille travaillent avec des probabilités au lieu d'une segmentation.
Les techniques de l'approche *image* sont décrites dans le sous-chapitre 2.1 et le sous-chapitre 2.2 est consacrée à l'approche *mélange*. Nous concluons le chapitre par une discussion sur les caractéristiques de ces deux familles.

2.1 Approche *image*

Les algorithmes de segmentation markovienne couramment utilisés requièrent idéalement que les paramètres du modèle soient connus. Cependant, il n'existe pas toujours de jeu de données test ou d'avis d'experts pour les estimer au préalable. Dans ce cas, les algorithmes mettent à jour la segmentation et les paramètres alternativement. La valeur courante de la segmentation est considérée comme la vraie scène dans l'étape d'estimation des paramètres, puis une nouvelle segmentation est calculée à partir des paramètres ainsi obtenus et des données observées. Nous plaçons ces méthodes non supervisées dans une approche *image* de la segmentation car elles sont des adaptations de méthodes supervisées classiques pour l'image. L'objectif principal est la segmentation, et l'estimation des paramètres peut être considérée comme une étape intermédiaire dans le processus d'estimation des classes. Nous présentons les principales méthodes de segmentation markovienne supervisée dans la partie 2.1.1. Les méthodes d'estimation des paramètres d'un modèle de champ de Markov caché et les algorithmes de segmentation-estimation alternées sont décrits dans la partie 2.1.2.

2.1.1 Segmentation supervisée

Lorsque les classes ou les couleurs des pixels d'une image ne sont pas directement observables, la définition de l'erreur commise en remplaçant la vraie image par une estimation n'est pas unique. La notion de “bonne segmentation” est en effet subjective car elle est liée aux critères donnés par l'utilisateur ou l'expert. La quantification de l'erreur et les caractéristiques de la segmentation estimée dépendent de la fonction de coût choisie. Dans cette partie deux méthodes de segmentation classiques sont décrites, correspondant à la minimisation du coût moyen sachant les observations, pour deux fonctions de coûts différentes : la première considère une erreur globale (paragraphe 2.1.1.1), la seconde une erreur locale (paragraphe 2.1.1.4). Ces méthodes sont les deux méthodes les plus largement utilisées pour la segmentation des champs de Markov cachés discrets. Leur mise en œuvre n'étant cependant pas possible de manière directe pour de tels modèles, nous présentons deux approximations particulièrement adaptées aux distributions de Gibbs, l'une de nature stochastique (paragraphe 2.1.1.2), l'autre déterministe (paragraphe 2.1.1.3).

2.1.1.1 Maximum a posteriori (MAP)

Soit $\mathbf{z}^*$ la vraie scène que l'on cherche à retrouver à partir des données et soit Ω l'ensemble des configurations possibles du champ aléatoire $\mathbf{Z}$.

Définition 2.1 *Une fonction de coût est une fonction $C : \Omega \times \Omega \to I\!\!R^+$ telle que*

$$\begin{aligned} \forall \mathbf{z}, \mathbf{z}^* \in \Omega, \quad C(\mathbf{z},\mathbf{z}^*) &\geq 0, \\ C(\mathbf{z},\mathbf{z}^*) = 0 \Leftrightarrow \mathbf{z} = \mathbf{z}^*. \end{aligned}$$

Le coût a posteriori est la fonction $\bar{C} : \Omega \to I\!\!R^+$ égale au coût moyen de l'estimation de $\mathbf{z}^*$ par la configuration $\mathbf{z}$ sachant les observations $\mathbf{y}$,

$$\bar{C}(\mathbf{z}) = I\!\!E[C(\mathbf{z},\mathbf{Z}) \mid \mathbf{y}].$$

Considérons la fonction de coût en $0-1$, qui quantifie de la même manière le coût de l'erreur lorsque $\mathbf{z}^*$ est remplacée par $\mathbf{z}$, quelle que soit $\mathbf{z} \neq \mathbf{z}^*$,

$$C(\mathbf{z},\mathbf{z}^*) = I_{\{\mathbf{Z} \neq \mathbf{Z}^*\}}(\mathbf{z},\mathbf{z}^*)\,. \tag{2.1}$$

La méthode de segmentation markovienne par maximum a posteriori (MAP) consiste à choisir la configuration qui minimise le coût a posteriori $\bar{C}$ correspondant à la fonction de coût (2.1). Dans ce cas, la fonction $\bar{C}$ est égale à,

$$\begin{aligned}\bar{C}(\mathbf{z}) &= \sum_{\mathbf{z}^*} C(\mathbf{z},\mathbf{z}^*) P_G(\mathbf{z}^* \mid \mathbf{y}) \\ &= 1 - P_G(\mathbf{z} \mid \mathbf{y}).\end{aligned}$$

L'estimateur de la vraie scène est donc la configuration qui a la plus grande probabilité conditionnellement aux données observées

$$\mathbf{z}^{\mathrm{MAP}} = \arg\max_{\mathbf{z}} P_G(\mathbf{z} \mid \mathbf{y}). \tag{2.2}$$

De manière équivalente il s'agit de la configuration qui minimise la fonction énergie $H(\mathbf{z} \mid \mathbf{y})$ de la distribution conditionnelle des données cachées. Par exemple, si l'on considère que la distribution marginale de $\mathbf{Z}$ est celle du modèle de Potts simple à K couleurs (voir partie 1.2), l'équation (2.2) devient

$$\mathbf{z}^{\mathrm{MAP}} = \arg\max_{\mathbf{z}} \beta \sum_{i\sim j} I_{\{Z_i=Z_j\}}(z_i,z_j) + \sum_i \log g(y_i \mid z_i,\theta). \tag{2.3}$$

Il est intéressant de noter comment l'estimateur évolue avec le paramètre spatial. Si $\beta = 0$, le modèle de Potts simple correspond à la distribution uniforme sur les K états possibles. Dans ce cas,

$$P_G(\mathbf{z} \mid \mathbf{y}) \propto P_G(\mathbf{y} \mid \mathbf{z}),$$

et $\mathbf{z}^{\mathrm{MAP}}$ est égal à l'argument du maximum de la distribution conditionnelle des données observées pour un modèle de mélange fini indépendant avec proportions égales,

$$\begin{aligned}&\forall\ i \in S,\ z_i^{\mathrm{MAP}} = e_l, \\ &\text{avec } l = \arg\max_{1\le k\le K} g(y_i \mid Z_i = e_k,\theta).\end{aligned} \tag{2.4}$$

Par contre, si le terme de dépendance spatiale devient prépondérant sur l'information fournie par les données, la segmentation par MAP tend à fournir une image à une seule couleur, valeur limite de $\mathbf{z}^{\mathrm{MAP}}$ lorsque β tend vers l'infini. À l'exception de ces deux cas extrêmes, le calcul de l'estimateur du MAP ne peut pas se faire de manière directe pour une distribution de Gibbs car cela implique le calcul de la fonction énergie de la distribution conditionnelle des données cachées pour toutes les configurations possibles. Nous décrivons dans les paragraphes suivants deux méthodes d'approximation possibles, le *recuit simulé* et l'algorithme *Iterated Conditional Modes* (ICM).

2.1.1.2 Algorithme de recuit simulé

L'algorithme de recuit simulé (Metropolis, Rosenbluth, Rosenbluth, Teller, et Teller 1953, Geman et Geman 1984) est une méthode de relaxation stochastique pour le calcul du ou de l'un des modes conditionnels d'une distribution de Gibbs. Considérons la distribution de probabilité suivante, dépendant d'un paramètre $\mathcal{T}$ interprétable comme une température dans un formalisme physique,

$$\begin{aligned} P_{\mathcal{T}}(\mathbf{z} \mid \mathbf{y}) &\propto (P_G(\mathbf{z} \mid \mathbf{y}))^{1/\mathcal{T}} \\ &\propto \exp\left(-\frac{H(\mathbf{z} \mid \mathbf{y})}{\mathcal{T}}\right). \end{aligned} \tag{2.5}$$

Lorsque $\mathcal{T} \to \infty$, la distribution $P_{\mathcal{T}}$ tend vers la distribution uniforme sur l'ensemble des configurations possibles du champ de Markov caché $\mathbf{Z}$. Lorsque $\mathcal{T} = 1$, on retrouve la distribution conditionnelle du champ caché $\mathbf{Z}$. D'autre part, la distribution $P_{\mathcal{T}}$ a la bonne propriété de se concentrer sur l'ensemble des minimiseurs de la fonction énergie $H(. \mid \mathbf{y})$ lorsque $\mathcal{T} \to 0$. L'algorithme de recuit simulé consiste à simuler une chaîne suivant la distribution (2.5), en faisant décroître la température au cours des itérations. Cette méthode s'inspire des techniques de recuit en physique pour isoler les états d'énergie minimale. Tant que la température reste élevée, l'algorithme autorise des passages dans des configurations qui font augmenter la fonction énergie H, ce qui permet d'éviter qu'il ne reste bloqué dans un minimum local. Puis lorsque la température diminue, la chaîne évolue progressivement vers le ou les minima globaux.
Cependant, comme il a déjà été remarqué, il n'est pas possible de simuler directement la distribution d'un champ de Markov. Geman et Geman (1984) proposent alors d'utiliser un échantillonneur de Gibbs (décrit dans la partie 1.4) pour simuler selon la loi $P_{\mathcal{T}}$. Soit $\mathcal{T}^{(q)}$ et i_q respectivement la valeur de la température et l'indice du pixel mis à jour à l'itération q de l'échantillonneur de Gibbs, le procédé dit de recuit simulé se déroule de la manière suivante :

(1) initialisation de la chaîne à une valeur $\mathbf{z}^{(0)}$;

(2) itération de l'étape suivante jusqu'à convergence ou test d'arrêt positif :
simuler $\mathbf{z}^{(q)}$ suivant la loi

$$P(\mathbf{Z}^{(q)} = \mathbf{z} \mid \mathbf{Z}^{(q-1)} = \mathbf{z}', \mathbf{Y} = \mathbf{y}) \quad = \quad \begin{cases} 0 & \text{si } \exists i \neq i_q \,/\, z_i \neq z'_i, \\ P_{\mathcal{T}^{(q)}}(Z_{i_q} = z_{i_q} \mid \mathbf{Z}_{N(i_q)} = \mathbf{z}'_{N(i_q)}) & \text{sinon.} \end{cases}$$

La convergence de l'algorithme dépend du schéma d'évolution de la température. Cela peut nécessiter une décroissance relativement lente. La température peut rester constante sur plusieurs itérations. Geman et Geman (1984) montrent que si les conditions suivantes sont vérifiées,

$$\begin{aligned} &\lim_{q \to \infty} \mathcal{T}^{(q)} = 0, \\ &\forall t \geq t_0 \geq 2, \quad \mathcal{T}^{(q)} \geq n \, \frac{H_{max} - H_{min}}{\log(t)}, \end{aligned} \tag{2.6}$$

où H_{max} et H_{min} sont respectivement le maximum et le minimum de la fonction énergie $H(. \mid \mathbf{y})$, alors quelle que soit l'initialisation $\mathbf{z}^{(0)}$, la distribution de $\mathbf{Z}^{(q)}$ converge vers la distribution uniforme sur l'ensemble des minimiseurs globaux de $H(. \mid \mathbf{y})$. Ce schéma n'est pas utilisé en pratique car l'évolution de l'algorithme correspondant est très lente. Des schémas d'évolution de la température déterminés expérimentalement comme efficaces et plus rapides sont en général utilisés (Geman et Geman 1984).

2.1.1.3 Algorithme ICM

L'algorithme *Iterated Conditional Modes* (ICM) a d'abord été proposé par Besag (1983) comme une méthode itérative pour approximer l'estimateur du MAP puis comme une amélioration de cette méthode (Besag 1986). Le principe consiste à mettre à jour les pixels de manière itérative et les uns après les autres à partir de l'information contenue dans les observations $\mathbf{y}$ et dans les valeurs courantes du champ $\mathbf{z}$ sur les autres pixels. Ainsi, si $\mathbf{z}^{(q-1)}$ représente la configuration courante, à l'itération q la classe d'un seul pixel, disons i_q, est mise à jour par

$$\mathbf{z}_{i_q}^{(q)} \quad = \arg\max_{z_i} P_G(z_i \mid \mathbf{Z}_{N(i_q)} = \mathbf{z}_{N(i_q)}^{(q-1)}, \mathbf{y}),$$

et $\mathbf{z}_j^{(q)} = \mathbf{z}_j^{(q-1)}$ pour tout $j \neq i_q$. Le calcul de ces probabilités locales est facile car il ne fait intervenir que les différents états possibles d'un pixel, au lieu de toutes les configurations de l'image comme dans le calcul de l'estimateur du MAP. La segmentation par l'algorithme ICM est donc plus rapide que la segmentation par MAP (avec éventuellement recuit simulé). Par contre, elle n'a plus le fondement mathématique de cette dernière.
Un cycle de l'algorithme est effectué lorsque tous les pixels ont été visités. Lorsque le parcours se fait de manière séquentielle, il est facile de voir que la probabilité de la segmentation courante ne décroît jamais,

$$\begin{aligned} P_G(\mathbf{z}^{(q)}) &= P_G(z_{i_q}^{(q)} \mid \mathbf{z}_{N(i_q)}^{(q)}) P_G(\mathbf{z}_{S\backslash\{i_q\}}^{(q)}) \\ &= P_G(z_{i_q}^{(q)} \mid \mathbf{z}_{N(i_q)}^{(q-1)}) P_G(\mathbf{z}_{S\backslash\{i_q\}}^{(q-1)}) \\ &\geq P_G(z_{i_q}^{(q-1)} \mid \mathbf{z}_{N(i_q)}^{(q-1)}) P_G(\mathbf{z}_{S\backslash\{i_q\}}^{(q-1)}) \\ &= P_G(\mathbf{z}^{(q-1)}). \end{aligned}$$

Cette propriété assure la convergence de l'algorithme vers un maximum local de la distribution de Gibbs conditionnelle $P_G(\mathbf{z} \mid \mathbf{y})$. En pratique, on observe qu'après quelques cycles, l'algorithme a convergé. La segmentation ainsi obtenue dépend fortement du choix de l'initialisation $\mathbf{z}^{(0)}$. En outre, le choix de l'ordre de visite des pixels peut aussi influer. Garder le même ordre à chaque itération peut induire des effets directionnels dans l'image segmentée (**?**). D'autre part la convergence n'est plus assurée si la visite des pixels se fait de manière synchrone.
Pour un modèle de Potts simple à K couleurs, une itération de ICM s'écrit

$$\mathbf{z}_i^{(q)} = e_l,$$

$$\text{avec } l = \arg\max_{1\leq k\leq K} \beta \sum_{j\in N(i)} I_{\{e_k\}}(z_j^{(q-1)}) + \log g(y_i \mid Z_i = e_k,\theta). \tag{2.7}$$

Comme pour l'estimateur du MAP, la segmentation fournie par ICM est influencée par la valeur du paramètre spatial β. Lorsque celui-ci est nul, on retrouve d'après la règle de mise à jour (2.7) un mode de la distribution des données complètes $(\mathbf{z},\mathbf{y})$, pour un modèle de mélange fini avec proportions égales, comme avec la méthode du MAP. Par contre, lorsque β tend vers l'infini, ICM opère à chaque itération un vote à la majorité, les observations $\mathbf{y}$ n'intervenant que pour départager dans le cas de plusieurs classes majoritaires dans le voisinage d'un pixel. Cette règle de mise à jour peut conduire à des *effets damier*: dans la segmentation apparaîssent des zones à l'aspect de damier à l'intérieur desquelles les classes des pixels oscillent d'une itération à l'autre.

2.1.1.4 Maximisation des probabilités marginales (MPM)

La méthode du MAP cherche à reconstruire la vraie scène de manière globale avec un coût qui porte sur l'image intégrale. Le coût en 0-1 est grossier, il attribue la même pénalité à une configuration qui diffère par un pixel de la vraie image $\mathbf{z}^*$ et à une configuration quelconque. Il est souvent plus judicieux de chercher à maximiser la proportion moyenne de pixels bien classés (Marroquin, Mitter, et Poggio 1987). Cela correspond au coût suivant, défini comme une somme de coûts locaux,

$$C(\mathbf{z},\mathbf{z}^*) \;=\; \sum_{i\in S} I_{\{Z_i\neq Z_i^*\}}(z_i,z_i^*).$$

La minimisation du coût a posteriori est équivalente à la maximisation pour tout i dans S des probabilités suivantes

$$P_G(z_i \mid \mathbf{y}) \;=\; \sum_{\mathbf{z}_{S\setminus\{i\}}} g(\mathbf{y} \mid \mathbf{z}) P_G(\mathbf{z}). \tag{2.8}$$

L'estimateur de $\mathbf{z}^*$ correspondant porte le nom d'*estimateur du maximum des probabilités marginales* (MPM, Besag 1986). Si l'on reprend l'exemple d'un modèle de Potts simple à K classes, on constate que l'estimateur du MPM a le même comportement que l'estimateur du MAP pour les deux valeurs extrêmes du paramètre spatial (0 et l'infini). Pour des valeurs intermédiaires, l'estimateur du MPM est plus robuste vis-à-vis du paramètre spatial que ne l'est l'estimateur du MAP, mais il l'est moins que ICM (Guyon 1993). Du point de vue de la mise en œuvre, les deux méthodes rencontrent les mêmes difficultés, puisque l'on voit avec l'équation (2.8) que la distribution marginale des données cachées intervient dans le calcul de l'estimateur du MPM. Une méthode de simulation de Monte-Carlo peut être utilisée pour approcher les probabilités markoviennes d'intérêt (Besag 1986, Marroquin, Mitter, et Poggio 1987) comme l'échantillonneur de Gibbs, mais la méthode de segmentation reste longue car il est nécessaire de générer une chaîne de Markov jusqu'à ce que la mesure stationnaire soit atteinte pour ensuite approcher les probabilités $P_G(z_i \mid \mathbf{y})$ par la fréquence empirique des occurences de z_i.

2.1.2 Segmentation non supervisée

Les algorithmes de segmentation d'images présentés dans la partie 2.1.1 demandent de connaître les paramètres du modèle. Ce n'est en général pas le cas. Pour obtenir de manière automatique une segmentation lorsque les paramètres sont inconnus (segmentation non supervisée), les techniques de segmentation supervisée sont adaptées et les paramètres et la segmentation sont mis à jour alternativement. Cela conduit à des algorithmes dits de segmentation-estimation alternées. Lorsque l'étape d'estimation se fait par maximum de vraisemblance, des difficultés surviennent pour la mise à jour du paramètre spatial et des approximations sont nécessaires. Une première solution, assez lourde, consiste à utiliser des simulations de Monte-Carlo (Younes 1988, Descombes, Morris, Zerubia, et Berthod 1999). Une autre solution largement utilisée pour pallier la complexité des modèles de champs de Markov est l'approximation par la pseudo-vraisemblance (Besag 1975). Les méthodes de segmentation non supervisée présentées dans cette partie sont des versions adaptées de la segmentation par MAP et de l'algorithme ICM et qui utilisent l'une ou l'autre de ces deux méthodes d'approximation lors de l'étape d'estimation. Les versions stochastiques, basées sur des simulations, permettent de retrouver asymptotiquement les estimateurs fournis par les méthodes exactes, mais c'est au prix de longs temps de calcul. Les méthodes de segmentation avec approximation, de nature déterministe, ont l'avantage d'être beaucoup plus rapides. Il n'y a plus alors de garantie sur la qualité de l'estimateur, mais en pratique les résultats fournis par ces dernières méthodes sont très satisfaisants en terme de segmentation.

2.1.2.1 Estimateur du maximum de vraisemblance

Définition 2.2 *Soit* $\mathbf{y}$ *une réalisation de la loi de densité* $g_G(\mathbf{y} \mid \Psi)$*, on appelle vraisemblance du paramètre* Ψ *la fonction* $L(\Psi,\mathbf{y})$ *définie par*

$$L(\Psi,\mathbf{y}) \quad = \quad g_G(\mathbf{y} \mid \Psi). \tag{2.9}$$

Soit $(\mathbf{z},\mathbf{y})$ *une réalisation de la distribution* $P_G(\mathbf{z},\mathbf{y} \mid \Psi)$*, la vraisemblance complète du paramètre* Ψ *est la fonction* $L_c(\Psi,\mathbf{z},\mathbf{y})$ *égale à*

$$L_c(\Psi,\mathbf{z},\mathbf{y}) \quad = \quad P_G(\mathbf{z},\mathbf{y} \mid \Psi). \tag{2.10}$$

Lorsque le champ caché $\mathbf{z}$ est disponible, un estimateur possible du paramètre Ψ est l'estimateur du maximum de la vraisemblance complète, paramètre pour lequel le couple $(\mathbf{z},\mathbf{y})$ est le plus probable sous le modèle $P_G(\mathbf{z},\mathbf{y} \mid \Psi)$,

$$\Psi^{\mathrm{MV}} \quad = \quad \arg\max_{\Psi} L_c(\Psi,\mathbf{z},\mathbf{y}). \tag{2.11}$$

De manière équivalente, il est plus simple de maximiser le logarithme de la vraisemblance complète, fonction convexe des paramètres dans le cas des champs de Markov cachés (Chalmond 2000). Plus précisement, l'estimation par maximum de vraisemblance complète du paramètre spatial et du paramètre de la distribution conditionnelle des données se fait de manière indépendante,

$$\theta^{\mathrm{MV}} = \arg\max_{\theta} \log g(\mathbf{y} \mid \mathbf{z},\theta), \tag{2.12}$$

$$\beta^{\mathrm{MV}} = \arg\max_{\beta} \log P_G(\mathbf{z} \mid \beta). \tag{2.13}$$

Si la maximisation décrite par (2.12) ne pose pas de problème, la résolution numérique de (2.13) n'est pas possible directement, sauf cas particuliers (*e.g.* modèles mesh). En effet, le gradient n'est pas calculable de manière analytique, car cela nécessite l'évaluation de la fonction de partition W de la distribution markovienne. Nous présentons dans les paragraphes suivants plusieurs solutions proposées à ce problème et les algorithmes de segmentation-estimation alternées correspondants.

2.1.2.2 Estimateur du maximum de pseudo-vraisemblance

Les difficultés liées au calcul de l'estimateur β^{MV} étant dues à la complexité du modèle, une solution consiste à approximer le champ de Markov par un système de variables conditionnellement indépendantes. C'est le principe de la méthode des ensembles de codage proposée par Besag (1974). Pour un système de voisinage donné, considérons une partition en R sous-ensembles $S_1,S_2,\ldots,S_R$ des pixels de l'ensemble S, de sorte que la propriété suivante de dépendance conditionnelle soit vérifiée pour tout r dans $\{1,\ldots,R\}$:

$$P(\mathbf{z}_{S_r} \mid \{z_j, j \in S \setminus S_r\}, \beta) = \prod_{i \in S_r} P(z_i \mid \{z_j, j \in S \setminus S_r\}, \beta). \tag{2.14}$$

Il suffit pour cela que deux sites d'un même sous-ensemble ne soient pas voisins. Dans le cas d'un voisinage d'ordre 1, on peut ainsi séparer les sites en deux sous-ensembles (*cf.* figure 2.1 (a)). Pour un voisinage d'ordre 2, quatre sous-ensembles sont nécessaires pour que la propriété (2.14) soit assurée (*cf.* figure 2.1 (b)). Plus le système de voisinage est complexe, plus la partition doit contenir de sous-ensembles.
Pour chacun des sous-ensembles S_r, on a d'après la propriété markovienne (1.1)

$$\forall\ 1 \le r \le R, \quad P_G(\mathbf{z}_{S_r} \mid z_j, j \in S \setminus S_r, \beta) = \prod_{i \in S_r} P_G(z_i \mid \mathbf{z}_{N(i)}, \beta). \tag{2.15}$$

Les probabilités conditionnelles intervenant dans le produit (2.15) sont facilement calculables et la maximisation de (2.15) fournit un estimateur $\hat{\beta}_r$ du paramètre β. Afin de prendre en compte toute l'information disponible et pas seulement celle contenue dans le sous-ensemble S_r, une solution consiste à considérer une combinaison des différents estimateurs obtenus pour chaque sous-ensemble de codage.
La pseudo-vraisemblance (Besag 1975) est une approximation de la distribution P_G, dérivée de la méthode de codage par extention du produit (2.15) à l'ensemble des pixels de S,

Définition 2.3 *On appelle pseudo-vraisemblance, et on note* PV*, le produit des probabilités locales suivant,*

$$\mathrm{PV}(\mathbf{z} \mid \beta) = \prod_{i \in S} P_G(z_i \mid \mathbf{z}_{N(i)}, \beta). \tag{2.16}$$

La pseudo-vraisemblance n'est plus une distribution de probabilité contrairement à la distribution (2.15), sauf dans le cas trivial d'indépendance des pixels. L'estimateur du maximum de pseudo-vraisemblance de β est

$$\beta^{\mathrm{PV}} \quad = \quad \arg\max_{\beta} \mathrm{PV}(\mathbf{z} \mid \beta). \tag{2.17}$$

La fonction de pseudo-vraisemblance, comme son logarithme, est encore concave. Par rapport à la vraisemblance complète L_c, le gradient est calculable, puisque la fonction de partition de la distribution de Gibbs n'intervient plus. Quelques itérations d'un algorithme de descente du gradient fournissent l'estimateur β^{PV}. Cette technique est moins coûteuse que la méthode originale de codage, puisqu'un seul estimateur est calculé au lieu de R. L'existence, l'unicité et la consistence de l'estimateur du maximum de pseudo-vraisemblance (lorsque la taille de l'image augmente) ont été démontrées par Geman et Graffigne (1987) pour une famille de champs de Markov à nombre d'états fini, incluant notamment le modèle de Potts.

Chalmond (1989) propose une variante astucieuse aux algorithmes de descente de gradient, basée sur une reparamétrisation de la pseudo-vraisemblance. Notons $\mathcal{N}$ l'ensemble de toutes les segmentations possibles du voisinage d'un pixel donné. Si on choisit un modèle homogène pour la probabilité marginale $P_G(\mathbf{z}|\,\beta)$, la probabilité conditionnelle $P_G(k|\,v)$ qu'un pixel soit affecté à la classe k sachant que son voisinage est de type $v \in \mathcal{N}$ ne dépend pas de la position de ce pixel. L'expression de la pseudo-vraisemblance peut être réécrite comme un produit de ces probabilités,

$$\mathrm{PV}(\mathbf{z} \mid \beta) = \prod_{(k,v)} P_G(k \mid v)^{n_{k,v}(\mathbf{z})}, \tag{2.18}$$

où $n_{k,v}(\mathbf{z})$ représente le nombre de pixels dans la classe k dont le voisinage est de type v dans l'image segmentée $\mathbf{z}$. La maximisation de la pseudo-vraisemblance sous la contrainte que pour tout $v \in \mathcal{N}$, $\sum_{k=1}^{K} P_G(k \mid v) = 1$ conduit à l'estimateur suivant

$$\forall 1 \leq k \leq K, \forall v \in \mathcal{N}, \quad \hat{P}(k \mid v) = \frac{n_{k,v}(\mathbf{z})}{\sum_{l=1}^{K} n_{l,v}(\mathbf{z})}. \tag{2.19}$$

Le rapport de deux probabilités conditionnelles est proportionnel à β après transformation logarithmique :

$$\log\left(\frac{P_G(k \mid v)}{P_G(l \mid v)}\right) \propto \beta.$$

Les différentes constantes de proportionnalité étant connues, on obtient ainsi un système surdéterminé et la méthode des moindres carrés fournit un estimateur de β. Cette procédure échoue cependant si une des quantités $n_{k,v}(\mathbf{z})$ est nulle pour l'image $\mathbf{z}$ considérée.

Qian et Titterington (1992) et Rydén et Titterington (1998) généralisent la pseudo-vraisemblance avec la notion de pseudo-vraisemblance par blocs. Au lieu de regarder la

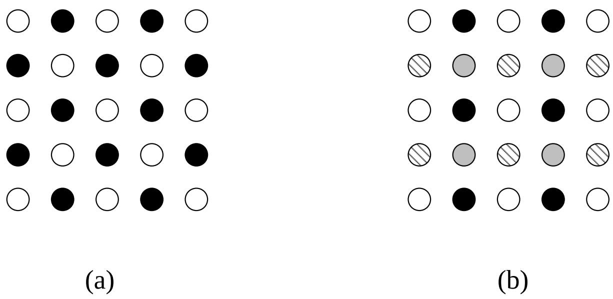


FIG. 2.1 – *méthode de codage: partition des pixels en sous-ensembles conditionnellement indépendants (a) pour un voisinage d'ordre 1, (b) pour un voisinage d'ordre 2.*

partition la plus fine de S, où chaque sous-ensemble est composé d'un élement, ils considèrent des partitions plus grossières, $B = \{B_1,B_2,\ldots,B_R\}$, par exemples des blocs carrés de 4 pixels et définissent

$$\mathrm{PV}_B(\mathbf{z}) \quad = \prod_{1 \le r \le R} P_G(\mathbf{z}_{B_r} \mid \mathbf{z}_{N(B_r)}),$$

où $N(B_r)$ désigne l'ensemble des pixels qui appartiennent aux voisinages des pixels de B_r, à l'exception des pixels de B_r eux-même. Ici, la partition n'est pas supposée vérifier la propriété (2.14), contrairement au cas de la méthode de codage. Rydén et Titterington (1998) utilisent l'approximation de la pseudo-vraisemblance par blocs pour l'inférence bayésienne sur les paramètres du modèle et observent sur leurs expériences que cela améliore significativement les résultats par rapport à l'utilisation de la pseudo-vraisemblance simple. Cependant, la taille des blocs doit rester raisonnable pour que les calculs mis en jeu ne deviennent pas trop lourds.

2.1.2.3 Algorithmes de segmentation-estimation de type MAP

Lakshmanan et Derin (1989) ont proposé une adaptation de la méthode de segmentation par recuit simulé dans le cas non supervisé. De manière récursive, l'algorithme de segmentation est interrompu et les paramètres sont réestimés d'après la règle (2.11) à partir de la segmentation courante. L'algorithme correspondant se nomme *Adaptative Segmentation Algorithm* (ASA). Lakshmanan et Derin (1989) ont démontré que, pour un schéma d'évolution de la température du recuit simulé vérifiant (2.6), la suite $\{(\mathbf{z}^{(q)},\Psi^{(q)})\}_{q\in\mathbb{N}}$ générée par ASA converge en loi vers un couple $(\mathbf{z}^*,\Psi^*)$ satisfaisant

$$\begin{aligned}\mathbf{z}^* &= \arg\max_{\mathbf{z}} P_G(\mathbf{z},\mathbf{y} \mid \Psi^*),\\ \Psi^* &= \arg\max_{\Psi} P_G(\mathbf{z}^*,\mathbf{y} \mid \Psi).\end{aligned}$$

En pratique, comme nous l'avons vu précédemment, l'estimateur (2.13) du paramètre spatial β ne peut pas être calculé directement. Lakshmanan et Derin (1989) utilisent l'approximation par la pseudo-vraisemblance pour mettre à jour β au cours de l'étape d'estimation de

l'algorithme.
Une approche similaire pour l'adaptation de l'estimation du MAP par recuit simulé utilise une technique de gradient stochastique au lieu de la pseudo-vraisemblance pour la mise à jour du paramètre spatial (Younes 1988). Ce principe utilise le fait que le gradient de $\log P_G(\mathbf{z} \mid \beta)$, incalculable directement, s'écrit comme l'espérance d'une fonction de $\mathbf{Z}$ suivant P_G. A chaque pas de la descente de gradient, cette espérance est remplacée par la valeur de la fonction pour une réalisation du champ marginal. Nous décrivons ci-dessous la méthode dans le cas d'un modèle de Potts simple. On remarque que l'énergie $H(\mathbf{z} \mid \beta)$ pour ce modèle est égale à $\beta n_{ch}(\mathbf{z})$ où $n_{ch}(\mathbf{z})$ est le nombre de cliques homogènes d'ordre 2 dans la réalisation $\mathbf{z}$.
On a alors l'égalité suivante :

$$\frac{\partial \log(P_G(\mathbf{z} \mid \beta))}{\partial \beta} = -n_{ch}(\mathbf{z}) - \mathbb{E}_{P_G}[n_{ch}(\mathbf{Z}) \mid \beta]. \tag{2.20}$$

À l'itération q de la descente de gradient, l'espérance $\mathbb{E}_{P_G}[n_{ch}(\mathbf{Z})| \beta^{(q)}]$ est approximée par $n_{ch}(\mathbf{z}^{(q)})$, $\mathbf{z}^{(q)}$ étant simulé suivant $P_G(\mathbf{z}| \beta^{(q)})$.
Ainsi, une itération de la méthode est donnée par :

$$\beta^{(q+1)} = \beta^{(q)} + \frac{\tau}{q+1}(n_{ch}(\mathbf{z}) - n_{ch}(\mathbf{z}^{(q)})),$$

où $\tau/(q+1)$ est le pas d'apprentissage.
La convergence de cet algorithme a été démontrée et l'estimateur ainsi obtenu est un maximum local de la vraisemblance, dès que le paramètre τ est assez petit. Cependant, la méthode est assez longue car à chaque itération de l'exploration du gradient, il est nécessaire d'utiliser un algorithme de simulation (*e.g.* échantillonneur de Gibbs). Younes (1989) adapte également cette technique de gradient stochastique pour l'estimation d'un champ de Markov caché. Dans ce cas, une simulation supplémentaire est nécessaire à chaque itération.

2.1.2.4 Algorithmes de segmentation-estimation de type ICM

Dans le cas où les paramètres ne peuvent pas être estimés à partir d'un jeu de données test, Besag (1986) propose de les estimer au cours de l'algorithme ICM. L'étape de segmentation de l'algorithme ICM non supervisé consiste à réaliser un cycle de la procédure ICM décrite dans le paragraphe 2.1.1.3 pour la valeur courante des paramètres (*i.e.* un seul passage en chaque site pour la mise à jour des classes). Au cours de l'étape d'estimation, le paramètre θ est réestimé par maximum de la log-vraisemblance conditionnelle d'après la règle (2.12) et le paramètre β par maximum de pseudo-vraisemblance d'après la règle (2.17), pour la valeur courante de la segmentation. Peu de résultats théoriques sont disponibles pour cette adaptation de ICM. En pratique, on observe que la convergence est atteinte en quelques itérations (une dizaine), comme pour la version supervisée. Dans la discussion de Besag, York, et Mollié (1991), Besag fait la remarque suivante : si β^* est la vraie valeur du paramètre spatial, si $\mathbf{z}_{\mathrm{ICM}}$ est la restauration du champ caché par ICM pour $\beta = \beta^*$, l'estimateur de β obtenu par maximisation de la pseudo-vraisemblance à partir de $\mathbf{z}_{\mathrm{ICM}}$ tend à être plus grand que l'estimateur que l'on obtiendrait à partir d'une réalisation de la distribution conditionnelle $P_G(\mathbf{z} \mid \mathbf{y},\beta^*)$. Cette remarque est valable également lorsque β est estimé à partir d'une segmentation obtenue par MAP. D'autre part, la méthode d'estimation du paramètre spatial n'est pas complètement satisfaisante car elle ne prend pas directement en compte les données observées. Cette information est utilisée seulement à travers $\mathbf{z}_{\mathrm{ICM}}$.

Dans le cas du modèle de Potts simple, Besag (1986) propose également que le paramètre β soit augmenté graduellement au cours de l'algorithme au lieu d'être estimé à chaque itération. Dans ce cas, β joue un rôle similaire à celui de l'inverse de la température dans l'algorithme de recuit simulé. En commençant par des petites valeurs du paramètre, on évite de se trouver trop tôt dans des modèles avec une forte dépendance spatiale, ce qui aurait pour conséquence de restaurer des images monochromes. Sur les expériences réalisées on observe que cette solution conduit à des segmentations satisfaisantes, avec de faibles taux de pixels mal classés, comparables à ceux obtenus par la version de ICM avec estimation par maximum de pseudo-vraisemblance. Notons enfin que Besag (1986) utilise un schéma de croissance du paramètre spatial jusqu'à $\beta = 1.5$. Lorsque, au cours d'un algorithme de segmentation, ce paramètre est fixé et non pas estimé, une valeur est généralement choisie entre 0.1 et 2.5 suivant la nature de l'image traitée et la structure de voisinage associée au modèle (*e.g.* Cocquerez et Philipp (1995), Besag (1986), Ambroise et Govaert (1998) ou encore Comer et Delp (2000)).

2.2 Approche *mélange*

Ce sous-chapitre est consacré aux méthodes de segmentation d'images par champ de Markov caché discret qui abordent le problème de la reconstruction des étiquettes des pixels comme un problème de données manquantes. Plus particulièrement, le modèle peut être interprété comme un modèle de mélange fini particulier, pour lequel la caractérisation des probabilités d'appartenance aux classes induit des dépendances complexes entre les étiquettes. Les techniques d'estimation et de classification pour un modèle de mélange, pour lequel il y a indépendance des étiquettes, peuvent alors être reprises pour la segmentation non supervisée de données spatiales. Nous avons qualifié cette approche d'approche *mélange* pour la segmentation markovienne non supervisée. En particulier, l'algorithme EM (Estimation-Maximisation) (Dempster, Laird, et Rubin 1977) est un outil puissant pour le calcul de l'estimateur du maximum de vraisemblance des paramètres en présence de données manquantes et a largement prouvé son efficacité pour l'estimation des paramètres d'un modèle de mélange fini indépendant. Il ne s'agit pas, comme dans le sous-chapitre précédent, d'une méthode spécifiquement destinée à l'image, mais d'une méthode générale dont le champ d'application est très vaste (données censurées, structures à variables latentes, réseaux neuronaux, etc.). Cette approche se distingue également de l'approche *image* par son objectif premier. L'algorithme EM est un algorithme d'estimation des paramètres, et non un algorithme de segmentation. Une segmentation des observations peut ensuite être obtenue par une méthode de segmentation supervisée, à partir des estimateurs fournis par EM. Cet algorithme maximise la vraisemblance $L(\Psi,\mathbf{y})$ et il ne requiert pas de connaître la classe des pixels, alors que les méthodes conduisant aux algorithmes de segmentation-estimation de l'approche *image* maximisent la vraisemblance complète $L_c(\Psi,\mathbf{z},\mathbf{y})$. Dans le contexte de l'analyse markovienne d'images, les difficultés habituelles propres au modèle se retrouvent pour l'algorithme EM. De nombreuses solutions ont été proposées pour rendre l'algorithme réalisable. Elles font appel aux principes d'approximation adaptés aux champs de Markov, comme la pseudo-vraisemblance ou les méthodes de Monte-Carlo.

2.2.1 Algorithme EM

L'algorithme EM (Dempster, Laird, et Rubin 1977) est une procédure largement utilisée pour calculer l'estimateur du maximum de vraisemblance pour des modèles à données incomplètes. Nous décrivons dans cette partie le principe de l'algorithme EM dans un contexte général. Par souci de simplicité, les variables correspondant aux données observées et aux données manquantes sont toujours notées $\mathbf{Y}$ et $\mathbf{Z}$ respectivement. Le vecteur $\mathbf{X}$ désigne les données complètes $(\mathbf{Z},\mathbf{Y})$. On suppose que le modèle probabiliste décrivant les variables dépend d'un paramètre Ψ et on rappelle que $L(\Psi,\mathbf{y})$ est la vraisemblance du paramètre, et $L_c(\Psi,\mathbf{x})$ la vraisemblance complète du paramètre. Puisque rien n'est précisé sur la nature des variables (discrètes ou continues), la notation P désigne indifféremment une densité ou une probabilité. Un estimateur possible pour le paramètre Ψ est l'estimateur du maximum de vraisemblance (voir la partie 2.1.2.1) défini par

$$\begin{aligned}\Psi^{\mathrm{MV}} &= \arg\max_{\Psi} \log P(\mathbf{y} \mid \Psi) \\ &= \arg\max_{\Psi} \log L(\Psi,\mathbf{y}). \quad (P_1)\end{aligned}$$

Dans beaucoup de cas, la solution à ce problème n'est pas directement accessible. On s'intéresse alors au problème suivant,

$$\begin{aligned}\Psi^{\mathrm{MVC}} &= \arg\max_{\Psi} \log P(\mathbf{x} \mid \Psi) \\ &= \arg\max_{\Psi} \log L_c(\Psi,\mathbf{z},\mathbf{y}). \quad (P_2)\end{aligned}$$

Le principe de l'algorithme EM est basé sur les deux idées suivantes :

- dans certaines situations où la distribution $P(\mathbf{x}|\,\Psi)$ est de forme simple le problème (P_2) est plus facile à résoudre que le problème (P_1),
- les données $\mathbf{z}$ n'étant pas disponibles, pour la résolution du problème (P_2) la log-vraisemblance complète du paramètre Ψ est remplacée par son espérance conditionnellement aux données observées,

$$\mathbb{E}[\log L_c(\Psi,\mathbf{Z},\mathbf{y}) \mid \mathbf{y},\Psi].$$

L'algorithme EM maximise la vraisemblance complète L_c par maximisation itérative de la fonction Q définie ainsi,

$$Q(\Psi \mid \Psi') \quad = \quad \mathbb{E}[\log L_c(\Psi,\mathbf{Z},\mathbf{y}) \mid \mathbf{y},\Psi']. \tag{2.21}$$

Soit $\Psi^{(q-1)}$ la valeur courante du paramètre, une itération de EM se divise en deux étapes

(E) : calcul des quantités ne faisant pas intervenir Ψ dans $Q(\Psi \mid \Psi^{(q-1)})$;

(M) : mise à jour du paramètre Ψ par résolution de

$$\Psi^{(q)} = \arg\max_{\Psi} Q(\Psi \mid \Psi^{(q-1)}). \tag{2.22}$$

On montre que l'estimateur solution du problème (P_1) est un point fixe de la fonction

$$\mathbf{\Phi}(\Psi') \quad = \quad \arg\max_{\Psi} Q(\Psi \mid \Psi').$$

En effet la log-vraisemblance et la log-vraisemblance complète vérifient la relation

$$\log L(\Psi,\mathbf{y}) = \log L_c(\Psi,\mathbf{x}) - \log P(\mathbf{z} \mid \mathbf{y},\Psi). \tag{2.23}$$

Le calcul de l'espérance selon la loi $P(\mathbf{z}|\, \mathbf{y},\Psi')$ des termes de part et d'autre de l'égalité (2.23) conduit à

$$\log L(\Psi,\mathbf{y}) = Q(\Psi \mid \Psi') - \mathcal{H}(\Psi \mid \Psi'), \tag{2.24}$$

où la fonction $\mathcal{H}$ est définie par

$$\mathcal{H}(\Psi \mid \Psi') \quad = \quad \mathbb{E}[\log P(\mathbf{Z} \mid \mathbf{y},\Psi) \mid \mathbf{y},\Psi'].$$

En applicant l'inégalité de Jensen, Dempster, Laird, et Rubin (1977) montrent que la fonction $\mathcal{H}(\Psi|\Psi')$ atteint son maximum pour $\Psi = \Psi'$,

$$\Psi' = \arg\max_{\Psi} \mathcal{H}(\Psi \mid \Psi'). \tag{2.25}$$

On déduit des égalités (P_1) et (2.25) que l'estimateur du maximum de vraisemblance vérifie la condition de point fixe suivante,

$$\Psi^{\mathrm{MV}} = \arg\max_{\Psi} Q(\Psi \mid \Psi^{\mathrm{MV}}).$$

L'algorithme EM peut donc être interprété comme un algorithme itératif de recherche de point fixe. Neal et Hinton (1998) ont montré que l'algorithme EM peut également être vu comme un algorithme de maximisation alternée d'une fonction des paramètres du modèle et de la distribution des données cachées. Une propriété attrayante de cet algorithme, qui découle directement de la définition de l'étape (E) et des égalités (2.24) et (2.25), est la croissante de la suite de vraisemblances générée $\{L(\Psi^{(q)},\mathbf{y})\}_{q\in\mathbb{N}}$. Ce point est crucial pour la démonstration des propriétés de convergence de l'algorithme EM. Par ses travaux, Wu (1983) a montré que sous des conditions suffisantes de régularité, l'estimateur obtenu par EM converge vers un maximum local de la vraisemblance $L(\Psi,\mathbf{y})$. Cependant, l'algorithme EM présente quelques limites : la valeur de l'estimateur à la convergence peut dépendre fortement de la position initiale et la convergence peut être très lente. L'algorithme peut se trouver bloquer dans un point selle ou un plateau de la vraisemblance $L(\Psi,\mathbf{y})$. Afin de pallier ces limitations, des versions stochastiques de l'algorithme EM ont été proposées. L'algorithme *Stochastic* EM (SEM, Broniatowski, Celeux, et Diebolt 1983) comprend ainsi une étape (S) de simulation entre les étapes (E) et (M) qui consiste à générer une segmentation $\mathbf{z}^{(q)}$ selon la distribution $P(\mathbf{z} \mid \mathbf{y},\Psi^{(q-1)})$ pour la valeur courante du paramètre. À l'étape (M) de SEM, les paramètres sont mis à jour par maximisation de la vraisemblance pour les données complètes $(\mathbf{z}^{(q)},\mathbf{y})$. La chaîne des estimateurs est une chaîne de Markov homogène. Si elle est de plus ergodique, elle converge en loi vers l'unique mesure stationnaire. L'algorithme *Simulated Annealing* EM (SAEM, Celeux et Diebolt 1990) est une variante de SEM, au cours de laquelle les paramètres sont estimés à l'étape (M) comme une moyenne pondérée entre l'estimateur EM et l'estimateur SEM. Les pondérations évoluent au cours de l'algorithme (d'où le nom de la méthode) de sorte que le poids de l'estimateur EM augmente en tendant vers 1 avec les itérations. Les algorithmes SEM et SAEM peuvent être vus comme des perturbations stochastiques de EM. Ainsi, ils respectent la dynamique de EM et on peut s'attendre à ce qu'ils ne restent pas bloqués dans des éventuels points selles, plateaux ou maxima locaux peu significatifs de la vraisemblance.

2.2.2 EM et modèle de mélange fini indépendant

L'algorithme EM trouve sa contribution la plus importante dans le contexte de l'estimation des paramètres d'un modèle de mélange. La description du problème comme un problème de données manquantes a permis d'utiliser l'algorithme qui est maintenant un outil standard. Cela a eu pour conséquence de renforcer l'intérêt pour les modèles de mélange en terme de développement méthodologique et de domaines d'application. Les structures définissant les dépendances entre les données manquantes peuvent être relativement complexes, comme pour les chaînes ou les champs de Markov. Nous décrivons ici l'algorithme EM pour le modèle simple de mélange indépendant.

Définition 2.4 *On dira qu'un système de variables $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ suit une loi de mélange fini si les variables sont identiquement distribuées suivant la loi*

$$g(y_i \mid \mu) \;=\; \sum_{k=1}^{K} \pi_k g_k(y_i \mid \theta_k), \tag{2.26}$$

où g_k est une densité paramétrée par θ_k, et les poids π_k sont tous positifs et vérifient $\sum_{k=1}^{K} \pi_k = 1$. Le paramètre μ du mélange désigne l'ensemble $\{(\pi_k,\theta_k), 1 \leq k \leq K\}$.

Le poids π_k représente donc la probabilité qu'une observation y_i de la variable Y_i soit une réalisation de la loi de densité g_k. Soit $\mathbf{y} = \{y_i, \ldots, y_n\}$ un échantillon que l'on suppose issu

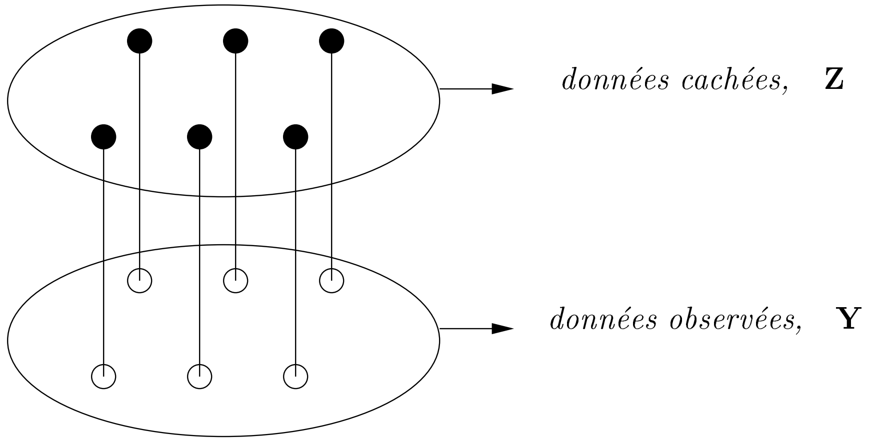


FIG. 2.2 – *représentation des dépendances entre l'ensemble des données observées* **Y** *et des données manquantes* **Z** *pour un modèle de mélange indépendant.*

d'une loi de mélange (2.26) de paramètre μ. Le problème de classification consiste à associer chaque composante y_i à une et une seule des lois de densité g_k parmi les K possibles. La classe k est formée de l'ensemble des observations associées à la loi de densité g_k. On peut interpréter la classification comme un problème de données manquantes en associant à chaque variable Y_i une variable Z_i, cachée, codant la classe de Y_i. Le poids π_k est alors égal à la probabilité $P(Z_i = e_k)$, qui ne dépend pas de i dans ce modèle. Nous appelons *modèle de mélange fini indépendant* le cas où les variables $\{Y_i,\dots,Y_n\}$ sont indépendantes conditionnellement aux classes $\{Z_i,\dots,Z_n\}$ et où les classes sont indépendantes entre elles. Les dépendances entre le système des données observées **Y** et le système des données manquantes **Z** pour ce modèle est représenté sur la figure 2.2.

Sous ce formalisme, on peut donc utiliser l'algorithme EM pour estimer les paramètres du mélange. À l'itération q, la fonction Q donnée par l'équation (2.21) s'écrit

$$Q(\mu \mid \mu^{(q-1)}) \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(q)} \log(\pi_k g_k(y_i \mid \theta_k)),$$

où le coefficient $t_{ik}^{(q)}$ désigne la probabilité $P(Z_i = e_k \mid y_i,\mu^{(q-1)})$. L'étape (E) de l'algorithme EM consiste à calculer ces coefficients selon la formule

$$\forall i \in S,\, \forall k \in \{1,\dots,K\}, \quad t_{ik}^{(q)} \;=\; \frac{\pi_k^{(q-1)} g_k(y_i \mid \theta_k^{(q-1)})}{\sum\limits_{l=1}^{K} \pi_l^{(q-1)} g_l(y_i \mid \theta_l^{(q-1)})}. \tag{2.27}$$

La résolution de l'étape (M) ne pose pas de problème. La mise à jour des poids est donnée par la formule explicite

$$\forall k \in \{1,\dots,K\}, \quad \pi_k^{(q+1)} \;=\; \frac{\sum\limits_{i=1}^{n} t_{ik}^{(q)}}{n}.$$

Le paramètre $\theta = \{\theta_k, 1 \leq k \leq K\}$ des lois de densité g_k est réestimé par la règle suivante

$$\theta^{(q)} \quad = \quad \arg\max_{\theta} \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}^{(q)} \log g_k(y_i \mid \theta_k).$$

Par exemple, dans le cas où les observations y_i sont de dimension 1 et les densités g_k sont des densités gaussiennes, les nouvelles valeurs $m_k^{(q)}$ et $(\sigma_k^2)^{(q)}$ des moyennes et variances de la classe k sont obtenues en pondérant chaque observation par sa probabilité $t_{ik}^{(q)}$ d'appartenance à cette classe

$$\forall k \in \{1,\ldots,K\}, \quad m_k^{(q)} = \frac{\sum_{i=1}^{n} t_{ik}^{(q)} y_i}{\sum_{i=1}^{n} t_{ik}^{(q)}},$$
$$\forall k \in \{1,\ldots,K\}, \quad (\sigma_k^2)^{(q)} = \frac{\sum_{i=1}^{n} t_{ik}^{(q)} (y_i - m_k^{(q)})^2}{\sum_{i=1}^{n} t_{ik}^{(q)}}.$$

Beaucoup d'auteurs se sont intéressés aux propriétés pratiques et théoriques de l'algorithme EM pour modèle de mélange fini indépendant et de ses variantes stochastiques SEM et SAEM (Redner et Walker 1984, Titterington, Smith, et Makov 1985, Celeux et Diebolt 1985, Celeux et Diebolt 1992). Redner et Walker (1984) ont montré que, sous des conditions de régularité assez peu contraignantes, l'algorithme EM converge vers l'unique estimateur du maximum de vraisemblance fortement consistant μ^{MLE} quel que soit l'initialisation de la chaîne choisie suffisamment proche de μ^{MLE}. Celeux et Diebolt (1992) ont démontré l'ergodicité de la suite des estimateurs générée par SEM et ils ont établi que la suite des estimateurs générée par SAEM converge vers un maximum local de la vraisemblance quel que soit le point de départ, pour le cas des mélanges finis indépendants.

2.2.3 Algorithme Neighborhood EM

Hathaway (1986) a montré l'équivalence entre une itération de l'algorithme EM pour modèle de mélange fini (une étape (E) suivie d'une étape (M)) et la maximisation alternée de la *vraisemblance classifiante floue* associée à ce modèle, définie par

$$D(\mu,\mathbf{c}) \quad = \quad \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log(\pi_k g(y_i \mid \theta_k)) - \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log c_{ik},$$

où $\mathbf{c} = \{c_{ik},\ i \in S,\ k \in \{1,\ldots,K\}\}$ est une matrice de classification floue dont les composantes vérifient

$$\forall i \in S, \forall k \in \{1,\ldots,K\}, \quad c_{ik} \in [0,1],$$
$$\forall i \in S, \quad \sum_{k=1}^{K} c_{ik} = 1.$$

Pour l'analyse de données spatiales, Ambroise (1996) propose d'ajouter le terme de pénalisation suivant favorisant les classements géographiquement homogènes

$$G(\mathbf{c}) = \sum_{i \sim j} c_i^t c_j,$$

conduisant au critère

$$U_\beta(\mu,\mathbf{c}) \quad = \quad D(\mu,\mathbf{c}) + \beta G(\mathbf{c}). \tag{2.28}$$

Le coefficient β, fixé, contrôle l'importance de l'information spatiale dans le nouveau critère. Ambroise (1996) suggère alors l'utilisation d'un algorithme de maximisation alternée en $\mathbf{c}$ et μ de $U_\beta(\mu,\mathbf{c})$, appelé *Neighborhood* EM (NEM, Ambroise, Dang, et Govaert 1997). Soit $\mathbf{c}^{(q-1)}$ la valeur courante de la matrice de classification floue, à l'itération q la mise à jour de celle-ci conduit à l'équation de point fixe suivante

$$\forall i \in S, \forall k \in \{1,\ldots,K\}, \quad c_{ik} \quad = \quad \frac{\pi_k g_k(y_i \mid \theta_k) \exp(\beta \sum\limits_{j \in N(i)} c_{jk})}{\sum_{l=1}^K \pi_l g_l(y_i \mid \theta_l) \exp(\beta \sum\limits_{j \in N(i)} c_{jl})}\,. \tag{2.29}$$

L'algorithme NEM permet d'ajouter sous une forme simple une information spatiale dans le traitement de données lorsque celles-ci sont représentées par un modèle de mélange fini indépendant. Il s'agit d'une solution moins complexe, mais aussi moins riche, que celle qui consiste à insérer directement la composante spatiale dans le modèle, comme pour les champs de Markov cachés.

2.2.4 EM et modèle de champ de Markov caché

Le modèle de champ de Markov caché discret (*cf.* partie 1.3) peut être interprété comme un modèle de mélange fini particulier pour lequel la classe d'une observation n'est plus indépendante de la classe des autres observations (*cf.* figures 1.3 et 2.2). On revient ainsi à la notion de sites (ou pixels pour l'image) organisés spatialement et à la notion de voisinage, qui n'existent pas dans le modèle de mélange fini indépendant. Si l'on pose $g_k(.|\theta_k) = g(.|Z_i = e_k,\theta)$ pour tout k dans $\{1,\ldots,K\}$, la loi de densité (2.26) a pour équivalent dans le modèle de champ de Markov caché la loi de densité suivante

$$g_G(y_i \mid \Psi) \quad = \quad \sum_{k=1}^K P_G(Z_i = e_k \mid \beta) g(y_i \mid Z_i = e_k,\theta).$$

Cependant, dans le modèle de champ de Markov caché, les variables manquantes sont maintenant liées par une dépendance markovienne. Pour un tel modèle, les poids des différentes classes peuvent varier avec les pixels. Par exemple, si les potentiels dépendent effectivement de i, le champ n'est pas homogène et chaque pixel n'a pas la même probabilité d'être associé à une classe particulière. Les variables Y_i ne sont pas identiquement distribuées. Il ne s'agit pas alors d'un modèle de mélange au sens de la définition 2.4. Si le modèle de champ de Markov est homogène alors on retrouve bien un mélange. Par exemple, si les potentiels $V_i(e_k)$ ne dépendent pas de i, le modèle de mélange fini défini par (2.26) est équivalent au modèle gibbsien (1.6) lorsque le paramètre spatial β est nul. Les poids des classes et les potentiels sur les singletons sont alors reliés par la formule

$$\forall k \in \{1,\ldots,K\}, \quad \pi_k \quad = \quad \frac{\exp(-V(e_k))}{\sum_{l=1}^{K} \exp(-V(e_l))},$$

où $V(e_k)$ est la valeur commune des $V_i(e_k)$.
La prise en compte de la composante spatiale avec le modèle markovien ajoute une complexité par rapport au modèle de mélange fini indépendant. Les quantités mises en jeu ne sont plus calculables directement. De plus, on va voir que l'estimation du paramètre β est bien moins évidente que celle des poids π_k du modèle de mélange fini indépendant. L'espérance $Q(\Psi \mid \Psi^{(q)})$ se décompose en deux termes,

$$Q(\Psi \mid \Psi^{(q-1)}) = Q_\theta(\theta \mid \Psi^{(q-1)}) + Q_\beta(\beta \mid \Psi^{(q-1)}), \tag{2.30}$$

où

$$Q_\theta(\theta \mid \Psi^{(q-1)}) \quad = \quad \sum_{i\in S}\sum_{z_i} P_G(z_i \mid \mathbf{y},\Psi^{(q-1)}) \log g(y_i \mid z_i,\theta), \tag{2.31}$$

$$Q_\beta(\beta \mid \Psi^{(q-1)}) \quad = \quad -\log W(\beta) - \sum_{c\in\mathcal{C}}\sum_{\mathbf{z}_c} P_G(\mathbf{z}_c \mid \mathbf{y},\Psi^{(q-1)}) V_c(\mathbf{z}_c \mid \beta). \tag{2.32}$$

À chaque itération de EM, l'estimation des paramètres θ et β se résout donc de manière indépendante. Au cours de l'étape (E), l'évaluation des termes indépendants de θ et β dans $Q_\theta(\theta| \Psi^{(q-1)})$ et $Q_\beta(\beta| \Psi^{(q-1)})$ demande de calculer les probabilités conditionnelles $P_G(z_i| \mathbf{y},\Psi^{(q-1)})$ et $P_G(\mathbf{z}_c| \mathbf{y},\Psi^{(q-1)})$. On a vu que cela n'est pas possible directement. Supposons que les probabilités markoviennes conditionnelles concernées sont connues ou estimées, alors le calcul de $Q_\theta(\theta \;|\theta^{(q-1)})$ ne pose pas de problème. Par contre, l'estimation de β apporte une difficulté supplémentaire car le calcul du maximum de Q_β fait intervenir le calcul de la constante de normalisation $W(\beta)$. Plusieurs solutions ont été proposées pour rendre les étapes (E) et (M) réalisables. Il n'est pas possible d'en faire une description exhaustive ici. Les versions de l'algorithme EM pour champ de Markov caché qui ont été inspiratrices des méthodologies et des résultats théoriques développés au cours de ce travail de thèse sont décrites dans la partie suivante. D'autres versions sont proposées dans Qian et Titterington (1991) et Qian et Titterington (1992).

2.2.5 Procédures de type EM pour champ de Markov caché

Les outils envisageables pour réaliser les étapes (E) et (M) de l'algorithme EM sont de même nature que ceux adoptés dans l'approche *image* : approximation par pseudo-vraisemblance et simulations de Monte-Carlo (MC). Si l'algorithme EM par essence ne demande pas le calcul des étiquettes des pixels pour mettre à jour les paramètres et utilise seulement les probabilités marginale et conditionnelle du champ de Markov, les approximations nécessaires pour la mise en pratique demandent de simuler des réalisations des variables cachées. Ces techniques conduisent à l'algorithme EM *gibbsien* et au *Monte-Carlo* EM.

2.2.5.1 EM gibbsien

L'algorithme EM gibbsien (Chalmond 1989) utilise l'approximation par la pseudo-vraisemblance (2.16) pour remplacer les probabilités conditionnelles $P_G(\mathbf{z}_c|\,\mathbf{y},\Psi^{(q-1)})$ dans l'étape (E). Cela permet d'éviter le calcul de la constante de normalisation à l'étape (M) pour l'estimation de β. Le terme $Q_\theta(\theta|\,\Psi^{(q-1)})$ n'est pas modifié et $Q_\beta(\beta|\,\Psi^{(q-1)})$ devient

$$\begin{aligned} Q_\beta(\beta|\,\Psi^{(q-1)}) &\approx \sum_{i\in S} I\!E_{P_G}[\log P_G(Z_i \mid \mathbf{Z}_{N(i)},\beta) \mid \mathbf{y},\Psi^{(q-1)}] \\ &\approx \sum_{i\in S}\sum_{\mathbf{z}_{\overline{N(i)}}} P_G(\mathbf{z}_{\overline{N(i)}} \mid \mathbf{y},\Psi^{(q-1)}) \, \log P_G(z_i \mid \mathbf{z}_{N(i)},\beta), \end{aligned}$$

où $\overline{N(i)} = \{N(i) \cup \{i\}\}$. Les probabilités conditionnelles $P_G(z_i|\,\mathbf{y},\ \Psi^{(q-1)})$ sont elles approximées par simulations de Monte-Carlo (simulations MC) de la manière suivante. À chaque itération q de l'algorithme, un échantillon $\mathbf{z}(q,1),\ldots,\mathbf{z}(q,T)$ de taille T est simulé selon la méthode de l'échantillonneur de Gibbs pour la valeur courante du paramètre, $\Psi^{(q-1)}$. Les distributions conditionnelles $P_G(z_i|\,\mathbf{y},\Psi^{(q-1)})$ sont estimées par la moyenne empirique,

$$\hat{P}(Z_i = e_k \mid \mathbf{y},\ \Psi^{(q-1)}) = \frac{1}{T}\sum_{t=1}^{T} I_{\{e_k\}}(z_i(q,t)).$$

À l'étape (M) le paramètre θ est mis à jour suivant les mêmes formules que pour l'algorithme EM dans lesquelles on utilise les approximations ci-dessus. Au lieu d'estimer également $P_G(\mathbf{z}_{\overline{N(i)}} \mid y,\Psi^{(q-1)})$ par la méthode de l'échantillonneur de Gibbs, pour ensuite mettre à jour le paramètre β, Chalmond (1989) utilise la formulation (2.18) de la pseudo-vraisemblance. Sous cette écriture, le paramètre β n'intervient plus. Le terme $Q_\beta(\beta|\,\Psi^{(q-1)})$ s'écrit en fonction des probabilités $P_G(k|\,v)$

$$Q_\beta(\beta|\,\Psi^{(q-1)}) \approx I\!E_{P_G}[\log \prod_{(k,v)} P_G(k \mid v)^{n_{k,v}(\mathbf{Z})} \mid \mathbf{y},\Psi^{(q-1)}].$$

À l'étape (M), les probabilités $P_G(k|\,v)$ sont estimées directement par les valeurs qui maximisent le terme de droite dans l'approximation ci-dessus. Ces estimations seront utilisées au cours de l'étape (E) de l'itération suivante pour déterminer les probabilités de transition de l'échantillonneur de Gibbs.

2.2.5.2 Monte-Carlo EM

Lorsque le calcul des probabilités conditionnelles des données manquantes dans l'étape (E) de l'algorithme EM est trop complexe pour être effectué directement, une solution consiste à utiliser une méthode dite de Monte-Carlo en générant des réalisations des distributions d'intérêt. La quantité $Q(\Psi \mid \Psi^{(q-1)})$ est approximée par sa moyenne empirique. La procédure correspondante est appelée Monte-Carlo EM (MCEM). Originellement introduite par Wei et Tanner (1990), elle ne concerne pas seulement les champs de Markov cachés, mais s'applique à tout problème de données manquantes où l'étape (E) est difficile à réaliser. Soit $T_{(q-1)}$ le nombre de réalisations générées à l'itération $q-1$ et soit $\Psi^{(q-1)}_{T_{(q-1)}}$ la valeur courante du paramètre du modèle, l'itération q de MCEM se déroule en deux étapes :

(MC) : générer T_q réalisations $\mathbf{z}(q,1), \mathbf{z}(q,2), \ldots, \mathbf{z}(q,T_q)$ selon la loi conditionnelle des données manquantes $P(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)}_{T_{(q-1)}})$. L'espérance conditionnelle $Q(\Psi \mid \Psi^{(q-1)}_{T_{q-1}})$ est alors approximée par

$$Q_{T_q}(\Psi \mid \Psi^{(q-1)}_{T_{q-1}}) = \frac{1}{T_q} \sum_{t=1}^{T_q} \log P(\mathbf{y}, \mathbf{z}(q,t) \mid \Psi).$$

et l'étape de maximisation devient

(M) : mettre à jour Ψ d'après

$$\Psi^{(q)}_{T_q} = \arg\max_{\Psi} Q_{T_q}(\Psi \mid \Psi^{(q-1)}_{T_{q-1}}).$$

On retrouve l'algorithme SEM dans le cas où $T_q = 1$ à chaque itération. Wei et Tanner (1990) recommandent d'utiliser de petites valeurs de T_q lors des premières itérations de l'algorithme puis d'augmenter T_q lorsque l'approximation courante s'approche du vrai maximum.
Dans le cas du modèle de champ de Markov caché, avec $P(\mathbf{z} \mid \mathbf{y}, \Psi) = P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$ une distribution de Gibbs, une simulation directe à l'étape (MC) est impossible et en pratique on génère T_q réalisations d'une chaîne de Markov ergodique dont la distribution invariante est $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$. Comer et Delp (2000) propose ainsi d'utiliser l'échantillonneur de Gibbs. L'étape (M) s'écrit

$$\theta^{(q)}_{T_q} = \arg\max_{\theta} \frac{1}{T_q} \sum_{t=1}^{T_q} \log g(\mathbf{y} \mid \mathbf{z}(q,t), \theta),$$
$$\beta^{(q)}_{T_q} = \arg\max_{\beta} \frac{1}{T_q} \sum_{t=1}^{T_q} \log P_G(\mathbf{z}(q,t) \mid \beta).$$

L'estimation de β nécessite le calcul de la distribution marginale du champ de Markov. Les modifications apportées par l'algorithme MCEM permettent donc de mettre à jour θ dans l'étape (M). Par contre la difficulté liée à l'estimation du paramètre β n'est pas levée. L'algorithme n'est donc pas complètement satisfaisant. L'algorithme EM gibbsien, bien qu'antérieur à l'algorithme MCEM, peut être interprété comme une amélioration de ce

dernier pour l'estimation du paramètre spatial. Comer et Delp (2000) ont appliqué MCEM pour le modèle de champ de Markov caché en supposant le paramètre β connu et appellent l'algorithme correspondant algorithme EM/MPM. Ils ont démontré des propriétés de convergence de l'algorithme dans ce cas. Nous reviendrons plus en détail sur ces résultats théoriques dans le chapitre 5, où nous étendons les résultats de Comer et Delp (2000).
Une famille d'algorithmes utilisant des simulations MC pour l'estimation des paramètres d'un champ de Markov caché et généralisant MCEM a été proposé par Qian et Titterington (1991). Supposons que $\hat{\Psi} = h(\mathbf{z},\mathbf{y})$ soit un estimateur de Ψ lorsque les données complètes sont disponibles, pour une certaine fonction h. Par exemple, la fonction h suivante définit l'estimateur du maximum de vraisemblance complète,

$$h(\mathbf{z},\mathbf{y}) = \arg\max_{\Psi} \log P_G(\mathbf{z},\mathbf{y} \mid \Psi). \tag{2.33}$$

Dans le cas où le champ $\mathbf{z}$ est inconnu, Qian et Titterington (1991) considèrent la procédure suivante, appelée *Stochastic Restoration-Estimation* (SRE). Une itération est définie par une étape (SR) dite de *restauration stochastique* et une étape (M) d'estimation par maximisation :

(SR) : générer T réalisations $\mathbf{z}(q,1), \mathbf{z}(q,2), \ldots, \mathbf{z}(q,T)$ de la distribution $P_G(\mathbf{z} \mid \mathbf{y},\Psi^{(q-1)})$;

(M) : mettre à jour le paramètre Ψ d'après

$$\begin{aligned} \text{version SRE}_\text{A} : \Psi^{(q)}_{\text{SRE}_\text{A}} &= \frac{1}{T}\sum_{t=1}^{T} h(\mathbf{z}(q,t),\mathbf{y}) \\ \mathbf{z}(q,t) \mid \Psi). & \\ \text{version SRE}_\text{B} : \Psi^{(q)}_{\text{SRE}_\text{B}} &= h(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}(q,t),\mathbf{y}) \end{aligned} \tag{2.34}$$

La version SRE_B pour la fonction h définie par (2.33) et l'algorithme MCEM sont équivalents dans le cas de modèles pour lesquels $\log P_G(\mathbf{z},\mathbf{y} \mid \Psi)$ est linéaire en $\mathbf{z}$. Les deux versions de l'algorithme SRE nécessitent en pratique d'utiliser un algorithme de simulation de la distribution de Gibbs à l'étape (SR) et de remplacer le maximum de vraisemblance par le maximum de pseudo-vraisemblance dans l'étape (M). Des variantes de l'algorithme SRE_B sont également étudiées dans Qian et Titterington (1991) et Qian et Titterington (1992) et illustrées pour la segmentation de données satellitaires à partir d'un modèle de Potts simple.

2.2.6 Segmentation non supervisée

Les algorithmes que nous avons décrits jusqu'à présent dans ce sous-chapitre sont des algorithmes de type EM pour l'estimation des paramètres d'un champ de Markov caché. Dans l'optique de retrouver les étiquettes des pixels de l'image dans un cadre non supervisé, ces

procédures forment donc la première étape de l'analyse. Une fois un estimateur $\hat{\Psi}$ obtenu, une segmentation peut être fournie par l'une des méthodes décrites dans la partie 2.1.1. Chalmond (1989) obtient ainsi une segmentation par MPM à partir des estimateurs fournis par l'algorithme EM gibbsien. L'algorithme MCEM donne une image segmentée par maximisation des fréquences empiriques calculées sur les simulations générées lors de la dernière étape (MC) réalisée. Ces estimations convergeant vers $P_G(z_i \mid \mathbf{y},\hat{\Psi})$ lorsque la longueur de la chaîne de Markov tend vers l'infini, la segmentation ainsi obtenue converge vers la segmentation MPM pour le même paramètre $\hat{\Psi}$.
L'algorithme *Point Pseudo-Likelihood* EM (PPL-EM) (Qian et Titterington 1991) est un algorithme d'estimation-maximisation alternées de type EM. Il se différencie des algorithmes de segmentation-estimation alternées de l'approche *image* par le fait que la segmentation calculée à chaque itération n'est pas considérée comme la vraie scène pour mettre à jour les paramètres. Elle n'est utilisée que pour approximer les probabilités markoviennes conditionnelles, de sorte que l'algorithme est encore un algorithme de maximisation de la vraisemblance des paramètres. Plus spécifiquement, l'itération q de l'algorithme PPL-EM est définie par une étape (R) de restauration et une étape (M) de maximisation de la *Point Pseudo-Likelihood* $\mathrm{PPL}(\mathbf{y} \mid \mathbf{z},\Psi)$,

(R) : générer une segmentation $\mathbf{z}^{(q)}$ (par exemple par ICM).

La Point Pseudo-Likelihood est l'espérance suivante

$$\begin{aligned}\mathrm{PPL}(\mathbf{y} \mid \mathbf{z}^{(q)},\Psi) &\approx \prod_{i\in S} \mathbb{E}_{P_G}[g(y_i \mid Z_i,\theta) \mid \mathbf{Z}_{N(i)} = \mathbf{z}^{(q)}_{N(i)},\Psi], \\ &= \prod_{i\in S} \sum_{z_i} P_G(z_i \mid \mathbf{z}^{(q)}_{N(i)},\beta) g(y_i \mid z_i,\theta). \qquad (2.35)\end{aligned}$$

On voit que la réalisation $\mathbf{z}^{(q)}$ n'est utilisée que comme une segmentation partielle, puisque seules les configurations des voisinages $\mathbf{z}^{(q)}_{N(i)}$ sont supposées exactes. On reconnaît dans l'espérance (2.35) la distribution d'un système $\mathbf{Y}$ de variables indépendantes conditionnellement à leur classe et pour un modèle où les variables $\{Z_i, \ldots ,Z_n\}$ sont indépendantes entre elles. On est donc ramené à un modèle de mélange fini indépendant particulier dont les poids sont donnés par les probabilités $P_G(z_i \mid \mathbf{z}^{(q)}_{N(i)},\beta)$ et peuvent donc dépendre de i. L'étape (M) de l'algorithme PPL-EM consiste à

(M) : mettre à jour Ψ d'après

$$\Psi^{(q)} = \arg\max_{\Psi} PPL(\mathbf{y} \mid \mathbf{z}^{(q)},\Psi).$$

En pratique, Qian et Titterington (1991) utilisent quelques itérations de l'algorithme EM pour modèle de mélange fini indépendant pour réaliser l'étape (M). Les auteurs constatent que le fait de ne considérer que les étiquettes $\mathbf{z}^{(q)}_{N(i)}$ des voisins d'un pixel comme exactes, mais pas celle du pixel lui-même, pour estimer les paramètres conduit à de meilleurs résultats qu'une estimation par ICM non supervisé. Lorsqu'un nombre suffisant d'itérations de PPL-EM a été réalisé pour que l'estimateur des paramètres ait convergé, le champ courant $\mathbf{z}^{(q)}$ fournit alors une segmentation des données.

On peut envisager de se ramener à l'approche *image* pour la segmentation non supervisée à partir de l'algorithme EM en ajoutant une étape (C) de classification entre les étapes (E) et (M). Il s'agit de l'algorithme *Classification* EM (CEM) (Celeux et Govaert 1992). Au cours de l'étape (C), une segmentation $\mathbf{z}^{(q)}$ est générée suivant la règle du MAP. Pour le modèle de mélange fini indépendant, cela correspond à maximiser les coefficients $t_{ik}^{(q)}$ puisqu'il y a indépendance entre les étiquettes Z_i. Pour le modèle de champ de Markov caché, on a vu que l'estimateur du MAP n'est pas accessible directement. Il faut utiliser l'algorithme de recuit simulé ou ICM. L'étape (M) devient

$$\Psi^{(q)} = \arg\max_{\Psi} \log P_G(\mathbf{y}, \mathbf{z}^{(q)} \mid \Psi).$$

L'algorithme CEM converge vers un maximiseur de la vraisemblance complète des paramètres, et non plus de la vraisemblance comme EM. Ainsi, même s'il s'agit d'une variante de l'algorithme EM, il se place dans l'approche *image*.

2.3 Conclusion

Nous avons présenté les principales méthodes pour l'estimation des paramètres d'un modèle de champ de Markov caché et pour la segmentation markovienne d'images. Nous les avons classées dans deux approches différentes, l'approche *image* et l'approche *mélange*. La première approche fournit des algorithmes de segmentation-estimation alternées et fait appel à des techniques d'estimation qui demandent de connaître les étiquettes des pixels de l'image. Ces méthodes associent ainsi une segmentation basée sur les principes du MAP, de ICM ou du MPM et une estimation par maximum de la vraisemblance complète des paramètres. En terme de pertinence de la segmentation, les algorithmes correspondants fournissent des images satisfaisantes, parfois trop lisses. Cependant, les limites de ces méthodes apparaissent avec l'étape d'estimation. En effet, l'utilisation d'une segmentation obtenue par une méthode de la famille *image* pour le calcul du maximum de vraisemblance induit un biais sur les estimateurs des paramètres. Pour des observations qui se trouvent dans la zone de recouvrement des densités de deux classes, la restauration des étiquettes revient à séparer les données en deux groupes correspondant aux deux moitiés gauche et droite de la zone de recouvrement. Cela conduit à une surestimation de la distance entre les moyennes des deux densités. Ce problème est soulevé notamment dans Titterington (1984) dans le cas d'un modèle de mélange indépendant. L'auteur recommande de maximiser la vraisemblance au lieu de maximiser la vraisemblance complète. Dans ce cas, la méthode d'estimation ne nécessite pas de disposer d'une segmentation. On retrouve alors les méthodes de l'approche *mélange*, qui sont donc des méthodes d'estimation. Les algorithmes correspondants sont l'algorithme EM et ses variantes stochastiques. Ils utilisent les probabilités marginale et conditionnelle du champ de Markov, ce qui permet des fluctuations autour de la segmentation des observations lorsque celle-ci n'est pas sûre. Cette approche conduit à de meilleurs estimateurs que l'approche *image*. Cependant, la complexité des modèles de champs de Markov cachés impose en pratique de travailler avec des approximations obtenues par simulations MC. Ces méthodes d'approximation sont précises car elles ont de bonnes propriétés de convergence et les algorithmes correspondants ne perdent pas la qualité des algorithmes exacts. Cependant, ce sont des procédures lourdes car à chaque itération des simulations sont réalisées, notamment en comparaison de l'algorithme ICM non supervisé, algorithme de référence dans l'approche *image*. Le schéma de la figure 2.3 reprend les principales méthodes de l'approche *mélange* et de l'approche *image* en soulignant les différences dans les principes adoptés par chacune.

Parmi les méthodes couramment utilisées pour la segmentation markovienne non supervisée il y a donc un choix à faire entre précision des estimateurs et rapidité de l'algorithme d'estimation. Dans ce travail de thèse nous avons choisi de nous intéresser à l'approche *mélange*, car l'utilisation de l'algorithme EM pour l'estimation des paramètres permet d'éviter le problème de biais constaté avec les méthodes de l'approche *image*. La difficulté réside à l'étape (E) dans le calcul des probabilités markoviennes conditionnelles et à l'étape (M) dans le calcul des probabilités markoviennes marginales lorsque le paramètre spatial doit être estimé. Notre objectif est de développer de nouvelles procédures s'inscrivant dans l'approche *mélange* et telles que d'une part les principes d'approximation utilisés préservent la qualité des résultats fournis par EM et d'autre part les algorithmes correspondants soient plus rapides que les algorithmes basés sur des simulations MC. Pour cela nous avons étudié puis généralisé la méthode d'approximation en champ moyen d'un champ de Markov (Chandler 1987) utilisé par Zhang (1992) dans l'algorithme EM avec des résultats encourageants. Les deux chapitres suivants décrivent le principe du champ moyen et la méthodologie développée autour de ce principe pour l'estimation des paramètres d'un modèle de champ de Markov caché et la segmentation.

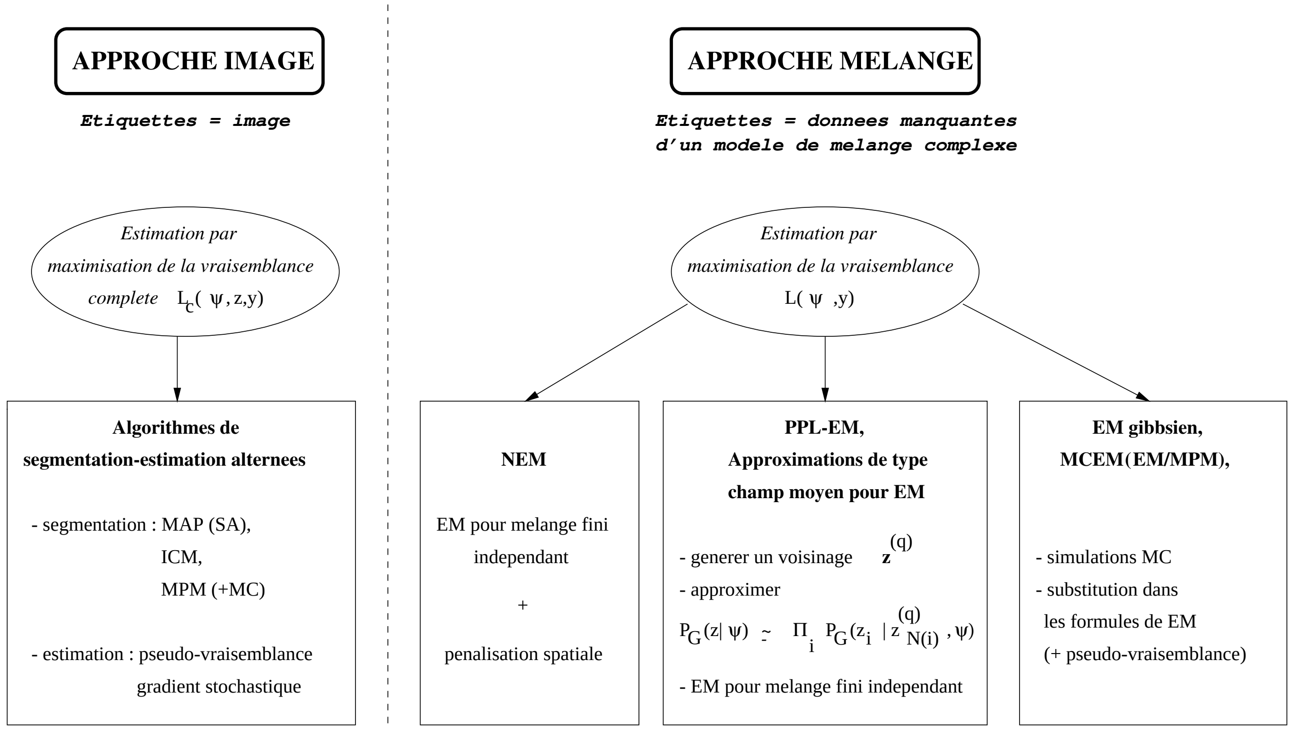


FIG. 2.3 – *Approches "image" et "mélange" pour l'estimation des paramètres d'un modèle de champ de Markov caché.*

Chapitre 3

Principe du champ moyen

Le champ moyen est à l'origine une méthode d'approximation pour le calcul de l'espérance d'un champ de Markov. Évaluer cette espérance est un problème de complexité équivalente à celle du calcul de la constante de normalisation du modèle. Cette méthode est issue de la mécanique statistique (*e.g.* Chandler 1987, Parisi 1988) où elle est utilisée comme outil pour l'étude des phénomènes de transition de phase. Plus récemment, les applications se sont étendues au domaine de la vision par ordinateur pour la reconstruction de surfaces (*e.g.* Geiger et Girosi 1991) ou la détection de contours (*e.g.* Zerubia et Chellappa 1993), au domaine des modèles graphiques (*e.g.* Jaakkola et Jordan 1997, Weiss 2000 ou Yuille 2001) ou encore des réseaux de neurones (*e.g.* Peterson et Anderson 1987, Kappen et Rodriguez 1998, Humphreys et Titterington 2000). L'approximation en champ moyen fournit également une approximation de la distribution d'un champ de Markov. Plus précisement, la méthode permet d'approximer un système markovien, avec des interactions complexes, par un système de variables indépendantes, plus facile à manipuler.
Dans la partie 3.1 nous présentons le principe du champ moyen et nous décrivons les approximations correspondantes de l'espérance et de la distribution d'un champ de Markov. Nous mettons ensuite en avant dans la partie 3.2 l'origine variationnelle de l'approximation en champ moyen et l'optimalité de cette méthode en un sens que nous précisons. Il en découle une approximation en champ moyen à l'ordre 1 de la fonction de partition d'une distribution de Gibbs, que nous comparons à l'approximation à l'ordre zéro sur un exemple simple dans la partie 3.3. Les extensions et variantes autour du principe du champ moyen qui ont été étudiées dans le cadre des champs de Markov cachés sont décrites dans la partie 3.4.

3.1 Champ de Markov et champ moyen

3.1.1 Principe

En physique, la théorie microscopique des systèmes de particules soumis à des interactions (*e.g.* gaz non parfait), décrits par une distribution de Gibbs, montre l'existence de transition de phase dans les modèles. L'étude de tels systèmes hautement corrélés conduit à

considérer des approximations, par exemple pour le calcul de l'entropie. Une méthode fondatrice est l'approximation en champ moyen (Chandler 1987, Parisi 1988). Cette méthode simple consiste, lorsque l'on se focalise sur une particule, à supposer que le rôle des particules voisines est identique à celui d'un champ moléculaire moyen qui agirait sur la particule cible. Cela revient à négliger les fluctuations des particules qui interagissent avec cette particule cible, en fixant celles-ci à leur valeur moyenne. Cette hypothèse est raisonnable lorsque le système est à l'équilibre car alors les fluctuations se compensent effectivement entre les particules. Cette méthode n'est jamais exacte pour des petites dimensions et pour des interactions de faible portée, mais elle est souvent de bonne qualité et sa simplicité fait qu'elle est largement utilisée. En effet, il est facile de voir (et nous le détaillons dans la suite) que le système résultant de l'approximation est un système de particules indépendantes.
L'équivalence entre distribution de Gibbs et champ de Markov permet d'appliquer le principe du champ moyen au calcul de l'espérance d'un champ de Markov. Regardons comment sont approximées la fonction énergie H associée au champ de Markov, puis l'espérance $\mathbb{E}_{P_G}[Z_i]$ avec cette méthode. Soit i le pixel cible de l'ensemble S, pour tout j différent de i, les variables Z_j sont fixées à leur valeur moyenne $m_j = \mathbb{E}_{P_G}[Z_j]$. Notons $\mathbf{m}$ le champ des valeurs moyennes, $\mathbf{m} = \{m_i, i \in S\}$. Remplacer les réalisations z_j par leur valeur moyenne m_j dans l'expression de la fonction énergie (1.4) conduit à la définition d'une nouvelle fonction énergie pour le pixel i,

$$H_i^*(z_i) \;=\; H(\mathbf{z})|_{z_j=m_j, j\neq i}\,. \tag{3.1}$$

Remarquons que les moyennes m_j sont des grandeurs continues. Il est donc nécessaire pour pouvoir appliquer le principe du champ moyen dans le cas d'un champ de Markov discret que la fonction énergie H, a priori définie sur le sous-ensemble $V^n = \{e_1, e_2, \ldots e_K\}^n$ de $\{0,1\}^{K\times n}$, puisse être étendue à $\mathcal{V}^n$ pour $\mathcal{V} = \{\mathbf{u} \in [0,1]^K,\ \sum_{k=1}^K u_k = 1\}$. Pour la même raison, il est pratique de considérer que les variables Z_j sont à valeurs dans $\mathcal{V}$ plutôt que dans V. À la fonction énergie (3.1) est associée de manière naturelle une mesure de probabilité P_i^* concentrée sur la variété $\{Z_j = m_j, \forall\, j \neq i\}$,

$$P_i^*(\mathbf{z}) \;=\; W_i^{*\,-1} \exp(-H_i^*(z_i)) \prod_{j\neq i} I_{\{m_j\}}(z_j), \tag{3.2}$$

où $W_i^* \;=\; \sum_{z_i} \exp(-H_i^*(z_i))$. Dans l'expression (1.4), les termes dépendant de z_i peuvent être isolés. Cela conduit à la décomposition de l'énergie en champ moyen (3.1) au pixel i en un terme $H_i^{*loc}(z_i)$ correspondant à l'énergie en champ moyen locale au pixel i et un terme $R_i^{*loc}(\mathbf{m}_{S\backslash\{i\}})$ indépendant de z_i,

$$H_i^*(z_i) \;=\; H_i^{*loc}(z_i) + R_i^{*loc}(\mathbf{m}_{S\backslash\{i\}}). \tag{3.3}$$

La constante de normalisation locale correspondante est

$$W_i^{*loc} \;=\; \sum_{z_i} \exp(-H_i^{*loc}(z_i)). \tag{3.4}$$

Le principe du champ moyen suggère que la distribution marginale du champ au pixel i,

$$P_G(z_i) \;=\; W^{-1} \sum_{\mathbf{z}_{S\setminus\{i\}}} \exp(-H(\mathbf{z})),$$

peut être approximée par

$$\begin{aligned} P_i^*(z_i) \;&=\; W_i^{*-1} \exp(-H_i^*(z_i)) \\ &=\; W_i^{*loc^{-1}} \exp(-H_i^{*loc}(z_i)), \end{aligned} \tag{3.5}$$

qui est aussi la distribution conditionnelle de Z_i sachant $\mathbf{Z}_{N(i)} = \mathbf{m}_{N(i)}$,

$$P_i^*(z_i) \;=\; P_G(z_i \mid \mathbf{m}_{N(i)}). \tag{3.6}$$

3.1.2 Condition de cohérence

Pour utiliser l'approximation définie par (3.5) il est nécessaire de connaître les valeurs moyennes m_j pour les pixels différents de i. Or, ces moyennes sont inconnues et l'objectif de l'approximation en champ moyen est justement de les approximer. Soit z_i^{mf} l'approximation en champ moyen de m_i. Les approximations en champ moyen de la fonction énergie, de la fonction de partition et de la distribution de Gibbs s'obtiennent en remplaçant dans les expressions (3.1) à (3.6) les valeurs moyennes exactes $\{m_j, j \in S\}$ par $\{z_j^{mf}, j \in S\}$ Par exemple, $H_i^{mf}(z_i)$, la fonction énergie au pixel i sous l'approximation en champ moyen est définie par

$$H_i^{mf}(z_i) \;=\; H(\mathbf{z})\big|_{z_j = z_j^{mf}, j \neq i},$$

et l'on définit de la même manière les quantités W_i^{mf}, H_i^{mfloc} ,W_i^{mfloc}, $P_i^{mf}(\mathbf{z})$ et $P_i^{mf}(z_i)$. Soit $\mathbb{E}_i^{mf}$ l'espérance pour la loi P_i^{mf}. La méthode d'approximation doit vérifier une condition de cohérence : les valeurs moyennes calculées à partir de l'approximation en champ moyen doivent être égales aux valeurs moyennes utilisées pour définir l'approximation. Pour les pixels j différents de i, il est clair d'après (3.2) que $\mathbb{E}_i^{mf}[Z_j] = z_j^{mf}$. L'approximation est donc cohérente. Pour le pixel i on a

$$\mathbb{E}_i^{mf}[Z_i] \;=\; W_i^{mf^{-1}} \sum_{z_i} z_i \exp(-H_i^{mf}(z_i)) \;=\; W_i^{mfloc^{-1}} \sum_{z_i} z_i \exp(-H_i^{mfloc}(z_i)). \tag{3.7}$$

La dernière expression est une fonction de $\{z_j^{mf}, j \in N(i)\}$ que nous notons

$$h_i(\{z_j^{mf}, j \in N(i)\}).$$

La condition de cohérence conduit à

$$z_i^{mf} = \mathbb{E}_i^{mf}[Z_i] = h_i(\{z_j^{mf}, j \in N(i)\}). \tag{3.8}$$

Si l'on répète la démarche pour tous les pixels de S on obtient l'équation ci-dessous

$$\mathbf{z}^{mf} \quad = \quad h(\mathbf{z}^{mf}) \quad = \begin{cases} h_1(\{z_j^{mf}, j \in N(1)\}) \\ \vdots \\ h_n(\{z_j^{mf}, j \in N(n)\}) \end{cases} . \tag{3.9}$$

L'approximation en champ moyen consiste à résoudre ce problème de point fixe. La solution $\mathbf{z}^{mf} = \{z_i^{mf}, i \in S\}$ est alors l'estimation en champ moyen de l'espérance $\mathbf{m}$. L'existence et l'unicité des solutions de (3.9) ont été étudiées par Wu et Doerschuk (1995). Les propriétés établies par les auteurs sont les suivantes dans le cas d'une fonction énergie $H(\mathbf{z}) = \beta H'(\mathbf{z})$

- s'il existe une solution à l'équation (3.9) alors cette solution est dans $\mathcal{V}^{|S|}$;
- si la fonction énergie H' est continue sur $[0,1]^{K\times n}$ alors il existe au moins une solution à (3.9) ;
- sous des conditions de régularité sur la fonction h et pour des valeurs de β suffisament petites la fonction h est contractante et l'équation (3.9) a une unique solution.

Lorsque les conditions assurant la dernière propriété sont vérifiées, l'unique solution peut être calculée de manière itérative.

Exemple 3.1 (Modèle de Potts simple à K couleurs) *En adoptant la notation vectorielle pour les variables z_i, la distribution du modèle de Potts à K couleurs (voir partie 1.2) s'écrit*

$$\begin{aligned} P_G(\mathbf{z} \mid \beta) &= W(\beta)^{-1} \exp(\beta \sum_{i \sim j} z_i^t z_j), \\ &= W(\beta)^{-1} \exp(\frac{\beta}{2} \sum_i z_i^t \sum_{j \in N(i)} z_j). \end{aligned} \tag{3.10}$$

L'énergie en champ moyen locale $H_i^{mfloc}(z_i)$ pour ce modèle est égale à

$$H_i^{mfloc}(z_i) \quad = \quad \beta z_i^t \sum_{j \in N(i)} z_j^{mf}, \tag{3.11}$$

et l'équation de cohérence devient

$$\forall i \in S \;, \quad z_i^{mf} \quad = \quad \frac{\sum\limits_{z_i} z_i \exp\{\beta z_i^t \sum\limits_{j \in N(i)} z_j^{mf}\}}{\sum\limits_{z_i} \exp\{\beta z_i^t \sum\limits_{j \in N(i)} z_j^{mf}\}}. \tag{3.12}$$

Parisi (1988) montre que l'approximation en champ moyen de l'espérance d'un modèle d'Ising (modèle équivalent au modèle de Potts simple à 2 couleurs) devient exacte dans le cas d'interactions faibles mais à portée infinie. De manière intuitive, l'idée est que lorsque

la connectivité entre les pixels est grande il intervient des phénomènes de moyennisation qui rendent un pixel relativement insensible aux fluctuations de ses voisins. Plus concrètement, dans le cas du modèle d'Ising sans champ externe ou du modèle de Potts simple, les valeurs moyennes m_j sont identiques en tout pixel et égales à m_1 par exemple. D'après la loi des grands nombres, l'influence des pixels voisins, représentée par le terme $\sum_{j\in N(i)} z_j$, tend vers une influence moyenne $N_v m_1$ lorsque le cardinal N_v du voisinage $N(i)$ (que l'on suppose ici indépendant de i) tend vers l'infini.
Une solution triviale de l'équation (3.12) est la configuration $\mathbf{u}^K$ uniforme en chaque pixel, définie par $u_i^K = (K^{-1},\ldots,K^{-1})^t$ pour tout i dans S. Ce n'est pas une configuration à proprement parler puisqu'elle n'appartient pas à V^n. Dans le cas des voisinages classiques d'ordre 1 et d'ordre 2, le point fixe $\mathbf{u}^K$ de la fonction g est stable si $\beta \leq \frac{K}{N_v}$ et instable si $\beta > \frac{K}{N_v}$ (voir Elfadel et Yuille 1992 et annexe A). Dans le cas de deux couleurs, il existe un unique point fixe lorsque $\beta \leq \frac{2}{N_v}$ et si $\beta > \frac{2}{N_v}$ il existe en plus deux points fixes symétriques (voir par exemple Parisi 1988). Ambroise et Govaert (1998) ont démontré l'unicité du point fixe et la convergence de la procédure itérative $\mathbf{z}^{q+1} = h(\mathbf{z}^{q+1})$ vers ce point fixe lorsque le paramètre spatial β est inférieur à $\frac{1}{N_v}$, pour un voisinage et un nombre de couleurs K quelconques. Nous avons montré que ces propriétés sont encore vérifiées lorsque $\beta < \frac{4}{N_v K}$, ce qui améliore la borne lorsque $K \leq 3$ (voir annexe A). Cependant cela conduit à une solution peu intéressante puisqu'il s'agit alors du point fixe $\mathbf{u}^K$ et la distribution $P_i^{mf}(z_i)$ correspondante est la distribution uniforme sur l'ensemble V, indépendamment de la valeur du paramètre spatial.

L'existence du point fixe $\mathbf{u}^K$ dans l'exemple ci-dessus est due au fait que le modèle de Potts simple ne possède pas de potentiels sur les singletons (*i.e.* de champ externe). Le champ uniforme $\mathbf{u}^K$ n'est plus solution du problème de point fixe dès que l'on enlève l'hypothèse d'équiprobabilité des couleurs dans le modèle. C'est en particulier le cas du champ conditionnel $\mathbf{Z} \mid \mathbf{y}$ dans un modèle de champ de Markov caché, le champ externe apparaissant dans le terme d'attache aux données.

Exemple 3.2 (Modèle de Potts à K couleurs avec champ externe) *Considérons l'extension du modèle défini par (3.10) dans laquelle les potentiels sur les singletons ne sont pas nuls. Cela conduit à la distribution de Gibbs suivante,*

$$P_G(\mathbf{z} \mid \beta,\mathbf{V}) \;=\; W(\beta,\mathbf{V})^{-1} \exp\{\sum_{i\in S} V_i(z_i) + \beta \sum_{i\sim j} z_i^t z_j\}, \tag{3.13}$$

où $\mathbf{V} = \{V_i(e_k),\ i \in S,\ k \in \{1,\ldots,K\}\}$ est une matrice réelle de taille de $n \times K$. Dès que les coefficients de $\mathbf{V}$ sont tous égaux, on retrouve le modèle de Potts simple. Pour le modèle (3.13) l'équation de cohérence (3.12) est ainsi modifiée,

$$\forall i \in S\ , \quad z_i^{mf} \;=\; \frac{\sum_{z_i} z_i \exp\{V_i(z_i) + \beta z_i^t \sum_{j\in N(i)} z_j^{mf}\}}{\sum_{z_i} \exp\{V_i(z_i) + \beta z_i^t \sum_{j\in N(i)} z_j^{mf}\}}\,. \tag{3.14}$$

Comme pour le modèle sans champ externe, nous avons établi que la fonction h associée au problème de point fixe (3.14) est contractante si $\beta < \frac{4}{N_v K}$, *quel que soit le système de voisinage et le nombre K de classes. Pour de telles valeurs du paramètre spatial la solution de ces équations existe et est unique et la suite* $\mathbf{z}^{q+1} = h(\mathbf{z}^{q+1})$ *converge vers l'unique point fixe de h (voir annexe A).*

3.1.3 Approximation en champ moyen d'une distribution de Gibbs

Il découle des expressions établies dans la partie 3.1.1 que l'approximation en champ moyen de la distribution de Gibbs $P_G(\mathbf{z})$ est

$$\begin{aligned} P^{mf}(\mathbf{z}) &= \prod_{i \in S} P_i^{mf}(z_i) \\ &= {W^{mf}}^{-1} \exp(-H^{mf}(\mathbf{z})), \end{aligned}$$

avec

$$H^{mf}(\mathbf{z}) = \sum_{i \in S} H_i^{mfloc}(z_i), \tag{3.15}$$

et $H_i^{mfloc}(z_i)$ définie par le découpage (3.3), en remplaçant m_i par z_i^{mf} pour tout i dans S. Il apparaît clairement à travers l'expression de la fonction énergie en champ moyen (3.15) et l'expression de la distribution correspondante que le système sous P^{mf} se comporte comme un système indépendant. La fonction de partition du modèle se factorise

$$W^{mf} = \sum_{\mathbf{z}} \exp\{-H^{mf}(\mathbf{z})\} = \prod_{i \in S} W_i^{mfloc}, \tag{3.16}$$

avec W_i^{mfloc} donnée par (3.4) en remplaçant m_i par z_i^{mf} pour tout i dans S. La difficulté de calcul inhérente au modèle exact est donc supprimée dans l'approximation en champ moyen. Remarquons que dans le cas d'un champ de Markov discret à K couleurs les composantes de l'approximation en champ moyen $z_i^{mf} = \mathbb{E}^{mf}[Z_i]$ sont les probabilités marginales de la variable Z_i sous la distribution P^{mf} (*cf.* équations (3.7) et (3.8))

$$z_i^{mf} = \left\{ \begin{array}{c} P^{mf}(Z_i = e_1) \\ \cdots \\ P^{mf}(Z_i = e_K) \end{array} \right\} = \left\{ \begin{array}{c} P_G(Z_i = e_1 \mid \mathbf{Z}_{N(i)} = \mathbf{z}^{mf}_{N(i)}) \\ \cdots \\ P_G(Z_i = e_K \mid \mathbf{Z}_{N(i)} = \mathbf{z}^{mf}_{N(i)}) \end{array} \right\}. \tag{3.17}$$

Exemple 3.3 (Modèle de Potts à K couleurs avec champ externe, suite) *Reprenons l'exemple 3.2. L'énergie en champ moyen locale est égale à*

$$H_i^{mfloc}(z_i) = \sum_{i \in S} (V_i(z_i) + \beta z_i^t \sum_{j \in N(i)} z_j^{mf}) ,$$

et

$$P^{mf}(Z_i = e_k) \;=\; \frac{\exp(V_i(e_k) + \beta e_k^t \sum\limits_{j \in N(i)} z_j^{mf})}{\sum_{l=1}^K \exp(V_i(e_l) + \beta e_l^t \sum\limits_{j \in N(i)} z_j^{mf})} \, .$$

Dans le cas particulier de 2 couleurs, on peut définir $p_i = P_i^{mf}(Z_i = e_1) = 1 - P_i^{mf}(Z_i = e_2)$ *qui s'écrit encore*

$$\forall i \in S, \quad p_i \;=\; [1 + \exp\{V_i(e_1) - V_i(e_2) + \beta(|N(i)| - 2 \sum_{j \in N(i)} p_j)\}]^{-1}.$$

L'approximation en champ moyen de $P_G(\mathbf{z} \mid \beta)$ *est alors*

$$P^{mf}(\mathbf{z}) = \prod_{i \in S} p_i^{z_i^t e_1} (1 - p_i)^{z_i^t e_2}.$$

3.2 Approche variationnelle

L'approximation en champ moyen se justifie en introduisant un principe variationnel issu de la mécanique statistique : le principe de minimisation de la fonctionnelle *énergie libre* $F(P)$ (voir par exemple Parisi 1988), également appelée *énergie libre variationnelle* dans certaines références. Étant donné $P(\mathbf{z})$ une distribution de probabilité sur l'espace Ω des configurations sur S, et une fonction énergie $H(\mathbf{z})$, la fonctionnelle $F(P)$ est définie par

$$F(P) \;=\; \mathbb{E}_P[H(\mathbf{Z})] - \mathbf{S}(P) \; , \tag{3.18}$$

où $\mathbb{E}_P$ est l'espérance pour la distribution P et $\mathbf{S}(P) = -\mathbb{E}_P[\log P(\mathbf{Z})]$ est l'entropie associée à P. Si P_G est la distribution de Gibbs associée à H, comme dans (1.3), il vient

$$\begin{aligned} F(P) \;&=\; -\log W + \mathbb{E}_P[\log(\frac{P(\mathbf{Z})}{P_G(\mathbf{Z})})] \\ &=\; -\log W + KL(P,P_G) \; , \end{aligned}$$

où $KL(P_1,P_2) = \mathbb{E}_{P_1}[\log(\frac{P_1(\mathbf{Z})}{P_2(\mathbf{Z})})]$ est la divergence de Kullback-Leibler entre deux distributions P_1 et P_2 de même support. Cette quantité est positive et ne s'annule que si les distributions P_1 et P_2 sont identiques. À partir de ces propriétés, il est facile de voir que la distribution qui minimise $F(P)$ est la distribution de Gibbs P_G. L'énergie libre minimale vaut donc

$$F(P_G) = -\log W. \tag{3.19}$$

Considérons maintenant, non plus le vrai minimum de l'énergie libre (ou de manière équivalente le minimum de la divergence de Kullback-Leibler), mais le minimum parmi l'ensemble des distributions de probabilité $P^\pi(\mathbf{z})$ qui ont la propriété de factorisation : $P^\pi(\mathbf{z}) =$

$\prod_{i \in S} P_i^\pi(z_i)$. On se restreint ainsi à des systèmes de variables indépendantes. Plus spécifiquement, cela revient à considérer les systèmes de variables dont la fonction énergie associée est de la forme

$$H^\pi(\mathbf{z}) \;=\; \sum_{i \in S} z_i^t [V_i + \Delta V_i] \;, \tag{3.20}$$

où les vecteurs $V_i = \{V_i(e_1), \ldots, V_i(e_K)\}, i \in S$ sont les fonctions potentiels sur les singletons de l'énergie H et où la quantité ΔV_i doit être identifiée. Soit $\mathbb{E}^\pi$ l'espérance pour la distribution définie par (3.20). Dans le cas d'une distribution de Gibbs P_G avec interaction d'ordre au plus 2, l'expression de l'approximation en champ moyen P^{mf} obtenue dans la partie précédente correspond à $\Delta V_i = \{\sum_{j \in N(i)} V_{ij}(e_k, \mathbb{E}^{mf}[Z_j]), 1 \leq k \leq K\}$. Pour les systèmes avec une fonction énergie (3.20), la fonctionnelle F s'écrit

$$F(P^\pi) \;=\; -\log W^\pi + \mathbb{E}^\pi[H(\mathbf{Z}) - H^\pi(\mathbf{Z})] \;, \tag{3.21}$$

où W^π est la fonction de partition du modèle défini par (3.20). Les quantités ΔV_i satisfaisant le principe d'énergie libre minimale peuvent être identifiées par annulation du gradient de (3.21). On retrouve alors le problème de point fixe (3.9), et la solution du problème de minimisation de l'énergie libre sur l'ensemble restreint des distributions factorisables est la distribution P^{mf}. Ce choix est donc optimal au sens de la divergence de Kullback-Leibler : pour ce critère c'est la meilleure approximation de P_G par un système de variables indépendantes. Une approche équivalente consiste à apliquer la théorie de la perturbation dynamique (*e.g* Chandler 1987). Le problème du calcul de l'énergie libre minimale est équivalent au problème du calcul de la fonction de partition W d'après l'égalité (3.19). Prenons comme modèle de référence (ou modèle d'ordre zéro) le modèle P^π défini par l'énergie $H^\pi(\mathbf{z})$ et considérons la perturbation de l'énergie $\Delta H = H - H^\pi$. La fonction de partition de la distribution de Gibbs P_G se factorise de manière exacte comme suit,

$$\begin{aligned} W &= \sum_{\mathbf{z}} \exp\{-[H^\pi(\mathbf{z}) + \Delta H(\mathbf{z})]\}, \\ &= W^\pi \mathbb{E}^\pi[\exp\{-\Delta H(\mathbf{Z})\}]. \end{aligned} \tag{3.22}$$

Dans une approche en champ moyen, on suppose que les fluctuations sont négligeables et donc d'une certaine façon que la variation ΔH est petite. Sous cette hypothèse, il s'agit alors de déterminer les coefficients ΔV_i de sorte à avoir la meilleure approximation de W à l'ordre 1. L'espérance dans le terme de droite de l'équation (3.22) peut s'écrire sous une forme développée

$$\begin{aligned} \mathbb{E}^\pi[\exp\{-\Delta H(\mathbf{Z})\}] &= \mathbb{E}^\pi[1 - \Delta H(\mathbf{Z}) + \frac{\Delta H(\mathbf{Z})^2}{2} + \ldots], \\ &= 1 - \mathbb{E}^\pi[\Delta H(\mathbf{Z})] + \frac{1}{2}\mathbb{E}^\pi[\Delta H(\mathbf{Z})^2] + \ldots \quad . \end{aligned} \tag{3.23}$$

Puisque ΔH est petit, en négligeant les termes d'ordre supérieur à 1, on obtient le résultat suivant :

$$\mathbb{E}^\pi[\exp(-\Delta H(\mathbf{Z}))] \approx \exp\{-\mathbb{E}^\pi[\Delta H(\mathbf{Z})]\},$$

et

$$W \approx W^{\pi} \exp\{-\mathbb{E}^{\pi}[H(\mathbf{Z}) - H^{\pi}(\mathbf{Z})]\}. \tag{3.24}$$

Cette approximation constitue également une borne inférieure de la fonction de partition W. En effet, on obtient facilement d'après (3.22) et la convexité de la fonction exponentielle l'inégalité suivante, dite *borne de Gibbs-Bogoliubov-Feynman* (GBF)

$$W \geq \underline{W}^{\pi} = W^{\pi} \exp\{-\mathbb{E}^{\pi}[H(\mathbf{Z}) - H^{\pi}(\mathbf{Z})]\}. \tag{3.25}$$

Le développement (3.23) suggère une autre approximation de la fonction de partition,

$$W \approx W^{\pi}(1 - \mathbb{E}^{\pi}[\Delta H(\mathbf{Z})]).$$

Cependant, en utilisant la propriété $e^x \geq 1 + x$ pour tout x, on montre que cette approximation est moins fine que celle donnée par (3.25).
Une minoration de W, optimale au sens de la borne de GBF parmi les modèles ayant une fonction énergie de la forme (3.20), est déterminée par maximisation du terme de droite de l'inégalité (3.25) sur l'ensemble de ces modèles. Ce problème est équivalent au problème de minimisation de l'énergie libre sur l'ensemble des distributions factorisables P^{π} puisque $\underline{W}^{\pi} = e^{-F(P^{\pi})}$. La solution est donc $P^{\pi} = P^{mf}$. Soit W^{GBF} l'approximation à l'ordre 1 $\underline{W}^{\pi}$ pour $\pi = P^{mf}$. Cette quantité vérifie $W \geq W^{\mathrm{GBF}} \geq \underline{W}^{\pi}$ pour toute distribution P^{π} factorisable. On peut espérer qu'il s'agit d'une meilleure approximation que W^{mf}, qui correspond à l'approximation à l'ordre zéro. La comparaison de ces deux approximations dans le cas du modèle de Potts simple binaire effectuée dans la partie suivante montre qu'en effet, le gain au niveau de la qualité de l'approximation en utilisant l'ordre 1 n'est pas négligeable.

L'approche variationnelle permet de distinguer l'approximation en champ moyen parmi l'ensemble des approximations d'un champ de Markov correspondant à un système de variables indépendantes (comme par exemple celui proposé par Geiger et Girosi 1991, décrit dans la partie 3.4) et de montrer son optimalité au sens de la divergence de Kullback-Leibler. Ce principe fournit une approximation de la distribution $P_G(\mathbf{z})$ et de l'espérance $\mathbb{E}_G[Z_i]$ et aussi comme on vient de le voir une approximation optimale de la constante de normalisation W parmi les bornes de la forme (3.25).

3.3 Approximations à l'ordre zéro et à l'ordre 1 de la fonction de partition

Il est établi que l'approximation en champ moyen W^{GBF} est la meilleure approximation de la fonction de partition de la distribution d'un champ de Markov parmi celles de la forme

(3.25). Le principe du champ moyen fournit une autre approximation de W. En effet, puisque l'approximation en champ moyen de $P_G(\mathbf{z})$ est

$$\frac{\exp(-H^{mf}(\mathbf{z}))}{W^{mf}},$$

il pourrait paraître naturel d'approximer W par W^{mf} plutôt que par W^{GBF}. Les approximations W^{mf} et W^{GBF} sont reliées par l'égalité suivante :

$$W^{\mathrm{GBF}} = W^{mf} \exp\{-\mathbb{E}^{mf}[H(\mathbf{Z} \mid \mathbf{y}) - H^{mf}(\mathbf{Z} \mid \mathbf{y})]\}.$$

Nous reprenons ici les travaux présentés dans Forbes et Peyrard (2001) sur la comparaison de ces deux approximations, sur l'exemple d'un modèle de Potts simple binaire, afin de déterminer l'effet du terme $\exp\{-\mathbb{E}^{mf}[H(\mathbf{Z} \mid \mathbf{y})-H^{mf}(\mathbf{Z} \mid \mathbf{y})]\}$ sur la qualité de l'approximation. Les expressions des approximations W^{mf} et W^{GBF} pour ce modèle sont détaillées dans la partie 3.3.1. Nous comparons ensuite l'évolution de la fonction de partition W en fonction du paramètre spatial du modèle, à celles de W^{GBF} et W^{mf}, dans le cas du modèle considéré en dimension 1 pour lequel le calcul de W est possible de manière exacte (partie 3.3.2). Dans le cas du modèle de Potts binaire en dimension 2, le calcul direct de W devient très rapidement impossible lorsque la taille de l'image augmente. Nous présentons donc des résultats pour une image de taille réduite (partie 3.3.3). L'étude de ces deux cas particuliers montre que l'approximation à l'ordre 1, W^{GBF}, est bien meilleure que l'approximation à l'ordre zéro, W^{mf}, mettant en évidence l'importance du terme correcteur $\exp\{-\mathbb{E}^{mf}[H(\mathbf{Z} \mid \mathbf{y}) - H^{mf}(\mathbf{Z} \mid \mathbf{y})]\}$ dans l'approximation de W.

3.3.1 Cas du modèle de Potts simple binaire

Les calculs présentés supposent que tous les pixels de l'image ont le même nombre de voisins. Pour cela, on identifie l'image à un tore, afin de compléter le voisinage des pixels sur le bord de l'image. On considère ainsi des conditions de bord périodiques. Dans le cas particuler où $\mathbf{z}$ est une chaîne de longueur n (dimension 1), cela revient à supposer un cas circulaire, en posant $\mathbf{z}_{n+1} = \mathbf{z}_1$. On rappelle que l'énergie d'un modèle de Potts simple à n sites est donnée par

$$H(\mathbf{z} \mid \beta) = -\beta \sum_{i \sim j} z_i^t z_j = -\frac{\beta}{2} \sum_{i=1}^{n} z_i^t \sum_{j \in N(i)} z_j.$$

On considère un modèle favorisant les cliques homogènes et on suppose donc que le paramètre spatial β est positif. L'énergie sous l'approximation en champ moyen s'écrit

$$H^{mf}(\mathbf{z} \mid \beta) = -\beta \sum_{i=1}^{n} z_i^t \sum_{j \in N(i)} z_j^{mf},$$

avec $z_j^{mf} = \mathbb{E}^{mf}[Z_j]$, l'approximation en champ moyen de $\mathbb{E}_{P_G}[Z_j]$. Il vient alors

$$\mathbb{E}^{mf}[H(\mathbf{Z} \mid \beta)] = -\frac{\beta}{2}\sum_{i=1}^{n} z_i^{mf\,t} \sum_{j\in N(i)} z_j^{mf} = \frac{1}{2}\mathbb{E}^{mf}[H^{mf}(\mathbf{Z} \mid \beta)],$$

de sorte que

$$W^{mf} \;=\; \sum_{\mathbf{Z}} \exp(-H^{mf}(\mathbf{z} \mid \beta)) \;=\; \sum_{i=1}^{n}\sum_{z_i} \exp(\beta z_i^t \sum_{j\in N(i)} z_j^{mf}), \tag{3.26}$$

et

$$W^{\mathrm{GBF}} \;=\; W^{mf}\exp(\mathbb{E}^{mf}[H(\mathbf{Z} \mid \beta)]) \;=\; W^{mf}\exp(-\frac{\beta}{2}\sum_{i=1}^{n} z_i^{mf\,t} \sum_{j\in N(i)} z_j^{mf}). \tag{3.27}$$

Puisqu'il n'y a pas de champ externe, par symétrie le champ moyen est identique en tout pixel et l'on peut écrire $z_i^{mf} = \mathbf{m}$ pour tout i dans S avec, dans le cas de 2 couleurs, $\mathbf{m} = (m_1,m_2)^t$ avec $m_1 + m_2 = 1$. Le champ moyen $\mathbf{m}$ doit vérifier les conditions de cohérence suivantes

$$\left\{ \begin{array}{l} m_1 = \frac{\exp(\beta N_v m_1)}{\exp(\beta N_v m_1)+\exp(\beta N_v m_2)} \\ \\ m_2 = \frac{\exp(\beta N_v m_2)}{\exp(\beta N_v m_1)+\exp(\beta N_v m_2)} \end{array} \right. ,$$

où $N_v = |N(i)|$ est le nombre de voisins d'un pixel pour le sytème de voisinage choisi. Il est équivalent de résoudre l'équation suivante

$$\begin{aligned} m_1 \;&=\; \frac{\exp(\beta N_v m_1)}{\exp(\beta N_v m_1) + \exp(\beta N_v(1-m_1))} \\ &=\; \frac{1}{1+\exp(\beta N_v(1-2m_1))}\,. \end{aligned} \tag{3.28}$$

Remarquons que si m_1 est solution de (3.28), alors $1 - m_1$ l'est aussi. Lorsque $\beta \le \frac{2}{N_v}$, il existe une unique solution, $m_1 = \frac{1}{2}$. Lorsque $\beta > \frac{2}{N_v}$, il existe deux solutions supplémentaires qui sont symétriques : m_1 et $1 - m_1$ avec $m_1 > \frac{1}{2}$ (voir par exemple Parisi 1988). Les deux solutions non triviales sont des fonctions de β pour lesquelles il n'existe pas d'expressions explicites. Cependant, le paramètre β peut être exprimé en fonction de m_1 à partir de l'égalité (3.28). On obtient ainsi

$$\beta \;=\; f(m_1) \;=\; \frac{1}{N_v(1-2m_1)}\log(\frac{1-m_1}{m_1}). \tag{3.29}$$

Remarquons que $f(1 - m_1) = f(m_1)$, de sorte que les deux solutions symétriques de (3.28) conduisent bien à la même valeur de β. De plus, lorsque m_1 tend vers 1, $f(m_1)$ tend vers

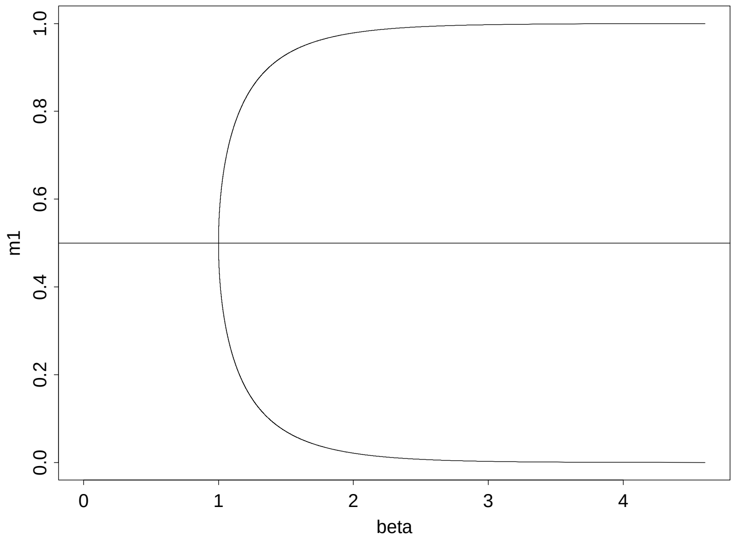


FIG. 3.1 – *solutions* $(m_1, 1-m_1)$ *des équations de cohérence du champ moyen pour le modèle de Potts binaire en dimension 1.*

l'infini, et lorsque m_1 tend vers la solution triviale $\frac{1}{2}$, $f(m_1)$ tend vers la valeur limite $\beta = \frac{2}{N_v}$. L'évolution des solutions de l'équation de point fixe (3.28) en fonction de la valeur de β est représentée sur la figure 3.1, dans le cas d'un modèle de Potts simple en dimension 1, pour lequel $N_v = 2$.

Les expressions (3.26) et (3.27) des approximations W^{mf} et W^{GBF} se réécrivent en fonction de β et m_1,

$$\begin{aligned} W^{mf} &= \left(\exp(\beta N_v m_1) + \exp(\beta N_v(1-m_1))\right)^n, \\ W^{\text{GBF}} &= W^{mf} \exp(-\frac{\beta N_v n}{2}(m_1^2 + (1-m_1)^2)), \end{aligned}$$

soit encore en utilisant (3.28) et en passant au logarithme

$$\begin{aligned} \log(W^{mf}) &= \beta N_v n m_1 + n \log(1 + \exp(\beta N_v(1-2m_1))) \qquad (3.30) \\ &= \beta N_v n m_1 - n \log(m_1), \\ \log(W^{\text{GBF}}) &= \frac{\beta N_v n}{2}(4m_1 - 2m_1^2 - 1) + n \log(1 + \exp(\beta N_v(1-2m_1))). \qquad (3.31) \end{aligned}$$

Remarquons que $\log(W^{mf})$ et $\log(W^{\text{GBF}})$ restent inchangés lorsque m_1 est remplacée par $1 - m_1$, de sorte que les deux solutions symétriques conduisent aux mêmes approximations

à l'ordre zéro et à l'ordre 1 de W.
Lorsque $\beta \leq \frac{2}{N_v}$, $m_1 = \frac{1}{2}$ et les expressions ci-dessus deviennent

$$\forall\, 0 \leq \beta \leq \frac{1}{2}, \qquad \begin{aligned} \log(W^{mf}) &= \frac{\beta N_v n}{2} + n\log(2), \\ \log(W^{\mathrm{GBF}}) &= \frac{\beta N_v n}{4} + \log(2). \end{aligned}$$

3.3.2 Comparaison des deux approximations en dimension 1

En général, le calcul de la fonction de partition W est exponentiel. Cependant, dans le cas du modèle de Potts simple binaire en dimension 1 on dispose d'une expression simple de W :

$$W = (\exp(\beta) + 1)^n + (\exp(\beta) - 1)^n ,$$

soit encore

$$\log(W) = n\beta + n\log(1 + \exp(-\beta)) + \log\left(1 + \left(\frac{1 - \exp(-\beta)}{1 + \exp(-\beta)}\right)^n\right).$$

La figure 3.2 montre l'évolution du logarithme de W et des deux approximations $\log W^{mf}$ et $\log W^{\mathrm{GBF}}$, en fonction de β, pour le modèle de Potts simple binaire en dimension 1 et pour $n = 100$ sites. Sur cet exemple, la courbe de l'évolution du logarithme de W^{mf} s'éloigne très rapidement de la courbe exacte et il est clair que la correction apportée par l'utilisation de l'approximation à l'ordre 1, W^{GBF}, améliore sensiblement l'approximation. Notons que lorsque $\beta \leq \frac{2}{N_v}$, on dispose directement des expressions des approximations $\log W^{mf}$ et $\log W^{\mathrm{GBF}}$ en fonction uniquement de β, à partir des équations (3.30) et (3.31) pour $m_1 = \frac{1}{2}$. Par contre, lorsque $\beta > \frac{2}{N_v}$, les courbes présentées ont été obtenues de la manière suivante : les approximations $\log W^{mf}$ et $\log W^{\mathrm{GBF}}$ ont été réécrites uniquement en fonction de m_1 grâce à la relation (3.29), puis nous nous sommes ramenés à une courbe en fonction de β en utilisant cette même relation.

3.3.3 Comparaison des deux approximations en dimension 2

En dimension 2, les expressions des approximations en champ moyen (3.30) et (3.31) sont toujours valables, par contre le calcul de W n'est pas possible directement à cause de l'explosion du nombre de configurations possibles avec le nombre de pixels. Afin de pouvoir comparer la qualité des approximations W^{mf} et W^{GBF}, nous nous sommes restreints à une grille de 3×3 pixels. Dans ce cas simple, la fonction de partition W est donnée pour un voisinage d'ordre 1 ($N_v = 4$) par

$$W = 102\exp(6\beta) + 144\exp(8\beta) + 198\exp(10\beta) + 48\exp(12\beta) + 18\exp(14\beta) + 2\exp(18\beta) ,$$

et pour un voisinage d'ordre 2 ($N_v = 8$) par

$$W = 252\exp(16\beta) + 168\exp(18\beta) + 72\exp(22\beta) + 18\exp(28\beta) + 2\exp(36\beta).$$

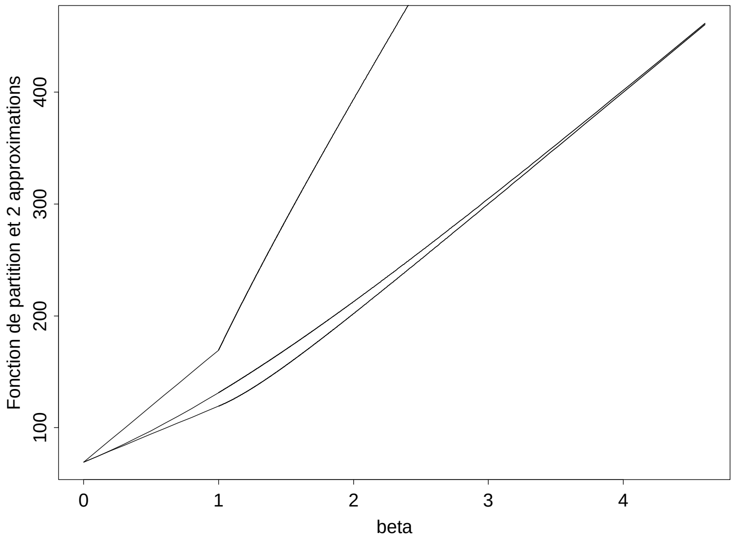


FIG. 3.2 – *modèle de Potts binaire en dimension 1 pour* $n = 100$ *sites : logarithme de la fonction de partition* $\log W$ *(courbe du milieu), logarithme de l'approximation à l'ordre zéro* $\log W^{mf}$ *(courbe du haut) et logarithme de l'approximation à l'ordre 1* $\log W^{\mathrm{GBF}}$ *(courbe du bas), en fonction du paramètre* β*.*

Les courbes du logarithme de la fonction de partition exacte et du logarithme des deux approximations à l'ordre zéro et à l'ordre 1, tracées en fonction de β, sont présentées sur la figure 3.3 pour les deux systèmes de voisinage. Ces courbes sont similaires à celles obtenues dans le cas de la dimension 1, et l'on retrouve sur cet exemple que l'approximation W^{GBF} est bien plus fine que l'approximation W^{mf}. Ceci nous fourni une information d'ordre qualitatif pour des images de taille plus réaliste: même si les valeurs numériques changent, il est raisonnable de penser que l'allure des courbes de la figure 3.3 reste iddentique lorsque le nombre de pixels augmente. Il est intéressant de noter que pour un nombre de pixels n quelconque fixé, il est possible de quantifier l'erreur induite $W - W^{\mathrm{GBF}}$ lorsque le paramètre β tend vers l'infini. En effet, W est alors équivalent au terme $K \exp(\frac{nN_v\beta}{2})$, qui est le terme dominant dans la somme sur toutes les configurations possibles du champ et correspond aux K configurations unicolores. On a donc

$$\log W \simeq \frac{nN_v\beta}{2} + \log K, \quad \text{lorsque } \beta \to \infty.$$

À partir des expressions (3.30) et (3.31), il vient

$$\log W^{mf} \simeq nN_v\beta, \quad \text{lorsque } \beta \to \infty,$$

et

$$\log W^{\mathrm{GBF}} \simeq \frac{nN_v\beta}{2}, \quad \text{lorsque } \beta \to \infty.$$

Pour des grandes valeurs de β, l'erreur commise avec l'approximation W^{GBF} est donc plus faible qu'avec W^{mf}. De plus elle est connue et égale à $\log K$, ce qui suggère une amélioration possible de l'approximation. Sur les exemples étudiés on observe que cette erreur asymptotique est déjà atteinte pour de petites valeurs du paramètre β (*cf.* figures 3.2 et 3.3).

3.4 Extensions et variantes

Plusieurs extensions du principe du champ moyen ont été étudiées (entre autre Kappen et Rodriguez 1998, Humphreys et Titterington 2000). Dans cette partie nous en présentons deux qui ont été appliquées dans le cadre de la segmentation markovienne d'images : la méthode *cavity field* et la méthode *cluster approximation* . Ces méthodes sont plus fines mais elles sont aussi plus complexes que l'approximation en champ moyen. Nous décrivons ensuite une variante du principe du champ moyen, utilisée dans le contexte de la reconnaissance d'images, qui permet elle aussi de se ramener au cas simple d'un système de variables indépendantes.

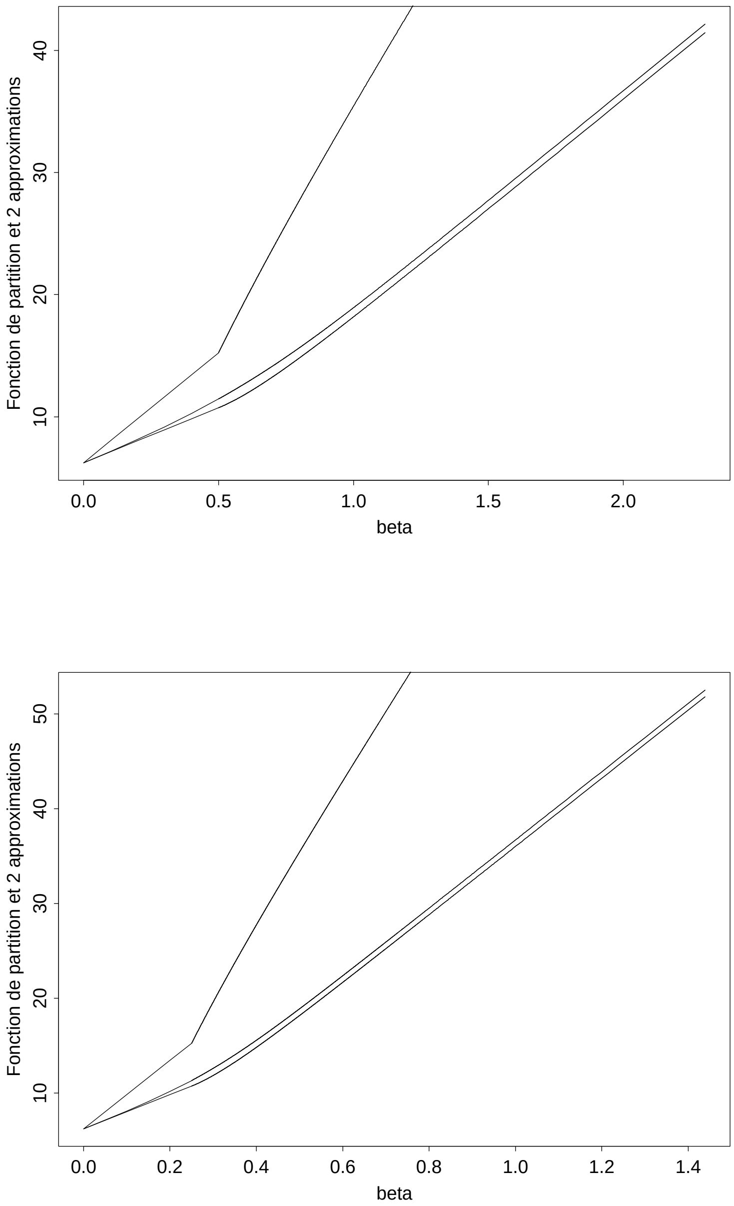


FIG. 3.3 – *modèle de Potts binaire sur une grille* 3×3 *et pour un voisinage d'ordre 1 (figure du haut) ou d'ordre 2 (figure du bas) : logarithme de la fonction de partition* $\log W$ *(courbe du milieu), logarithme de l'approximation à l'ordre zéro* $\log W^{mf}$ *(courbe du haut) et logarithme de l'approximation à l'ordre 1* $\log W^{\mathrm{GBF}}$ *(courbe du bas), en fonction du paramètre* β.

***Cavity field*:** l'espérance $m_i = \mathbb{E}_{P_G}[Z_i]$ d'un champ de Markov discret au site i est donnée par la formule (3.32) suivante

$$\begin{aligned} \forall i \in S, \quad m_i &= \sum_{z_i} z_i P_G(z_i) \\ &= \sum_{\mathbf{z}_{N(i)}} \sum_{z_i} z_i P_G(z_i \mid \mathbf{z}_{N(i)}) P_G(\mathbf{z}_{N(i)}) \\ &= \mathbb{E}_{P_G(\mathbf{z}_{N(i)})} \left[\frac{\exp \epsilon_i(\mathbf{Z}_{N(i)})}{\sum_{k=1}^K \exp \epsilon_{ik}(\mathbf{Z}_{N(i)})} \right], \end{aligned} \tag{3.32}$$

avec $\epsilon_i(\mathbf{Z}_{N(i)}) = \{\beta e_k^t \sum_{j \in N(i)} Z_j,\, 1 \leq k \leq K\}$ par exemple pour un modèle de Potts simple. Le principe du champ moyen revient à faire l'approximation suivante

$$\forall i \in S, \quad m_i \approx \frac{\exp \mathbb{E}_{P_G(\mathbf{z}_{N(i)})}[\epsilon_i(\mathbf{Z}_{N(i)})]}{\sum_{k=1}^K \exp \mathbb{E}_{P_G(\mathbf{z}_{N(i)})}[\epsilon_{ik}(\mathbf{Z}_{N(i)})]}. \tag{3.33}$$

Cette approximation est le développement à l'ordre 1 de l'espérance (3.32) autour de la valeur moyenne $\mathbb{E}_{P_G(\mathbf{z}_{N(i)})}[\epsilon_i(\mathbf{Z}_{N(i)})]$. La méthode dite *cavity field* (voir par exemple Dunmur et Titterington 1997) consiste à prendre aussi en compte les termes d'ordre 2 dans le développement afin de raffiner l'approximation. Cette méthode est aussi connu sous le nom d'approche TAP (*cf.* Thouless, Anderson, et Palmer 1977 et Opper et Winther 2000). Dunmur et Titterington (1998) ont appliqué la méthode *cavity field* pour la mise en œuvre de l'algorithme EM pour champ de Markov caché et ont comparé la procédure correspondante avec celle basée sur l'utilisation de l'approximation en champ moyen. Leurs expériences montrent des résultats très peu différents pour les deux méthodes, de sorte qu'en pratique l'approximation à l'ordre 2 *cavity field*, plus complexe à calculer, ne paraît pas apporter une amélioration. Notons que la méthode *cavity field* est très proche de l'approximation de Bethe (Bethe 1935) utilisée dans le contexte de l'inférence pour les modèles graphiques.

***Cluster approximation*:** il est possible d'imaginer une généralisation du principe du champ moyen en ne fixant pas tous les pixels autres que le pixel cible à leur valeur moyenne, mais en traitant exactement ceux qui se trouvent proches du pixel cible. C'est ce que Wu et Doerschuk (1995) proposent dans le cas de champs de Markov discrets et nomment *cluster approximation*. Soit i un pixel donné et $G_i \in S$ un ensemble de pixels contenant i. Pour calculer de manière approchée la moyenne de la variable Z_i, les interactions locales entre les pixels de G_i sont traitées exactement alors que les interactions "de longue distance" entre les pixels de G_i et les pixels de $S \backslash G_i$ sont approximées. Le principe d'approximation est le même que pour le champ moyen : les fluctuations des variables Z_j pour j dans $S \backslash G_i$ sont négligées et les variables sont fixées à leur valeur moyenne. Ces valeurs moyennes étant justement inconnues, l'approximation dépend d'une condition de cohérence qui se traduit là aussi par un problème de point fixe. L'approximation en champ moyen correspond au cas particulier $G_i = \{i\}$, et on retrouve alors l'équation (3.9). Dans le cas général, la solution du problème de point fixe fournit une approximation de l'espérance du champ de Markov $\mathbf{Z}$ et Wu et Doerschuk (1995) montrent que lorsqu'une telle solution existe elle appartient à l'ensemble $\mathcal{V}$. Les conditions nécessaires établies par les auteurs pour garantir l'existence et l'unicité de cette solution sont les mêmes que celles déjà décrites dans la partie 3.1.2 pour le cas particulier du champ moyen. Si $G_i = N(i)$, toutes les interactions du premier ordre sont prises en compte exactement. Ce cas ne conduit plus à une approximation de la distribution de Gibbs qui corresponde à un système indépendant. Pour que la méthode de Wu et Doerschuk (1995) reste intéressante il faut donc que l'ensemble G_i soit de taille raisonnable. En outre, dans leurs expériences les auteurs ont testé la méthode pour approximer l'espérance d'un champ de Markov discret conditionnellement aux observations. Les résultats obtenus pour $G_i = \{i\}$ (*i.e.* approximation en champ moyen) sont analogues à ceux obtenus dans le cas où G_i est étendu aux quatre plus proches voisins et dans le cas d'une approximation par simulations MC.

Approximation par un système de variables indépendantes : dans le contexte de la reconstruction de surface, Geiger et Girosi (1991) considèrent, comme pour le principe du champ moyen, une méthode d'approximation d'un champ de Markov continu par approximation de toutes les interactions. Le principe consiste, chaque fois qu'une interaction est concernée, à négliger les fluctuations. Ainsi, dans le cas d'un modèle de champ de Markov avec interaction au plus d'ordre 2, de fonction énergie

$$H(\mathbf{z}) \;=\; -\sum_{i \in S} \{V_i(z_i) + \frac{\beta}{2} \sum_{j \in N(i)} V_{ij}(z_i,z_j)\},$$

le potentiel $V_{ij}(z_i,z_j)$ est remplacé par $1/2[V_{ij}(z_i,m_j)+V_{ij}(m_i,z_j)]$, avec $m_i = \mathbb{E}_{P_G}[Z_i], \forall i \in S$, ce qui amène à définir l'énergie approximée $\tilde{H}(\mathbf{z}) = \sum_{i \in S} \tilde{H}_i(z_i)$ avec pour énergie locale

$$\tilde{H}_i(\mathbf{z}) \;=\; -\{V_i(z_i) + \frac{\beta}{2} \sum_{j \in N(i)} V_{ij}(z_i,m_j)\}\;. \tag{3.34}$$

Comme pour le principe du champ moyen, cette méthode conduit à une approximation du champ de Markov par un système de variables indépendantes, puisque l'on obtient comme approximation de la distribution de Gibbs $\tilde{P}(\mathbf{z}) \propto \exp\{-\tilde{H}(\mathbf{z})\}$ qui se factorise. Les deux méthodes se distinguent par l'approximation des potentiels d'ordre 2, le principe du champ moyen faisant l'hypothèse $V_{ij}(z_i,z_j) \approx V_{ij}(z_i,m_j) + V_{ij}(m_i,z_j)$. Ainsi, avec la méthode de Geiger et Girosi (1991), on a bien $\tilde{H}(\mathbf{z})|_{\mathbf{m}=\mathbf{z}} = H(\mathbf{z})$, cependant la distribution marginale correspondante $\tilde{P}_i(z_i)$ diffère de $P_G(z_i \mid \mathbf{m}_{N(i)})$ alors que l'approximation en champ moyen conduit à $P_i^*(z_i) = P_G(z_i \mid \mathbf{m}_{N(i)})$ (équation (3.6)). L'approche variationnelle (partie 3.2) permet d'établir qu'en théorie l'approximation $\tilde{P}(\mathbf{z})$ d'une distribution de Gibbs est moins fine que l'approximation en champ moyen, au sens de la divergence de Kullback-Leibler.

Zhang (1992) a utilisé les deux méthodes d'approximation pour la mise en œuvre de l'algorithme EM pour champ de Markov caché et reporte que l'approximation en champ moyen conduit à de meilleures estimations des paramètres du modèle.

Chapitre 4

Champ moyen et algorithme EM

Nous nous intéressons dans ce chapitre à l'outil *champ moyen* pour l'estimation des paramètres d'un modèle de champ de Markov caché. L'approximation est attrayante car elle permet de se ramener à un système indépendant. L'idée d'utiliser le principe du champ moyen pour la mise en œuvre de l'algorithme EM pour modèles de champs de Markov cachés est due à Zhang (1992). Les résultats encourageants obtenus par cette technique nous amène à proposer une généralisation du principe d'approximation en champ moyen d'une distribution de Gibbs. Le principe consiste à se ramener à une distribution factorisable en fixant le voisinage de chaque pixel à des constantes à définir. Nous proposons une famille d'algorithmes de type EM qui tirent parti de la simplicité du modèle résultant de l'approximation. Ces procédures s'interprètent comme un algorithme EM appliqué à un modèle de mélange fini indépendant adapté à chaque itération. Les principaux résultats obtenus sont reportés dans Celeux, Forbes, et Peyrard (2001) et nous présentons ici la méthodologie développée de manière plus précise ainsi que des expérimentations supplémentaires.
La partie 4.1 présente l'algorithme de Zhang (1992) ainsi que d'autres procédures proches liées à l'approximation en champ moyen d'un modèle de champ de Markov caché. La généralisation de l'approximation en champ moyen d'une distribution de Gibbs est décrite dans la partie 4.2 et la famille d'algorithmes EM *de type champ moyen* correspondante est définie et discutée dans les parties 4.3 à 4.8. Les performances des algorithmes proposés sont illustrées sur différents exemples d'images simulées et réelles dans la partie 4.9.

4.1 Algorithme de Zhang

Dans le cas de l'estimation des paramètres d'une chaîne de Markov cachée, Zhang (1992) décrit l'algorithme EM avec approximation en champ moyen des probabilités conditionnelles mises en jeu. L'algorithme correspondant est simple et peu coûteux. Sur les expériences réalisées par Zhang (1992) il présente des performances comparables à celles de l'algorithme de Baum (Baum, Soules, et Weiss 1970), algorithme exact d'estimation. Dans le cas de EM pour champ de Markov caché discret, Zhang (1992) propose une approximation *de type champ moyen* des distributions marginales et conditionnelles du champ gibbsien. À l'étape

E, dans l'espérance $Q(\Psi \mid \Psi^{(q)})$, les probabilités conditionnelles $P_G(z_i \mid \mathbf{y},\Psi^{(q-1)})$ sont ainsi remplacées par

$$P_G(z_i \mid \mathbf{y},\Psi^{(q-1)}) \;\approx\; P_G(z_i \mid \mathbf{z}_{N(i)}^{mf\,(q)},\mathbf{y},\Psi^{(q-1)}),$$

où $\mathbf{z}^{mf\,(q)}$ est l'approximation en champ moyen de l'espérance du champ conditionnel $P_G(\mathbf{z} \mid \mathbf{y},\Psi^{(q-1)})$, solution de

$$\forall i \in S, \quad z_i^{mf\,(q)} = \frac{\sum\limits_{z_i} z_i \exp\{-\sum\limits_{c\ni i} V_c(z_i,\mathbf{z}_{c\backslash\{i\}}^{mf\,(q)} \mid \beta^{(q-1)}) + \log g(y_i \mid z_i,\, \theta^{(q-1)})\}}{\sum\limits_{z_i} \exp\{-\sum\limits_{c\ni i} V_c(z_i,\mathbf{z}_{c\backslash\{i\}}^{mf\,(q)} \mid \beta^{(q-1)}) + \log g(y_i \mid z_i,\, \theta^{(q-1)})\}}\,. \tag{4.1}$$

Le terme dépendant uniquement du paramètre θ dans Q devient

$$Q_\theta(\theta \mid \Psi^{(q-1)}) \;\approx\; \sum_{i\in S}\sum_{z_i} P_G(z_i \mid \mathbf{z}_{N(i)}^{mf\,(q)},\mathbf{y},\Psi^{(q-1)}) \log g(y_i \mid z_i,\theta),$$

soit encore d'après (3.17),

$$Q_\theta(\theta \mid \Psi^{(q-1)}) \;\approx\; \sum_{i\in S}\sum_{k=1}^{K} z_{ik}^{mf\,(q)} \log g(y_i \mid e_k,\theta)\,. \tag{4.2}$$

Tous les termes intervenant dans Q_θ et Q_β et indépendants de θ et β sont maintenant calculables directement. En plus de cette approximation des probabilités conditionnelles, Zhang (1992) propose de remplacer à l'étape (M) la distribution marginale $P_G(\mathbf{z} \mid \beta)$ par l'approximation *de type champ moyen* suivante

$$\prod_{i\in S} P_G(z_i \mid \mathbf{z}_{N(i)}^{mf\,(q)},\beta). \tag{4.3}$$

La quantité (4.2) reste inchangée et le terme dépendant uniquement de β devient

$$\begin{aligned} Q_\beta(\beta \mid \Psi^{(q-1)}) \;&\approx\; \sum_{i\in S}\sum_{z_i} P_G(z_i \mid \mathbf{y},\Psi^{(q-1)})\; \log P_G(z_i \mid \mathbf{z}_{N(i)}^{mf\,(q)},\beta) \\ &\approx\; \sum_{i\in S}\sum_{k=1}^{K} z_{ik}^{mf\,(q)} \log P_G(e_k \mid \mathbf{z}_{N(i)}^{mf\,(q)},\beta). \end{aligned} \tag{4.4}$$

Cette méthode permet d'éviter le calcul de la constante de normalisation $W(\beta)$ à l'étape (M). Soit $\hat{\Psi}$ l'estimateur final fourni par cette procédure, Zhang (1992) obtient une segmentation $\hat{\mathbf{z}}$ d'après la règle suivante

$$\forall i \in S, \quad \hat{z}_i \;=\; e_k \text{ ssi } k = \arg\max_{1\le l\le K} P_G(e_l \mid \mathbf{z}_{N(i)}^{mf\,(q)},\mathbf{y},\hat{\Psi}).$$

Remarquons que l'approximation utilisée dans l'étape (E) nécessite de résoudre le problème de point fixe associé au champ conditionnel $P_G(\mathbf{z} \mid \mathbf{y},\Psi^{(q-1)})$. Pour cette distribution, la fonction énergie possède des potentiels sur les singletons définis par $V_i(z_i) = -\log g(y_i \mid z_i,\theta)$, de sorte que le problème de l'existence d'une solution uniforme $\mathbf{u}^K$ ne se pose pas. La distribution (4.3) est bien une distribution factorisable, mais elle ne correspond pas à l'approximation en champ moyen de $P_G(\mathbf{z} \mid \beta)$, puisque $\mathbf{z}^{mf\,(q)}$ est calculé à partir du champ conditionnel pour $\beta = \beta^{(q-1)}$. Le calcul de l'approximation en champ moyen de $P_G(\mathbf{z} \mid \beta)$ est impossible car cela demanderait de connaître la valeur du paramètre β. L'approximation (4.3) n'est donc pas optimale au sens de la divergence de Kullback-Leibler.
Par la suite, nous ferons référence à l'algorithme qui vient d'être décrit sous le nom d'algorithme de Zhang. Dans le cas d'images tests simulées, l'approche en champ moyen pour EM montre de bons résultats pour l'estimation des paramètres de la densité des observations, par contre les performances sont moins bonnes en ce qui concerne l'estimation des paramètres du champ de Markov. Sur ces images, on observe également que l'allure des segmentations est très semblable selon que le paramètre spatial β est estimé ou fixé. Cela conduit Zhang (1992) à la conclusion que le choix d'un paramètre β dans un intervalle de valeurs raisonnable est suffisant, une valeur précise n'étant pas vraiment importante pour la segmentation finale. En termes de segmentation d'images réelles, peu d'expériences sont publiées illustrant les performances de l'algorithme de Zhang. Dans Zhang (1992), pour traiter de telles images les paramètres spatiaux sont supposés connus et ne sont pas estimés. Dans Dunmur et Titterington (1998), l'approximation en champ moyen pour EM est comparée à l'algorithme ICM non supervisé sur des images simulées et réelles. Les résultats montrent que les estimateurs obtenus par champ moyen sont plus satisfaisants et peuvent conduire à de meilleures segmentations. Cette étude comparative est la seule disponible à notre connaissance, de sorte que l'on peut simplement parler de résultats encourageants pour le champ moyen.

Zhang a également étudié l'approximation en champ moyen pour l'estimation des paramètres d'un champ de Markov caché continu. Nous décrivons ci-dessous la version de l'algorithme EM correspondante. Le dernier paragraphe détaille les liens entre l'algorithme de Zhang et un algorithme d'estimation d'un modèle de mélange fini indépendant, l'algorithme NEM présenté dans la partie 2.2.3.

Cas continu et approximation du point selle : dans des travaux ultérieurs, Zhang (1993) étend l'utilisation de l'approximation en champ moyen à l'estimation des paramètres d'un champ de Markov continu associé à un processus de lignes dans le contexte de la restauration d'image. Le principe est le même que pour un champ de Markov discret. Cependant, en pratique, la méthode n'est plus efficace dans le cas continu car la résolution du problème de point fixe associé au champ moyen implique d'intégrer sur l'ensemble de définition des variables Z_i. Zhang (1993) propose alors d'utiliser l'approximation dite *du point selle* (*cf.* Geiger et Girosi 1991). Il s'agit, comme pour le champ moyen, d'approximer l'influence des pixels $j \in N(i)$ dans le calcul de $\mathbb{E}_{P_G}[Z_i]$ par l'influence d'un champ constant. L'approximation du point selle utilise le champ z^{sp} solution de

$$\forall\, i \in S, \quad \frac{\partial H(\{z_i \cup \mathbf{z}^{sp}_{S\setminus\{i\}}\})}{\partial z_i}\,|_{z_i = z_i^{sp}} \;=\; 0. \tag{4.5}$$

Cette approche peut être utilisée pour obtenir une approximation $P^{sp}(\mathbf{z}) = W^{sp^{-1}} \exp(-H^{sp}(\mathbf{z}))$ de $P_G(\mathbf{z})$ qui est factorisable, avec

$$\begin{aligned} H^{sp}(\mathbf{z}) &= \sum_{i\in S} H(\{z_i \cup \mathbf{z}^{sp}_{S\setminus\{i\}}\}), \\ W^{sp} &= \prod_{i\in S} \int_{z_i} H(\{z_i \cup \mathbf{z}^{sp}_{S\setminus\{i\}}\}) dz_i. \end{aligned}$$

Au cours de l'algorithme EM, les probabilités conditionnelles sont approximées à l'étape E par

$$P^{sp}(z_i \mid \mathbf{y}, \Psi^{(q-1)}) = P_G(z_i \mid \mathbf{Z}_{N(i)} = \mathbf{z}^{sp}_{N(i)}, \mathbf{y}, \Psi^{(q-1)}),$$

où $\mathbf{z}^{sp}$ est l'approximation du point selle du champ conditionnel. Zhang (1993) se plaçant dans le cas où les paramètres du champ de Markov sont connus, il n'est pas nécessaire d'approximer la distribution marginale. La dernière configuration $\mathbf{z}^{sp}$ calculée au cours de l'algorithme fournit une restauration de l'image.
Ces travaux sont à rapprocher de ceux de Leahy, Qi, et Zhou (1997) sur l'estimation des paramètres de la distribution de Gibbs marginale dans le cas d'un champ caché continu. Ils proposent une méthode pour estimer le paramètre spatial β par l'algorithme EM dans le cas où le paramètre du bruit θ est connu. La procédure correspondante est équivalente à celle proposée par Zhang (1992) dans le cas discret et pour θ connu, en remplaçant l'approximation en champ moyen $\mathbf{z}^{mf\,(q)}$ de l'espérance conditionnelle par l'approximation du point selle $\mathbf{z}^{sp}$ dans la relation (4.2) et la formule (4.3).

Lien entre NEM et l'algorithme de Zhang : Dang et Govaert (1998) ont montré que, dans le cas où les proportions π_k du modèle de mélange fini indépendant sont identiques, l'algorithme NEM (*cf.* partie 2.2.3) est équivalent à l'algorithme de Zhang pour un modèle de Potts simple de paramètre spatial β connu. Dans le cas plus général où les proportions du modèle de mélange sont fixes mais quelconques, l'algorithme NEM est équivalent à l'algorithme de Zhang pour un modèle de Potts avec champ externe (*cf.* exemple 3.2) dont le paramètre spatial est connu, ainsi que les potentiels sur les singletons qui doivent vérifier $V_i(e_k) = \log \pi_k$ pour tout $i \in S$, $k \in \{1, \ldots, K\}$. En effet, dans ce cas le problème de point fixe (4.1) qui doit être résolu au cours de l'itération q de l'algorithme de Zhang est donné par

$$\forall i \in S, \forall k \in \{1, \ldots, K\},$$
$$z_{ik}^{mf\,(q)} = \frac{\exp\{\log \pi_k + \beta e_k^t \sum_{j\in N(i)} z_j^{mf\,(q)} + \log g(y_i \mid Z_i = e_k,\, \theta^{(q-1)})\}}{\sum\limits_{l=1}^{K} \exp\{\log \pi_l + \beta e_l^t \sum_{j\in N(i)} z_j^{mf\,(q)} + \log g(y_i \mid Z_i = e_l,\, \theta^{(q-1)})\}}.$$

On retrouve bien l'égalité (2.29), et on a donc $\mathbf{c}^{(q)} = \mathbf{z}^{mf\,(q)}$ ou plus précisement, s'il n'y a pas unicité, les solutions de l'étape de mise à jour de $\mathbf{c}^{(q)}$ dans NEM sont celles du problème de point fixe ci-dessus. Ensuite, puisque les paramètres β et $\mathbf{V}$ du champ de Markov sont supposés connus, l'étape (M) de l'algorithme de Zhang consiste à mettre à jour le paramètre θ d'après (*cf.* équation (4.2))

$$\theta^{(q)} = \arg\max_{\theta} \sum_{i \in S} \sum_{k=1}^{K} z_{ik}^{mf\,(q)} \log g(y_i \mid e_k, \theta).$$

L'étape (M) de NEM est donnée par

$$\begin{aligned}\theta^{(q)} &= \arg\max_{\theta} \sum_{i \in S} \sum_{k=1}^{K} c_{ik}^{(q)} \log \pi_k g(y_i \mid e_k, \theta), \\ &= \arg\max_{\theta} \sum_{i \in S} \sum_{k=1}^{K} c_{ik}^{(q)} \log g(y_i \mid e_k, \theta).\end{aligned}$$

La formule de mise à jour du paramètre θ est donc bien la même dans les deux procédures.

Il n'est pas intéressant d'estimer le paramètre spatial β au cours de l'algorithme NEM par maximisation du critère $U_\beta(\mu, \mathbf{c})$. En effet, ce critère est linéaire en β et n'est donc pas borné selon ce paramètre (*cf.* équation 2.28). Dang (1998) propose une version de l'algorithme NEM où le paramètre β est mis à jour à chaque itération, dans le cas où les proportions π_k sont identiques. Les paramètres à estimer sont alors β et θ. En s'inspirant de l'équivalence qui vient d'être décrite, il propose de maximer le critère $F_{potts}(\mathbf{c}, \theta, \beta) = U_\beta(\theta, \mathbf{c}) - \log W(\beta)$, où $W(\beta)$ est la fonction de partition du modèle de Potts simple. Ce nouveau critère fait apparaître le terme $\beta \sum_{i \sim j} c_i^t c_j - \log W(\beta) = \log P_G(\mathbf{c} \mid \beta)$ qui peut s'interpréter comme une extension à $\mathcal{V}^n$ du logarithme de la distribution $P_G(\mathbf{z} \mid \beta)$ du modèle de Potts simple. La maximisation de F_{potts} en β n'est donc pas possible directement et Dang (1998) propose de remplacer $P_G(\mathbf{c} \mid \beta)$ par l'approximation de la pseudo-vraisemblance, de sorte que F_{potts} devient

$$\begin{aligned}F_{potts}(\mathbf{c}, \theta, \beta) \approx & \sum_{i \in S} \sum_{k=1}^{K} c_{ik} \log g_G(y_i \mid Z_i = e_k) - \sum_{i \in S} \sum_{k=1}^{K} c_{ik} \log c_{ik} \\ & + \beta \sum_{i \in S} c_i^t \sum_{j \in N(i)} c_j - \sum_{i \in S} \log \sum_{k=1}^{K} e^{\beta \sum_{j \in N(i)} c_{jk}}.\end{aligned}$$

Soient $\mathbf{c}^{(q-1)}$, $\theta^{(q-1)}$ et $\beta^{(q-1)}$ les valeurs courantes des paramètres, une étape de mise à jour des paramètres par maximisation alternée de $F_{potts}(\mathbf{c}, \theta, \beta)$ se déroule de la manière suivante :

(1) $\mathbf{c}^{(q)} = \arg\max_{\mathbf{c}} U_{\beta^{(q-1)}}(\theta^{(q-1)}, \mathbf{c})$;

(2) $\theta^{(q)} = \arg\max_\theta U_{\beta^{(q-1)}}(\theta,\mathbf{c}^{(q)})$;

(3)

$$\begin{aligned}
\beta^{(q)} &= \arg\max_\beta \beta \sum_{i\in S} c_i^{(q)t} \sum_{j\in N(i)} c_j^{(q)} - \sum_{i\in S} \log \sum_{k=1}^K e^{\beta \sum_{j\in N(i)} c_{jk}^{(q)}} \\
&= \arg\max_\beta \beta \sum_{i\in S} \sum_{k=1}^K c_{ik}^{(q)} [\sum_{j\in N(i)} c_{jk}^{(q)} - \log \sum_{k=1}^K e^{\beta \sum_{j\in N(i)} c_{jk}^{(q)}}] \\
&= \arg\max_\beta \sum_{i\in S} \sum_{k=1}^K c_{ik}^{(q)} \log P_G(Z_i = e_k \mid \mathbf{Z}_{N(i)} = \mathbf{c}_{N(i)}^{(q)},\beta). \qquad (4.6)
\end{aligned}$$

On retrouve l'approximation de la distribution markovienne marginale utilisée dans l'algorithme de Zhang dans le calcul de Q_β à l'étape (M) (équation (4.4)). Les deux procédures sont donc équivalentes lorsque le modèle de champ de Markov utilisé est le modèle de Potts simple puisque, comme il a été détaillé précédemment, l'étape (1) de mise à jour de θ par maximisation de $U_{\beta^{(q-1)}}(\theta^{(q-1)},\mathbf{c})$ est équivalente au calcul de l'approximation en champ moyen $\mathbf{z}^{mf(q)}$ de la distribution markovienne conditionnelle. La proposition de Dang (1998) pour la mise à jour de β au cours de NEM s'avère donc fondée puisque nous venons de montrer qu'elle permet d'étendre l'équivalence entre NEM et l'algorithme de Zhang au cas où le paramètre spatial est inconnu et les proportions du mélanges sont supposées égales.

4.2 Généralisation du principe du champ moyen

La méthode d'approximation en champ moyen repose sur le fait qu'en fixant les voisins du pixel cible à des constantes (les valeurs moyennes), les interactions sont supprimées. Cela conduit à un modèle plus simple que la distribution de Gibbs. On peut imaginer de généraliser le principe en définissant une famille de distributions sur S obtenues en fixant les voisins à des constantes quelconques. Ainsi, pour un champ voisin $\tilde{\mathbf{z}}$ arbitraire, nous considérons l'approximation suivante de $P_G(\mathbf{z})$,

$$P_{\tilde{\mathbf{z}}}(\mathbf{z}) = \prod_{i\in S} P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}). \qquad (4.7)$$

Toute distribution de la forme (4.7) est la distribution d'un système de variables indépendantes. Le problème est alors de choisir une stratégie pour définir un champ voisin $\tilde{\mathbf{z}}$. L'approximation en champ moyen choisit $\tilde{\mathbf{z}}$ en négligeant les fluctuations autour de sa moyenne de l'environnement de chaque pixel. Cela correspond à la distribution (4.7) avec $P_{\tilde{\mathbf{z}}}(\mathbf{z}) = P^{mf}(\mathbf{z})$ et $\tilde{\mathbf{z}} = \{\mathbb{E}^{mf}[Z_i], i \in S\}$. Si $\tilde{\mathbf{z}}$ est remplacé par $\mathbf{z}$ dans (4.7), on retrouve la pseudo-vraisemblance définie à la partie 2.1.2.2. Cependant dans la définition (4.7) le champ $\tilde{\mathbf{z}}$ est constant pour tout $\mathbf{z}$, de sorte que l'approximation de la pseudo-vraisemblance est d'une nature différente et n'est pas incluse dans la classe d'approximations proposées. Il ne s'agit d'ailleurs pas d'une distribution de probabilité. Nous allons présenter deux distributions particulières de la forme (4.7), correspondant à deux choix naturels pour $\tilde{\mathbf{z}}$: l'approximation *en champ modal* et l'approximation *en champ simulé*. La première s'obtient de la manière suivante. Au lieu de fixer les voisins du pixel cible à leur valeur moyenne, nous proposons de les fixer au mode (ou à l'un des modes s'il en existe plusieurs) de leur distribution. Ces modes, comme l'espérance d'un champ de Markov, sont inconnus. Cependant, de la même façon que pour le champ moyen, il est possible de définir l'approximation en champ modal par une condition de cohérence traduisant que le mode de la distribution approximée $P_{\tilde{\mathbf{Z}}}$ doit être égal au mode utilisé pour construire l'approximation. La distribution $P_G(\mathbf{z})$ est ainsi approximée par

$$P^{mode}(\mathbf{z}) = P_{\mathbf{Z}^{mode}}(\mathbf{z}) = \prod_{i \in S} P_G(z_i \mid \mathbf{z}_{N(i)}^{mode}),$$

où $\mathbf{z}_i^{mode}$ est un mode de la distribution $P^{mode}(z_i) = P_G(z_i \mid \mathbf{z}_{N(i)}^{mode})$, vérifiant l'équation de cohérence

$$z_i^{mode} \;=\; \arg\max_{z_i} P_G(z_i \mid \mathbf{z}_{N(i)}^{mode}).$$

Cette équation répétée pour tout i dans S conduit à un problème de point fixe qui peut être résolu itérativement

$$\forall i \in S, \quad z_i^{mode^{(q)}} \;=\; \arg\max_{z_i} P_G(z_i \mid \mathbf{z}_{N(i)}^{mode^{(q-1)}}). \tag{4.8}$$

Exemple 4.1 (Modèle de Potts simple à K couleurs (suite)) *Pour le modèle (3.10) de l'exemple 3.1, la condition de cohérence définissant l'approximation en champ modal devient*

$$\forall i \in S, \quad z_i^{mode} \;=\; \arg\max_{z_i} \exp(\beta z_i^t \sum_{j \in N(i)} z_j^{mode}).$$

La résolution itérative de cette équation conduit à choisir pour z_i^{mode} la couleur courante la plus souvent présente dans le voisinage $\mathbf{z}_{N(i)}^{mode}$.

Une autre méthode pour choisir le voisinage consiste à générer une réalisation de la distribution de Gibbs $P_G(\mathbf{z})$ puis à fixer $\tilde{\mathbf{z}}$ à cette valeur. Cette approximation ne dépend pas d'une condition de cohérence comme l'approximation en champ moyen ou en champ modal. Nous l'avons nommée *approximation en champ simulé*. Soit $\mathbf{z}^{sf}$ une réalisation de $P_G(\mathbf{z})$ obtenue par l'échantillonneur de Gibbs (Geman et Geman 1984 et partie 1.4), l'approximation en champ simulé correspondante de $P_G(\mathbf{z})$ est donc

$$P^{sf}(\mathbf{z}) = P_{\mathbf{Z}^{sf}}(\mathbf{z}) = \prod_{i \in S} P_G(z_i \mid \mathbf{z}_{N(i)}^{sf}).$$

Cette deuxième méthode d'approximation trouve son intérêt dans l'optique d'une utilisation pour l'estimation des paramètres d'un champ de Markov caché par EM. En effet, l'étape stochastique de mise à jour du voisinage laisse espérer la convergence de l'algorithme, ce que

nous démontrons effectivement dans la partie 4.5.
Dans la suite de ce mémoire nous parlerons de manière générale d'*approximation de type champ moyen* pour désigner une approximation d'une distribution de Gibbs de la forme (4.7), pour laquelle la valeur du champ au pixel i ne dépend plus de la valeur sur les autres pixels, qui sont fixés à des constantes indépendamment de la valeur au pixel i. L'utilisation de cette méthode pour résoudre les étapes (E) et (M) de l'algorithe EM conduit aux algorithmes EM *de type champ moyen* définis dans la partie suivante.

4.3 Algorithmes EM de type champ moyen

Lorsque les étiquettes d'un modèle de mélange ont pour loi une distribution factorisable, la mise en œuvre de l'algorithme EM est simple (*cf.* partie 2.2.2). Nous proposons donc de remplacer le modèle de champ de Markov caché défini par la distribution de Gibbs $P_G(\mathbf{z} \mid \beta)$ et la densité conditionnelle des observations $g(\mathbf{y} \mid \mathbf{z},\theta)$ par le modèle simplifié défini par $P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta)$ et $g(\mathbf{y} \mid \mathbf{z},\theta)$. Le champ voisin $\tilde{\mathbf{z}}$ dépendant dans le cas le plus général des paramètres (θ,β), celui-ci doit également être mis à jour au cours de la procédure d'estimation. Le principe des algorithmes que nous proposons consiste à alterner une étape de restauration du voisinage (*neighborhood restoration* (NR)), puis une étape (EM) d'estimation des paramètres du modèle de champ de Markov caché via EM pour le modèle de mélange fini indépendant correspondant au modèle simplifié courant. La forme générale d'une itération de ces procédures, que nous nommons algorithmes EM *de type champ moyen*, est donc la suivante,

(NR) **choix des voisins :** créer, à partir des observations $\mathbf{y}$ et de l'estimation courante du paramètre $\Psi^{(q-1)}$, un champ $\tilde{\mathbf{z}}^{(q)}$. Pour chaque pixel i, fixer les voisins à $\tilde{\mathbf{z}}^{(q)}_{N(i)}$ et remplacer la distribution marginale $P_G(\mathbf{z} \mid \beta)$ par

$$P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta) \quad = \quad \prod_{i \in S} P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)},\beta). \tag{4.9}$$

(EM) **estimation :** appliquer l'algorithme EM pour le modèle de mélange fini indépendant défini par (4.9) et $g(\mathbf{y} \mid \mathbf{z},\theta)$, avec pour valeurs initiales $\theta^{(q-1)}$ et $\beta^{(q-1)}$, pour obtenir la mise à jour des estimateurs $\theta^{(q)}$ et $\beta^{(q)}$.

La distribution des données complètes $P_G(\mathbf{y},\mathbf{z} \mid \Psi)$ est ainsi remplacée par

$$P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z},\mathbf{y} \mid \Psi) \quad = \quad \prod_{i \in S} \{g(y_i \mid z_i,\theta)\ P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)},\beta)\}, \tag{4.10}$$

qui correspond à une distribution des observations $\mathbf{Y}$ de la forme

$$g_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{y} \mid \Psi) \quad = \quad \sum_{\mathbf{z}} g(\mathbf{y} \mid \mathbf{z},\theta) P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta), \tag{4.11}$$

qui, du fait de l'hypothèse d'indépendance conditionnelle des observations, se réécrit

$$\begin{aligned} g_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{y} \mid \Psi) &= \prod_{i \in S} \sum_{z_i} g(y_i \mid z_i,\theta)\; P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)},\beta) \\ &= \prod_{i \in S} \sum_{z_i} g(y_i \mid z_i,\tilde{\mathbf{z}}^{(q)}_{N(i)},\theta)\; P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)},\beta) \\ &= \prod_{i \in S} g_G(y_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)},\Psi). \end{aligned} \tag{4.12}$$

Les algorithmes EM de type champ moyen abordent la résolution du problème de maximisation de la vraisemblance $L(\Psi,\mathbf{y}) = g_G(\mathbf{y} \mid \Psi)$, impossible directement, en remplaçant celle-ci par l'approximation $g_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi)$, que nous noterons aussi par la suite $L_{\tilde{\mathbf{z}}}(\Psi,\mathbf{y})$, pour un champ $\tilde{\mathbf{z}}$ donné. La vraisemblance $L_{\tilde{\mathbf{z}}}(\Psi,\mathbf{y})$ est de la forme de l'approximation PPL de Qian et Titterington (1991) (voir chapitre 2.2.6). Il s'agit donc de la vraisemblance d'un modèle de mélange fini indépendant particulier dont les poids dépendent de i selon la forme du modèle de champ de Markov initial et le choix de $\tilde{\mathbf{z}}$. Grâce à l'étape (NR), les problèmes propres à EM pour le modèle de champ de Markov caché disparaissent. Le calcul de la constante de normalisation de $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta)$ ne pose pas de difficulté du fait de la factorisation de cette distribution. L'étape (E) se réduit au calcul de l'approximation de la distribution markovienne conditionnelle $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y},\Psi)$ pour la valeur courante du paramètre, qui se déduit naturellement de l'approximation (4.9) de $P_G(\mathbf{z} \mid \beta)$:

$$P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y},\Psi) = \frac{g(\mathbf{y} \mid \mathbf{z},\theta)\; P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta)}{P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{y} \mid \Psi)}, \tag{4.13}$$

qui se simplifie encore, à partir des équations (1.8) et (4.12),

$$\begin{aligned} P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y},\Psi) &= \prod_{i \in S} \left\{ \frac{g(y_i \mid z_i,\theta)\; P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)},\beta)}{\sum_{z_i} g(y_i \mid z_i,\theta)\; P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)},\beta)} \right\} \\ &= \prod_{i \in S} P_G(z_i \mid y_i,\tilde{\mathbf{z}}^{(q)}_{N(i)},\Psi) \\ &= \prod_{i \in S} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid y_i,\Psi). \end{aligned} \tag{4.14}$$

Il en résulte l'approximation correspondante de $Q_\theta(\theta \mid \Psi^{(q-1)})$ et $Q_\beta(\beta \mid \Psi^{(q-1)})$, quantités définies par (2.31) et (2.32),

$$Q_\theta(\theta \mid \Psi^{(q-1)}) \approx \sum_{i \in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid y_i,\Psi^{(q-1)})\; \log g(y_i \mid z_i,\theta),$$

et

$$Q_\beta(\beta \mid \Psi^{(q-1)}) \approx \sum_{i \in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid y_i,\Psi^{(q-1)})\; \log P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \beta).$$

Remarquons que les approximations utilisées aux étapes (E) et (M) sont cohérentes dans le sens où la règle de Bayes est satisfaite pour les probabilités $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta)$ et $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y},\Psi)$ (équations (4.9) et (4.13)). Un algorithme EM de type champ moyen peut donc s'interpréter comme un algorithme EM appliqué à un modèle de mélange fini indépendant, différent à chaque itération et correspondant au modèle simplifié courant.
Ce n'est pas le cas dans les versions de EM pour champ de Markov caché qui combinent approximation de la pseudo-vraisemblance et simulations MC (*cf.* partie 2.2.5). En effet, pour ces procédures, d'une part la distribution marginale du champ de Markov est remplaçée par l'approximation de la pseudo-vraisemblance $\mathrm{PV}(\mathbf{z} \mid \beta)$ qui n'est pas une distribution en général. D'autre part l'approximation $\hat{P}(\mathbf{z} \mid \mathbf{y},\Psi)$ de la distribution conditionnelle est calculée par simulations MC de sorte que la loi de Bayes n'est pas vérifiée entre $(\mathrm{PV}\mathbf{z} \mid \beta)$ et $\hat{P}(\mathbf{z} \mid \mathbf{y},\Psi)$. Il n'est donc pas possible d'exhiber un modèle pour les données complètes $(\mathbf{z},\mathbf{y})$ qui découlerait des approximations.
Un avantage des algorithmes de type champ moyen est qu'ils permettent de se ramener au cas de l'estimation d'un modèle de mélange fini indépendant, tout en gardant l'information spatiale présente dans le modèle initial, à travers le champ des voisins $\tilde{\mathbf{z}}$.

Malgré l'étape (NR) qui pourrait apparaître comme une étape de segmentation, notre procédure générale vient se placer dans l'approche *mélange* pour l'estimation des paramètres d'un modèle de champ de Markov caché (voir le sous-chapitre 2.2). En effet, $\tilde{\mathbf{z}}^{(q)}$ est considéré comme un ensemble de valeurs utilisées pour approximer l'interaction du voisinage avec chaque pixel, plutôt que comme une possible restauration du champ caché $\mathbf{z}$. Il est utilisé pour construire les approximations des distributions markoviennes marginale et conditionnelle. Plus précisement, à chaque itération ces procédures traitent les voisinages $\tilde{\mathbf{z}}_{N(i)}^{(q)}$ comme vrais, mais le champ z_i au pixel i est toujours supposé inconnu. Il s'agit de la même approche que celle utilisée dans la procédure PPL-EM de Qian et Titterington (1991). Dans le cas où $\tilde{\mathbf{z}}^{(q)}$ est choisi comme l'approximation en champ moyen, les composantes ne sont pas nécessairement à valeur dans $\{0,1\}$, de sorte que cela ne fournirait même pas une configuration valide du champ de Markov.

4.4 Discussion sur le choix des voisins

La flexibilité de nos procédures réside dans le choix de la valeur du champ voisin $\tilde{\mathbf{z}}^{(q)}$. Un candidat doit conduire à une approximation raisonnable de $P_G(\mathbf{z},\mathbf{y} \mid \theta,\beta)$. Dans le modèle de champ de Markov caché, $P_G(\mathbf{z} \mid \beta)$ et $P_G(\mathbf{z} \mid \mathbf{y},\Psi)$ ne sont pas disponibles en pratique, contrairement à $g(\mathbf{y} \mid \mathbf{z},\theta)$ qui est généralement choisie comme une densité d'une loi simple à manipuler (*e.g.* loi gaussienne). Connaissant la densité conditionnelle des observations, il suffit d'approximer l'une des deux quantités inconnues, $P_G(\mathbf{z} \mid \beta)$ ou $P_G(\mathbf{z} \mid \mathbf{y},\Psi)$, pour en déduire une approximation de l'autre par application de la règle de Bayes (équations (4.9) et (4.13)). Notre sélection de $\tilde{\mathbf{z}}^{(q)}$ est donc guidée par la qualité de l'approximation de la distribution markovienne soit marginale soit conditionnelle. Si l'on adopte par exemple le principe du champ moyen, les approximations ne peuvent pas à la fois être toutes les deux optimales au sens de la divergence de Kullback-Leibler et satisfaire à la règle de Bayes (voir la description de l'agorithme de Zhang dans la partie 4.1). Il faut donc dans un premier temps choisir si le principe d'approximation de type champ moyen d'une distribution de Gibbs présenté dans la partie 4.2 doit être appliqué à la distribution marginale ou conditionnelle. Un premier argument pour sélectionner $\tilde{\mathbf{z}}^{(q)}$ en se basant sur $P_G(\mathbf{z} \mid \mathbf{y},\Psi)$ est que ce choix présente l'avantage de prendre en compte directement les données $\mathbf{y}$. À cette remarque s'ajoute une considération d'ordre pratique. L'étude du modèle de Potts simple (*cf.* exemple 3.1) dissuade d'utiliser l'approximation en champ moyen du champ markovien marginal pour estimer le paramètre spatial du champ marginal dans le cas où la segmentation $\mathbf{z}$ est connue (voir Archer et Titterington 2000 pour le cas de la dimension 1). En effet, l'approximation dans ce cas présente peu d'intérêt à cause de l'existence de la solution uniforme $\mathbf{u}^K$ à l'équation de cohérence, et ce quelle que soit la valeur de β. Dans le cas de l'estimation des paramètres d'un champ de Markov caché par un algorithme EM de type champ moyen, si à l'étape (NR) $\tilde{\mathbf{z}}^{(q)}$ est choisi comme l'approximation en champ moyen de la distribution $P_G(\mathbf{z} \mid \beta)$ pour la valeur courante de β, la procédure peut s'avérer inopérante pour estimer β au cours de l'étape (EM). La mise à jour de ce paramètre est obtenue par la maximisation suivante

$$\beta^{(q)} \;=\; \arg\max_{\beta} \sum_{i\in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \mathbf{y},\Psi^{(q-1)}) \;\log P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \beta). \qquad (4.15)$$

Supposons $\tilde{\mathbf{z}}$ tel que $\tilde{z}_i$ ne dépend pas de i, hypothèse naturelle puisqu'il n'y a pas de champ externe dans le modèle. Soit $m = (m_1,m_2,\ldots,m_K)$ cette valeur commune et $\mathbf{m}$ le champ sur S égal à m en tout site, l'étape de maximisation (4.15) devient pour le modèle de Potts simple défini par (3.1)

$$\beta^{(q)} \;=\; \arg\max_{\beta} N_v\beta \sum_{k=1}^{K} m_k \sum_{i\in S} P_{\mathbf{m}}(e_k \mid \mathbf{y},\Psi^{(q-1)}) \;-\; n\log(\sum_{k=1}^{K} e^{N_v\beta m_k}), \qquad (4.16)$$

où $N_v = |N(i)|$ est supposé indépendant de i. Si de plus toutes les composantes de m sont égales (nécessairement à K^{-1}), la fonction (4.16) à maximiser en β se simplifie en $-n\log K$, qui ne dépend plus de β ! le paramètre spatial ne peut donc pas être mis à jour dans ce cas. Comme nous l'avons rappelé ci-dessus, un tel champ $\mathbf{m}$ est solution du problème de point fixe associé au modèle de Potts sans champ externe (*cf.* exemple 3.1).

Si le choix de $\tilde{\mathbf{z}}$ est fait selon $P_G(\mathbf{z} \mid \mathbf{y},\Psi)$, le problème de l'existence de la solution uniforme ne se pose plus puisque le terme d'attache aux données ajoute un champ externe au modèle de Potts, avec $V_i(z_i) = -\log g(y_i \mid z_i,\theta)$ pour tout i dans S. Notons également que si le paramètre β est connu, seule une approximation de $P_G(\mathbf{z} \mid \mathbf{y},\Psi)$ est nécessaire à l'étape (E). Ces considérations nous ont conduit à rejeter la solution d'une sélection de $\tilde{\mathbf{z}}$ basée sur la distribution markovienne marginale. Remarquons que c'est également la stratégie adoptée par Zhang (1992) puisque l'approximation (4.3) est de la forme (4.7) avec $\tilde{\mathbf{z}} = \mathbf{z}^{mf\,(q)}$, approximation en champ moyen de l'espérance de la distribution markovienne conditionnelle.

Ce choix fait, nous avons considéré différentes manières d'approximer la distribution conditionnelle des données cachées. Le choix de référence parmi les distributions de la forme

(4.7) est bien sûr l'approximation en champ moyen. Nous nommerons *algorithme en champ moyen* l'algorithme correspondant à choisir pour $\tilde{\mathbf{z}}^{(q)}$ à l'étape (NR) l'approximation en champ moyen de la distribution markovienne conditionnelle pour les paramètres courants.
Nous proposons de considérer également les approximations en champ modal et en champ simulé du champ de Markov conditionnel, initialement décrites dans la partie 4.2 dans le cas du champ de Markov marginal. Soit z_i^{mode} un mode au site i de la distribution P^{mode} du modèle factorisable correspondant à l'approximation en champ modal de $P_G(\mathbf{z} \mid \mathbf{y},\Psi)$. La distribution $P_G(z_i \mid \mathbf{y},\Psi)$ est alors approximée par $P_G(z_i \mid \mathbf{z}_{N(i)}^{mode},\mathbf{y},\Psi)$ de sorte que z_i^{mode} vérifie l'équation de cohérence associée,

$$\forall i \in S, \quad z_i^{mode} \;=\; \arg\max_{z_i} P_G(z_i \mid \mathbf{z}_{N(i)}^{mode},\mathbf{y},\Psi). \tag{4.17}$$

Si ce problème de point fixe est résolu de manière itérative, avec mise à jour séquentielle des pixels, à l'étape (NR) le champ voisin $\tilde{\mathbf{z}}$ est en fait la solution de la segmentation supervisée par ICM pour un modèle de champ de Markov caché de paramètres égaux aux paramètres courants (voir la partie 2.1.1.3). La distribution marginale $P_G(z_i \mid \beta)$ est ensuite approximée par

$$P_i^{mode}(z_i \mid \beta) = P_G(z_i \mid \mathbf{z}_{N(i)}^{mode},\beta)\,.$$

Nous appellerons *algorithme en champ modal* l'algorithme EM de type champ moyen dont l'étape (NR) est définie par l'équation (4.17) .
Nous avons également considéré le cas où $\tilde{\mathbf{z}}^{(q)}$ est une simulation de la distribution conditionnelle $P_G(\mathbf{z} \mid \mathbf{y},\Psi^{(q-1)})$, obtenue par l'échantillonneur de Gibbs. Cela conduit à l'*algorithme en champ simulé*. Le choix de cette méthode de restauration des voisins est d'esprit analogue à l'étape (S) de l'algorithme SEM. L'utilisation d'une restauration stochastique à l'étape (NR) permet à l'algorithme de se déplacer plus librement dans l'espace d'états, par rapport aux algorithmes en champ moyen et en champ modal.
Les détails de mise en œuvre des algorithmes sont décrits dans la partie 4.7. En résumé l'étape (NR) pour les algorithmes en champ moyen, en champ modal et en champ simulé consiste à

(NR) choix des voisins

Algorithme en champ moyen :	fixer $\tilde{\mathbf{z}}^{(q)}$ à l'estimation en champ moyen de l'espérance de la distribution conditionnelle $P_G(\mathbf{z} \mid \mathbf{y},\Psi^{(q-1)})$;
Algorithme en champ modal :	fixer $\tilde{\mathbf{z}}^{(q)}$ à l'estimation en champ modal du mode de la distribution conditionnelle $P_G(\mathbf{z} \mid \mathbf{y},\Psi^{(q-1)})$;
Algorithme en champ simulé :	simuler $\tilde{\mathbf{z}}^{(q)}$ selon la distribution conditionnelle $P_G(\mathbf{z} \mid \mathbf{y},\Psi^{(q-1)})$, via l'échantillonneur de Gibbs.

La structure générale définissant les algorithmes EM de type champ moyen permet de donner une interprétation simple du fonctionnement des algorithmes en champ moyen, en champ modal et en champ simulé : à chaque itération le modèle de champ de Markov caché est approximé par un modèle où les données manquantes sont indépendantes, puis les paramètres sont estimés par EM sur le modèle simplifié. Pour les trois algorithmes étudiés, l'étape (NR) de choix des voisins nécessite la résolution d'une équation déterministe ou la réalisation d'une seule simulation. Cela conduit à des procédures moins coûteuses que les versions de EM pour champ de Markov caché basées sur des simulations MC (EM gibbsien, MCEM, SRE, voir partie 2.2.5).
Les propriétés théoriques des trois méthodes de mise à jour de $\tilde{\mathbf{z}}^{(q)}$ proposées ci-dessus sont de natures différentes. Dans le cas de l'approximation en champ moyen, la distribution $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$ avec $\tilde{\mathbf{z}}^{(q)}$ solution du problème de point fixe correspondant, est la meilleure approximation de $P_G(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$ au sens de la divergence de Kullback-Leibler. Pour l'approximation en champ modal, la procédure itérative de résolution du problème (4.17) converge vers un maximum local de $P_G(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$, mais nous ne savons pas si la distribution $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$ correspondante approche bien la distribution markovienne conditionnelle. Dans le cas de l'algorithme en champ simulé, $\tilde{\mathbf{z}}^{(q)}$ est une réalisation de la distribution d'intérêt $P_G(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$. La mise à jour stochastique des voisins au cours de l'algorithme assure la convergence de la procédure, comme nous le démontrons ci-après.

4.5 Convergence de l'algorithme en champ simulé

Considérons l'algorithme suivant, très proche de l'algorithme en champ simulé, dont l'itération q est définie par

(NR) simuler $\tilde{\mathbf{z}}^{(q)}$ selon la distribution conditionnelle $P_G(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$ via l'échantillonneur de Gibbs ;

(EM) calculer $\Psi^{(q)}_{\mathrm{MV}} = \arg\max_{\Psi} L(\Psi, \tilde{\mathbf{z}}^{(q)}, \mathbf{y})$ puis $\Psi^{(q)} = em(\tilde{\mathbf{z}}^{(q)}, \mathbf{y}, \Psi^{(q)}_{\mathrm{MV}})$,

où $\Psi^{(q)}_{\mathrm{MV}}$ est l'estimateur du maximum de vraisemblance complète pour les données complètes $(\tilde{\mathbf{z}}^{(q)}, \mathbf{y})$ (calculé éventuellement à partir de l'approximation de la pseudo-vraisemblance) et où $em(\tilde{\mathbf{z}}, \mathbf{y}, \Psi_0)$ est l'estimateur de Ψ obtenu par maximisation de la vraisemblance $L_{\tilde{\mathbf{z}}}(\Psi, \mathbf{y})$ par Q itérations de EM initialisé par une valeur donnée Ψ_0 du paramètre. Ainsi, pour cet algorithme, à chaque itération le champ $\tilde{\mathbf{z}}^{(q)}$ est considéré comme une réalisation du champ $\mathbf{Z}$ uniquement pour le calcul de l'initialisation de l'étape (EM). Ensuite, ce champ joue à nouveau le même rôle que dans les algorithmes EM de type champ moyen. Cet algorithme définit une chaîne de Markov $(\Psi^{(q)})_{q \in \mathbb{N}}$ qui possède les propriétés suivantes :

Proposition 1 *La chaîne $\{\Psi^{(q)}\}_{q\in I\!N}$ est une chaîne de Markov ergodique. De plus la convergence est uniformément géométrique.*

De manière plus générale, il suffit que la valeur du paramètre utilisée à l'itération q pour initialiser l'étape (EM) soit indépendante de $\Psi^{(q-1)}$ pour que l'algorithme correspondant vérifie la proposition 1. L'algorithme défini ci-dessus diffère légèrement de l'algorithme en champ simulé. En effet, dans la définition générale proposée dans la partie 4.2, l'algorithme EM est initialisé par la valeur courante des paramètres $\Psi^{(q-1)}$ et non pas par $\Psi^{(q)}_{\mathrm{MV}}$. Sans cette modification la suite $\{\tilde{\mathbf{z}}^{(q)}\}_{q\in I\!N}$ n'est pas une chaîne de Markov puisque le calcul de $\Psi^{(q)}$ dépend de $\Psi^{(q-1)}$, et la proposition 1 n'est pas vérifiée.
Si la convergence est établie seulement pour le cas où l'étape (EM) est initialisée par $\Psi^{(q)}_{\mathrm{MV}}$, on peut penser qu'en pratique $\Psi^{(q)}_{\mathrm{MV}}$ et $\Psi^{(q-1)}$ sont proches et qu'ainsi l'algorithme en champ simulé a un comportement similaire à l'algorithme décrit ci-dessus. L'algorithme en champ simulé permet alors d'éviter le calcul supplémentaire de $\Psi^{(q)}_{\mathrm{MV}}$ à chaque itération.
Notons également qu'à l'étape (NR) l'échantillonneur de Gibbs doit être initialisé par une valeur arbitraire $\mathbf{z}_0$ afin d'assurer que $\{\Psi^{(q)}\}_{q\in I\!N}$ est également une chaîne de Markov. En pratique ce n'est pas le cas, puisqu'au cours de l'algorithme en champ simulé l'échantillonneur de Gibbs est initialisé par $\tilde{\mathbf{z}}^{(q-1)}$ (voir partie 4.7).

Démonstration de la proposition 1 : il est évident, de part la définition de l'algorithme, que la chaîne $\{\tilde{\mathbf{z}}^{(q)}\}_{q\in I\!N}$ est une chaîne de Markov. Les coefficients de la matrice de transition de cette chaîne sont des produits de probabilités conditionnelles de la forme $P_G(\tilde{\mathbf{z}}_i \mid \tilde{\mathbf{z}}'_{N(i)}, y_i, \Psi)$ puisque $\tilde{\mathbf{z}}^{(q)}$ est obtenu par l'échantillonneur de Gibbs. Ces coefficients sont strictements positifs d'après la propriété de positivité des modèles de champs de Markov. On en déduit que la chaîne $\{\tilde{\mathbf{z}}^{(q)}\}_{q\in I\!N}$ est irréductible et apériodique. De plus, cette chaîne étant à espace d'états fini, elle possède une unique mesure invariante P^* et elle est géométriquement ergodique.
L'étape (EM) étant déterministe, la probabilité conditionnelle $P(\Psi^{(q)} \mid \mathbf{z}^{(q)})$ est la loi de Dirac $\delta_{em(\tilde{\mathbf{z}}^{(q)},\mathbf{y},\Psi_0)}(\Psi^{(q)})$ égale à 1 si $\Psi^{(q)} = em(\tilde{\mathbf{z}}^{(q)},\mathbf{y},\Psi_0)$ et nulle sinon. La chaîne $\{\Psi^{(q)}\}_{q\in I\!N}$ est donc une chaîne de Markov sur $\Gamma = \{em(\mathbf{z},\mathbf{y},\Psi_0),\ \mathbf{z} \in V^n\}$ qui est un ensemble fini. La probabilité de transition $P(\Psi \mid \Psi')$ est égale à $\sum_{\mathbf{z}} P(\mathbf{z} \mid \Psi')\delta_{em(\mathbf{z},\mathbf{y},\Psi_0)}(\Psi)$. Elle est donc strictement positive quel que soit Ψ et Ψ' dans Γ, puisque la probabilité de transition $P(\mathbf{z} \mid \Psi')$ est strictement positive. Ainsi, $\{\Psi^{(q)}\}_{q\in I\!N}$ possède une unique mesure invariante π^* et elle est géométriquement ergodique. Plus précisement, la mesure π^* est définie par

$$\begin{aligned}\forall A \in \Gamma, \quad \pi^*(A) &= \sum_{\mathbf{z}} P^*(\mathbf{z} \mid \mathbf{y})\delta_{em(\mathbf{z},\mathbf{y},\Psi_0)}(A) \\ &= P^*(\{\mathbf{z} \ / \ em(\mathbf{z},\mathbf{y},\Psi_0) \in A\}).\end{aligned}$$

□

Remarquons que ces propriétés sont valables quel que soit le nombre Q d'itérations de EM réalisées au cours de l'étape (EM).

La proposition 1 suggère un comportement rassurant de l'algorithme en champ simulé, alors que l'on ne peut assurer la convergence théorique dans le cas des algorithmes en champ moyen et en champ modal. Cependant elle ne renseigne pas sur la nature de la distribution stationnaire P^* et donc les caractéristiques de π^* restent inconnues. On peut espérer des résultats plus précis, notamment sur le comportement asymptotique de l'algorithme en champ simulé (*i.e.* lorsque la taille de l'image augmente), à l'instar des résultats existant pour l'algorithme SEM dans le cas indépendant. Ainsi, Celeux et Diebolt (1993) dans le cadre des mélanges finis indépendants et plus récemment Nielsen (2000) dans un contexte plus général où les données complètes $\{(Z_i,Y_i),i = 1,\ldots,n\}$ sont indépendantes et identiquement distribuées, ont montré que, lorsque la taille n de l'échantillon tend vers l'infini, alors la distribution stationnaire de la chaîne de Markov ergodique générée par SEM converge vers une distribution gaussienne dont la moyenne est l'estimateur du maximum de vraisemblance du paramètre du modèle. On peut penser s'inspirer de ces résultats dans un contexte avec dépendance, cependant on est loin de savoir le faire car le cas des modèles de champs de Markov cachés est difficile.

4.6 Variantes pour l'étape d'estimation

À la place de l'algorithme EM, d'autres algorithmes peuvent être envisagés à l'étape d'estimation des algorithmes EM de type champ moyen, comme CEM (Celeux et Govaert 1992) ou SEM (Celeux et Diebolt 1985). Ces deux procédures insèrent après l'étape (E) une étape au cours de laquelle une configuration $\mathbf{z}^{(q)}$ est générée puis utilisée comme image segmentée dans l'étape (M). Dans CEM, $\mathbf{z}^{(q)}$ est générée suivant le principe du MAP, alors qu'elle est simulée dans SEM. En supposant qu'une seule itération de l'algorithme EM est réalisée à l'étape d'estimation de l'algorithme EM de type champ moyen, celle-ci prend alors la forme suivante

- **(E)** identique à l'étape (E) de EM ;
- étape supplémentaire

(C) pour CEM

$$\forall i \in S, \quad z_i^{(q)} = \arg\max_{z_i} P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q)},y_i,\Psi^{(q-1)}),$$

(S) pour SEM,

pour tout i dans S, $z_i^{(q)}$ est simulé selon $P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q)},y_i,\Psi^{(q-1)})$,
qui est proportionnel à $g(y_i \mid z_i,\theta^{(q-1)}) P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q)},\beta^{(q-1)})$,

en considérant $\mathbf{z}^{(q)}$ comme une segmentation des observations, l'étape (M) devient

- **(M)** mettre à jour les estimateurs d'après

$$\begin{aligned}\theta^{(q)} &= \arg\max_{\theta} \sum_{i \in S} \log g(y_i \mid z_i^{(q)}, \theta), \\ \beta^{(q)} &= \arg\max_{\beta} \sum_{i \in S} \log P_{\tilde{\mathbf{z}}^{(q)}}(z_i^{(q)} \mid \beta).\end{aligned}$$

A priori, toute combinaison des méthodes de génération du champ voisin à l'étape (NR) et d'estimation des paramètres peuvent être considérées. Cependant certaines semblent plus naturelles, comme l'approximation en champ modal et CEM, car dans ce cas $\tilde{\mathbf{z}}^{(q)}$ et $\mathbf{z}^{(q)}$ sont tous les deux générés suivant le principe du MAP. Plus précisement, puisque $\tilde{\mathbf{z}}^{(q)}$ est la solution du problème de point fixe suivant,

$$\forall i \in S, \quad \tilde{z}_i^{(q)} = \arg\max_{z_i} P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q)}, y_i, \Psi^{(q-1)}), \tag{4.18}$$

s'il y a unicité de la solution de ce problème alors par définition de l'étape (C) nous avons $\tilde{\mathbf{z}}^{(q)} = \mathbf{z}^{(q)}$. Dans ce cas une itération de la procédure devient

(NR) résoudre (4.18) ;

(EM) calculer $P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q)}, y_i, \Psi^{(q-1)})$ pour tout i dans S et pour tout z_i et mettre à jour le paramètre Ψ d'après

$$\begin{aligned}\theta^{(q)} &= \arg\max_{\theta} \sum_{i \in S} \log g(y_i \mid \tilde{z}_i^{(q)}, \theta), \\ \beta^{(q)} &= \arg\max_{\beta} \sum_{i \in S} \log P_G(\tilde{z}_i^{(q)} \mid \tilde{\mathbf{z}}_{N(i)}^{(q)}, \beta).\end{aligned} \tag{4.19}$$

On retrouve ainsi une itération de l'algorithme ICM non supervisé (paragraphe 2.1.2.4). Comme il a été remarqué dans la partie précédente, l'étape (NR) est équivalente à l'étape de segmentation par ICM. On reconnaît ensuite la maximisation de la pseudo-vraisemblance avec l'équation (4.19), de sorte que l'étape (2) est identique à l'étape de mise à jour de paramètres dans ICM non supervisé. L'équivalence entre ICM non supervisé et l'algorithme EM de type champ moyen correspondant à la combinaison de l'approximation en champ modal pour l'étape (NR) et d'une itération de CEM pour l'étape d'estimation se retrouve également si $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$ est remplacée à l'étape (C) par la distribution de Dirac donnant une probabilité 1 à la configuration $\tilde{\mathbf{z}}^{(q)}$, puisque cela assure $\tilde{\mathbf{z}}^{(q)} = \mathbf{z}^{(q)}$.

4.7 Mise en œuvre

Dans cette partie, nous spécifions la stratégie utilisée pour mettre en œuvre les trois algorithmes EM de type champ moyen proposés. En principe, pour les algorithmes en champ

moyen, en champ modal et en champ simulé, l'étape (NR) de mise à jour des voisins consiste respectivement à résoudre un problème de point fixe de type (3.9), réaliser l'algorithme ICM non supervisé et simuler suivant une distribution de Gibbs. Chacune de ces trois opérations induit une procédure itérative pour laquelle la mise à jour de $\tilde{\mathbf{z}}^{(q)}$ peut être réalisée soit en parallèle soit séquentiellement. Dans le dernier cas, le champ en chaque pixel est mis à jour en utilisant immédiatement les nouvelles valeurs disponibles sur les autres pixels. Dans une mise à jour en parallèle (ou asynchrone), les nouvelles valeurs des voisins ne sont utilisées qu'une fois que tous les pixels ont été mis à jour. En pratique, Dang (1998) remarque que la résolution de l'équation de cohérence du champ moyen avec un schéma séquentiel conduit à une convergence plus rapide et évite de possibles phénomènes d'oscillations (effets de damier). En ce qui concerne l'algorithme ICM (voir Besag 1986 et Wu et Doerschuk 1995), la convergence est garantie dans le cas séquentiel mais pas dans le cas d'une mise à jour en parallèle. Dans l'algorithme en champ simulé, la version asynchrone n'est pas valide (Geman et Geman 1984) en ce sens qu'une distribution stationnaire existe mais que sa nature est inconnue. Pour ces raisons expérimentales et théoriques, nous avons adopté une mise à jour séquentielle dans nos algorithmes, prenant en compte les nouvelles valeurs de $\tilde{\mathbf{z}}^{(q)}_{N(i)}$ pour mettre à jour le champ au site i. Ce choix correspond à une stratégie globale : chaque nouvelle valeur dans l'ensemble $\{\tilde{\mathbf{z}},\theta,\beta\}$ est immédiatement prise en compte pour la mise à jour des autres. En particulier, une seule itération des procédures définissant l'étape (NR) est réalisée et celles-ci sont initialisées par $\tilde{\mathbf{z}}^{(q-1)}$. Comme dans Besag (1986), un seul *cycle* est parcouru pour la mise à jour de $\tilde{\mathbf{z}}^{(q)}$, un *cycle* correspondant au parcours de tous les pixels. De la même manière, nous utilisons une seule itération de EM dans l'étape (EM). En accord avec ces choix, les définitions des étapes (NR) et (EM) des algorithmes que nous avons effectivement utilisées dans nos expérimentations sont celles détaillées ci-dessous.
Soit $\tilde{\mathbf{z}}^{(q-1+\frac{i}{n})} = \{\tilde{z}_1^{(q)}, \ldots, \tilde{z}_i^{(q)}, \tilde{z}_{i+1}^{(q-1)}, \ldots, \tilde{z}_n^{(q-1)}\}$, la configuration mise à jour jusqu'au site i. Les étapes (NR) de choix des voisins pour les algorithmes en champ moyen, en champ modal et en champ simulé sont définies par

(NR) choix des voisins

Algorithme en champ moyen

$$\tilde{\mathbf{z}}^{(q)} = I\!E_{P_{\tilde{\mathbf{z}}^{(q-1)}}(.\mid \mathbf{y},\Psi^{(q-1)})}[\mathbf{Z}],$$

ce qui, d'après l'égalité (4.14) et pour une mise à jour séquentielle, est équivalent à

$$\forall i \in S, \quad \tilde{z}_i^{(q)} = I\!E_{P_G(\,.\mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})},y_i,\Psi^{(q-1)})}[Z_i],$$

$$i.e.\ \ \tilde{z}_i^{(q)} = \frac{\sum\limits_{z_i} z_i \exp\{-\sum\limits_{c\ni i} V_c(z_i,\tilde{\mathbf{z}}_{c\backslash\{i\}}^{(q-1+\frac{i-1}{n})} \mid \beta^{(q-1)}) + \log g(y_i \mid z_i,\, \theta^{(q-1)})\}}{\sum\limits_{z_i} \exp\{-\sum\limits_{c\ni i} V_c(z_i,\tilde{\mathbf{z}}_{c\backslash\{i\}}^{(q-1+\frac{i-1}{n})} \mid \beta^{(q-1)}) + \log g(y_i \mid z_i,\, \theta^{(q-1)})\}}\ .$$

Algorithme en champ modal

$\tilde{\mathbf{z}}^{(q)} = \arg\max_{\mathbf{z}} P_{\tilde{\mathbf{z}}^{(q-1)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$,

ce qui, d'après l'égalité (4.14) et pour une mise à jour séquentielle, conduit à

$\forall i \in S, \quad \tilde{z}_i^{(q)} = \arg\max_{z_i} P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, y_i, \Psi^{(q-1)})$,

i.e. $\tilde{z}_i^{(q)} = \arg\max_{z_i} g(y_i \mid z_i, \theta^{(q-1)}) P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, \beta^{(q-1)})$.

Algorithme en champ simulé

pour tout i dans S, $\tilde{z}_i^{(q)}$ est simulé selon $P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, y_i, \Psi^{(q-1)})$,
qui est proportionnel à $g(y_i \mid z_i, \theta^{(q-1)}) P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, \beta^{(q-1)})$.

Une seule itération de EM est ensuite réalisée dans l'étape (EM), utilisant la distribution $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z}, \mathbf{y} \mid \Psi)$ au lieu de $P_G(\mathbf{z}, \mathbf{y} \mid \Psi)$, de sorte que cette étape devient

(EM) estimation

(E) calculer $P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \mathbf{y}, \Psi^{(q-1)})$ pour tout i dans S,

(M) mettre à jour les estimateurs $\Psi^{(q)} = (\theta^{(q)}, \beta^{(q)})$ selon

$$\theta^{(q)} = \arg\max_{\theta} \sum_{i \in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \mathbf{y}, \Psi^{(q-1)}) \ \log g(y_i \mid z_i, \theta),$$

et

$$\beta^{(q)} = \arg\max_{\beta} \sum_{i \in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \mathbf{y}, \Psi^{(q-1)}) \ \log P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \beta). \tag{4.20}$$

4.8 Liens avec d'autres algorithmes

Nous avons, dans les sous-chapitres 2.1 et 2.2, distingué deux approches pour l'estimation des paramètres d'un modèle de champ de Markov caché. Les méthodes de l'approche *image* combinent segmentation suivant le principe du MAP et estimation par maximisation de la vraisemblance complète des paramètres $L_c(\Psi, \mathbf{z}, \mathbf{y})$ pour les observations et la segmentation courante. Cette approche induit un biais des estimateurs et une solution consiste à se placer dans une approche *mélange*. Les méthodes de cette dernière approche visent à maximiser la vraisemblance des paramètres $L(\Psi, \mathbf{y})$ et travaillent avec les probabilités conditionnelles du champ de Markov caché. Les algorithmes EM de type champ moyen décrits par la structure générale présentée dans la partie 4.3 se placent dans l'approche *mélange*. Au début de chaque itération est créée une *configuration de travail* $\tilde{\mathbf{z}}^{(q)}$ avant l'étape de mise à jour du paramètre Ψ par maximisation de l'approximation de la vraisemblance des paramètres, $L_{\tilde{z}^{(q)}}(\Psi, \mathbf{y})$ via

EM.
Il apparaît que des algorithmes connus d'estimation dérivés de l'approche *mélange* peuvent s'écrire dans la structure générale des algorithmes EM de type champ moyen, ce qui apporte une nouvelle vision de ces méthodes. Ainsi, si une seule itération de l'algorithme EM est réalisée dans l'étape (EM), l'algorithme en champ moyen est l'algorithme de Zhang (1992) (voir partie 4.1). En effet, les deux algorithmes se basent sur l'approximation en champ moyen de l'espérance conditionnelle du champ de Markov pour approximer les distributions marginale et conditionnelle du champ caché dans les formules de EM. La lecture de l'algorithme de Zhang dans notre structure permet de mettre en évidence la cohérence (*i.e.* vérification de la règle de Bayes) entre le choix des approximations de $P_G(\mathbf{z} \mid \beta)$ et de $P_G(\mathbf{z} \mid \mathbf{y},\theta)$ proposées à l'origine de manière heuristique. En effet, ces deux approximations définissent un modèle pour les données complètes et il apparaît qu'une itération de l'algorithme de Zhang revient à appliquer une itération de EM sur ce modèle simplifié.
D'autre part, comme il a été remarqué précédemment, l'approximation $L_{\tilde{\mathbf{z}}}(\Psi,\mathbf{y})$ de la vraisemblance $L(\Psi,\mathbf{y})$ utilisée dans les algorithmes EM de type champ moyen a la même forme que l'approximation PPL de Qian et Titterington (1991) (*cf.* équations (2.35) et (4.12)). Ainsi, la structure des algorithmes de type champ moyen est équivalente à la structure en deux étapes (restauration et maximisation) de la procédure générale PPL-EM proposée par Qian et Titterington (1991) (voir chapitre 2.2.6), puisque en pratique les auteurs suggérent de réaliser l'étape de maximisation avec EM. Ils suggérent également d'utiliser ICM ou une segmentation par MAP pour générer $\tilde{\mathbf{z}}$ à l'étape de restauration du voisinage, ce qui rejoint l'esprit de l'algorithme en champ modal.
L'algorithme de Zhang et la procédure générale de Qian et Titterington (1991) sont deux solutions pour la mise en œuvre de l'algorithme EM pour champ de Markov caché basées sur l'idée de se ramener à un système de variables indépendantes (soit par approximation en champ moyen des interactions, soit par approximation de la vraisemblance des paramètres). Il n'est donc pas surprenant de constater que ces deux méthodes s'interprètent dans une même structure unifiée : la structure des algorithmes EM de type champ moyen.
En ce qui concerne l'algorithme en champ simulé, l'intérêt de cette procédure par rapport à d'autres algorithmes stochastiques de l'approche *mélange* pour l'estimation des paramètres d'un champ de Markov caché (voir partie 2.2.5) est de ne nécessiter qu'une seule simulation $\tilde{\mathbf{z}}^{(q)}$ à chaque itération. Néanmoins, l'algorithme en champ simulé et MCEM sont deux procédures de structures très proches lorsque le paramètre spatial β est connu, ce qui devient d'autant plus évident lorsque l'on généralise la première en simulant plusieurs voisinages $\tilde{\mathbf{z}}_1^{(q)},\ldots,\tilde{\mathbf{z}}_{T_q}^{(q)}$ à chaque itération. Cette idée est développée dans le chapitre 5, ce qui conduit à un algorithme que nous nommons *algorithme en champ simulé généralisé* et pour lequel nous établissons qu'il possède les mêmes propriétés de convergence que MCEM.
Enfin, il est possible de se ramener à l'approche *image* en rajoutant une étape de classification dans l'étape (EM) de chaque itération d'un algorithme EM de type champ moyen. En particulier, on retrouve ainsi l'algorithme ICM non supervisé à partir de l'algorithme en champ modal, comme il a été détaillé dans la partie 4.6.

4.9 Étude expérimentale

Dans cette partie, nous illustrons les performances des algorithmes EM de type champ moyen dans le cas d'un modèle de Markov simple, le modèle de Potts sans champ externe. Nous rappelons la définition de la fonction énergie de ce modèle :

$$H(\mathbf{z} \mid \beta) = -\beta \sum_{i \sim j} z_i^t z_j ,$$

avec le paramètre spatial β supposé positif afin de favoriser les cliques homogènes. Nous avons considéré des distributions $g(y_i \mid z_i,\theta)$ gaussiennes. Si z_i appartient à la classe k alors $g(y_i \mid z_i,\theta)$ est une densité gaussienne de moyenne m_k et d'écart-type σ_k,

$$g(y_i \mid Z_i = e_k,\theta) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\{-\frac{(y_i - m_k)^2}{2\sigma_k^2}\}. \tag{4.21}$$

Les paramètres du modèle de champ de Markov caché sont donc le paramètre spatial β et le paramètre de bruit $\theta = \{(\mu_k,\sigma_k),1 \leq k \leq K\}$.
Spécifions pour ce modèle les étapes (NR) et (EM) décrites dans la partie 4.7. À l'étape (NR) de l'algorithme en champ moyen, une itération de l'algorithme de résolution du problème de point fixe s'écrit dans notre stratégie séquentielle

$$\forall i \in S \ , \quad \tilde{z}_i^{(q)} = \frac{\sum\limits_{z_i} z_i \exp\{\beta^{(q-1)} z_i^t \sum\limits_{j \in N(i)} \tilde{z}_j^{(q-1+\frac{i-1}{n})} + \log g(y_i \mid z_i,\, \theta^{(q-1)})\}}{\sum\limits_{z_i} \exp\{\beta^{(q-1)} z_i^t \sum\limits_{j \in N(i)} \tilde{z}_j^{(q-1+\frac{i-1}{n})} + \log g(y_i \mid z_i,\, \theta^{(q-1)})\}} \ .$$

Les quantités utilisées dans les algorithmes en champ modal et en champ simulé sont proportionnelles à

$$g(y_i \mid z_i,\theta^{(q-1)}) P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})},\beta^{(q-1)}) \ \propto \ g(y_i \mid z_i,\theta^{(q-1)}) \exp(\beta^{(q-1)} z_i^t \sum_{j \in N(i)} \tilde{z}_j^{(q-1+\frac{i-1}{n})}) \ .$$

Comme nous l'avons mentionné dans la partie 4.7, une seule itération de la procédure de mise à jour du voisinage $\tilde{\mathbf{z}}^{(q)}$ est réalisée au cours de l'étape (NR). Nous n'avons pas observé d'amélioration significative en utilisant plus d'itérations. De la même façon, une seule itération de l'algorithme EM pour modèle de mélange fini indépendant est ensuite réalisée et l'étape (EM) s'écrit

$$\forall 1 \leq k \leq K, \quad m_k^{(q)} = \frac{\sum_{i=1}^n y_i \, P_{\tilde{\mathbf{z}}^{(q)}}(Z_i = e_k \mid y_i,\Psi^{(q-1)})}{\sum_{i=1}^n P_{\tilde{\mathbf{z}}^{(q)}}(Z_i = e_k \mid y_i,\Psi^{(q-1)})},$$

$$\forall 1 \leq k \leq K, \quad (\sigma_k^2)^{(q)} = \frac{\sum_{i=1}^n (y_i - m_k^{(q)})^2 P_{\tilde{\mathbf{z}}^{(q)}}(Z_i = e_k \mid y_i,\Psi^{(q-1)})}{\sum_{i=1}^n P_{\tilde{\mathbf{z}}^{(q)}}(Z_i = e_k \mid y_i,\Psi^{(q-1)})},$$

$$\beta^{(q)} = \arg\max_{\beta} \sum_{i \in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \mathbf{y},\Psi^{(q-1)}) \ \log P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \beta).$$

Il n'existe pas de formule explicite pour la mise à jour du paramètre spatial, contrairement au cas du paramètre θ. Nous avons utilisé une procédure de recherche du zéro du gradient de la fonction à maximiser en β, dans l'intervalle [0,10], par dichotomie associée à une interpolation linéaire ou quadratique (Forsythe, Malcolm, et Moler 1980).
Les résultats reportés sont ceux obtenus pour les algorithmes en champ moyen, en champ modal et en champ simulé. Nous comparons leurs performances avec l'algorithme EM pour modèle de mélange fini indépendant (que l'on appellera aussi EM indépendant) et avec l'algorithme ICM. L'algorithme EM indépendant est le seul algorithme parmi les cinq étudiés qui ne prend pas en compte l'information spatiale (voir partie 2.2.2). Nous utilisons une version non supervisée de l'algorithme ICM, pour lequel à chaque itération le paramètre β est estimé par maximisation de la pseudo-vraisemblance (voir paragraphe 2.1.2.4), et cela en utilisant la fonction de recherche de zéro citée ci-dessus.
Pour chaque expérience, les algorithmes ont tous été initialisés avec la même segmentation obtenue par simple seuillage des observations $\mathbf{y}$: l'intervalle des valeurs dans l'image observée est divisé en K sous-intervalles réguliers et chacun est assigné à une classe. Afin d'étudier les algorithmes dans des conditions équivalentes, nous avons utilisé le même nombre N d'itérations pour chacun. Ce choix nous a paru plus pertinent qu'un critère d'arrêt basé sur les paramètres ou sur la segmentation. En effet, dans le cas de l'algorithme en champ moyen ou en champ modal nous n'avons pas de garantie de convergence de ces quantités. Le temps d'exécution observé pour la réalisation d'une itération des algorithmes en champ moyen, en champ modal, en champ simulé et ICM est sensiblement le même (de l'ordre de 4 secondes pour 10 itérations pour une image de taille 100×100). L'algorithme ICM non supervisé fournit directement une segmentation de l'image observée. Pour les algorithmes en champ moyen, en champ modal et en champ simulé, les segmentations reportées sont le résultat de la maximisation de la distribution conditionnelle fournie par les algorithmes, $P_{\tilde{\mathbf{z}}^{(N)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(N)})$:

$$\begin{aligned} &\forall i \in S, \quad \hat{z}_i = e_l, \\ &\text{avec } e_l = \arg \max_{1 \le k \le K} P_{\tilde{\mathbf{z}}^{(N)}}(Z_i = e_k \mid \mathbf{y}, \Psi^{(N)}). \end{aligned} \tag{4.22}$$

Dans le cas de l'algorithme EM indépendant, la segmentation est obtenue par maximisation des estimations des probabilités conditionnelles des étiquettes (*cf.* équation (2.27))

$$\begin{aligned} &\forall i \in S, \quad \hat{z}_i = e_l, \\ &\text{avec } e_l = \arg \max_{1 \le k \le K} t_{ik}^{(N)} . \end{aligned} \tag{4.23}$$

Lorsque la vraie segmentation $\mathbf{z}$ est connue, nous reportons également les taux d'erreur (*i.e.* pourcentage de pixels mal classés), qui donnent une information supplémentaire sur la capacité des algorithmes à retrouver $\mathbf{z}$, en plus du critère visuel.

Le code en langage C qui a été développé pour réaliser les expérimentations présentées dans ce mémoire se base sur un programme de classification spatiale de Dang (1998), disponible à *http://www.hds.utc.fr/ mdang/*. Il s'agit à l'origine d'un programme pour la segmentation utilisant l'algorithme NEM et que nous avons complété, entre autre, par les algorithmes

EM de type champ moyen et l'algorithme ICM.

Nous illustrons les différents algorithmes sur plusieurs exemples. Dans la partie 4.9.1 nous comparons les algorithmes en termes d'estimation des paramètres du modèle et nous étudions la stabilité des résultats sur des exemples simulés. Les performances en estimation et en segmentation des algorithmes sont ensuite testées sur des images avec bruit simulé pour différentes valeur de K et du paramètre du bruit (partie 4.9.2). Enfin, le dernier exemple traité est celui d'une image réelle (partie 4.9.3).

4.9.1 Simulations d'un modèle de Potts

Pour cette première série d'expérimentations, nous avons simulé des réalisations de taille 150×150 d'un modèle de Potts simple avec un voisinage d'ordre 1 et pour plusieurs valeurs de K et de β, que nous avons bruitées avec un bruit gaussien. Nous avons ensuite utilisé les cinq algorithmes considérés pour estimer les paramètres du modèle de champ de Markov caché. Les réalisations du modèle de Potts ont été obtenues par l'échantillonneur de Gibbs (*cf.* partie 1.4). Ensuite la procédure de création de l'image bruitée est la suivante : si l'on note $cl(z_i)$ l'étiquette du pixel i, c'est à dire $cl(z_i) = k$ si $z_i = e_k$ pour $k \in \{1, \ldots, K\}$, alors l'observation y_i se décompose de la manière suivante

$$\forall i \in S, \quad y_i = cl(z_i) - 1 + \epsilon_i, \tag{4.24}$$

avec ϵ_i réalisation d'une loi gaussienne d'espérance nulle et d'écart-type $\sigma = 0.5$ et $\{\epsilon_i, i \in S\}$ formant un système de variables indépendantes. Cependant, afin de juger la capacité des algorithmes à retrouver cette information, nous avons utilisé un modèle avec variance dépendant de la classe pour estimer les paramètres. Pour un nombre de classes K et une valeur du paramètre spatial β donnés, 20 réalisations indépendantes du modèle de Potts simple correspondant ont été générées puis bruitées. Pour chaque image nous avons ensuite utilisé $N = 100$ itérations des algorithmes pour estimer les paramètres du modèle. Pour des modèles à 2, 3 et 4 classes, nous avons reporté dans les tableaux 4.1 à 4.5 la valeur moyenne et la variance des estimateurs (entre parenthèses) sur les 20 expériences et pour chaque algorithme. Nous avons également réalisé ces expériences pour un modèle de Potts à 4 et à 5 classes. Les résultats en ce qui concerne l'estimation du paramètre de bruit θ n'apportant pas de nouvelles informations nous avons uniquement reporté la valeur des estimations de β dans le tableau 4.6 pour ces deux expériences.

Il ressort de ces expériences que les résultats fournis par les cinq algorithmes étudiés sont relativement stables. L'algorithme EM indépendant présente néanmoins les plus grandes variations sur les estimations des paramètres du bruit sur l'ensemble des 20 simulations. On note aussi une plus grande variabilité dans l'estimation du paramètre spatial avec l'algorithme ICM. Les performances des algorithmes restent également stables lorsque K varie. Nous observons que l'algorithme en champ moyen tend à surestimer le paramètre β, et ce d'autant plus que la vraie valeur de ce paramètre est petite. Nous n'expliquons pas vraiment

l'origine de ce phénomène. Une raison possible est que, de part son effet "moyennisateur", l'approximation en champ moyen d'un modèle de champ de Markov pourrait conduire à des réalisations plus homogènes par rapport au modèle initial, renforçant ainsi les dépendances spatiales à travers le paramètre β. L'algorithme en champ modal fournit les plus fines estimations de β, sauf pour les cas des simulations en deux classes avec $\beta = 0.2$ et $\beta = 0.6$. Après l'algorithme en champ modal, les meilleures estimations du paramètre spatial sont fournies par l'algorithme en champ simulé, même si on observe une tendance de cet algorithme à sous-estimer β. En termes d'estimation des paramètres de bruit, tous les algorithmes sauf ICM conduisent à des résultats satisfaisants. On retrouve donc sur ces exemples le phénomène de biais inhérent à l'algorithme ICM, déjà mentionné dans le chapitre 2.2.6. Les différentes méthodes retrouvent bien le fait que les variances des classes sont identiques lorsque $K = 2$. Elles y parviennent moins bien pour les exemples avec $K \geq 3$.
Nous avons également indiqué la moyenne et la variance du pourcentage de pixels mal classés (taux d'erreur) dans les tableaux 4.1 à 4.5. Pour $K = 2$, ces résultats peuvent être comparés avec le taux théorique d'erreur inconditionnelle de classement par modèle de mélange fini indépendant. Dans le cas d'un mélange de lois gaussiennes avec variances communes et proportions égales, ce taux est donné par la formule

$$Err = \phi(-\frac{|\mu_1 - \mu_2|}{2\sigma}),$$

où ϕ est la fonction de répartition de la loi normale centrée réduite et où σ est la variance commune (voir Celeux 1990). On obtient pour $\mu_1 = 0$, $\mu_2 = 1$ et $\sigma = 0.5$ un taux d'erreur théorique $Err = \phi(-1) \simeq 15.87\%$. Dans le cas d'un modèle de Potts simple binaire et pour les trois valeurs de β testées, l'algorithme EM indépendant permet d'atteindre ce taux théorique. Il est intéressant de noter que lorsque l'on utilise le modèle de champ de Markov caché pour restaurer la réalisation initiale, en moyenne les taux d'erreur sont bien plus faibles que le taux d'erreur théorique. Ceci met en avant le gain obtenu avec la prise en compte de l'information spatiale. Comme l'on pouvait s'y attendre, ce gain devient plus important lorsque la valeur du paramètre spatial augmente : pour $\beta = 0.2$ les taux d'erreur moyens obtenus par les quatres algorithmes intégrant l'information spatiale sont proches du taux d'erreur théorique (au mieux 15.64% d'erreur avec ICM), alors que pour $\beta = 0.78$ les taux d'erreur moyens pour ces mêmes algorithmes varient entre 10.14% et 13.86%. Les expériences réalisées pour des modèles à 3 et 4 classes confirment l'importance de la prise en compte de l'information spatiale : on constate aussi sur ces exemples une diminution sensible du taux d'erreur lorsque l'on utilise un modèle de champ de Markov caché.

	β	μ_1	μ_2	σ_1	σ_2	taux d'erreur
valeurs exactes	0.2	0	1	0.5	0.5	-
EM indépendant	-	0.00	1.00	0.50	0.50	15.88
	-	($1.2\ 10^{-}4$)	($1.6\ 10^{-}4$)	($2.3\ 10^{-}5$)	($5.3\ 10^{-}5$)	(0.04)
champ moyen	0.53	0.00	1.00	0.50	0.50	16.55
	(0.001)	($8.7\ 10^{-}5$)	($1.3\ 10^{-}4$)	($2.5\ 10^{-}5$)	($5.9\ 10^{-}5$)	(0.17)
champ modal	0.28	0.00	1.00	0.50	0.50	15.74
	($6.3\ 10^{-}4$)	($5.4\ 10^{-}5$)	($5.1\ 10^{-}5$)	($2.5\ 10^{-}5$)	($2.9\ 10^{-}5$)	(0.07)
champ simulé	0.22	0.00	1.00	0.50	0.49	15.86
	($3.3\ 10^{-}4$)	($7.5\ 10^{-}5$)	($3.5\ 10^{-}5$)	($1.9\ 10^{-}5$)	($2.9\ 10^{-}5$)	(0.05)
ICM	0.19	-0.08	1.08	0.40	0.40	15.64
	($5.3\ 10^{-}4$)	($7.8\ 10^{-}5$)	($7.5\ 10^{-}5$)	($1.3\ 10^{-}5$)	($0.9\ 10^{-}5$)	(0.04)

TAB. 4.1 – *moyenne et variance des estimations des paramètres d'un modèle de Potts simple et du pourcentage de pixels mal classés, pour* $K = 2$ *et* $\beta = 0.2$.

	β	μ_1	μ_2	σ_1	σ_2	taux d'erreur
valeurs exactes	0.6	0	1	0.5	0.5	-
EM indépendant	-	0.00	1.00	0.50	0.50	15.91
	-	($2.3\ 10^{-}4$)	($2.3\ 10^{-}4$)	($6.0\ 10^{-}5$)	($4.5\ 10^{-}4$)	(0.04)
champ moyen	0.86	0.01	1.00	0.51	0.51	13.86
	($5.8\ 10^{-}4$)	($4.6\ 10^{-}4$)	($3.3\ 10^{-}4$)	($1.5\ 10^{-}4$)	($7.3\ 10^{-}5$)	(0.30)
champ modal	0.66	-0.01	1.01	0.49	0.49	13.38
	($2.7\ 10^{-}4$)	($1.4\ 10^{-}4$)	($6.8\ 10^{-}5$)	($3.5\ 10^{-}5$)	($0.9\ 10^{-}5$)	(0.08)
champ simulé	0.57	-0.01	1.01	0.49	0.49	13.86
	($2.0\ 10^{-}4$)	($3.5\ 10^{-}5$)	($4.9\ 10^{-}5$)	($1.6\ 10^{-}5$)	($2.8\ 10^{-}5$)	(0.11)
ICM	0.62	-0.06	1.06	0.42	0.42	13.25
	($8.1\ 10^{-}4$)	($7.2\ 10^{-}5$)	($1.4\ 10^{-}4$)	($3.2\ 10^{-}5$)	($2.6\ 10^{-}5$)	(0.08)

TAB. 4.2 – *moyenne et variance des estimations des paramètres d'un modèle de Potts simple et du pourcentage de pixels mal classés, pour* $K = 2$ *et* $\beta = 0.6$.

	β	μ_1	μ_2	σ_1	σ_2	taux d'erreur
valeurs exactes	0.78	0	1	0.5	0.5	-
EM indépendant	-	-0.03	0.97	0.49	0.51	16.00
	-	(2.6 $10^{-}4$)	(2.9 $10^{-}4$)	(3.3 $10^{-}5$)	(9.1 $10^{-}5$)	(0.06)
champ moyen	0.93	0.01	1.01	0.51	0.50	10.14
	(3.4 $10^{-}4$)	(7.3 $10^{-}5$)	(9.0 $10^{-}5$)	(1.9 $10^{-}5$)	(4.2 $10^{-}5$)	(0.05)
champ modal	0.81	-0.02	1.00	0.49	0.50	10.44
	(3.2 $10^{-}4$)	(4.4 $10^{-}5$)	(7.6 $10^{-}5$)	(1.2 $10^{-}5$)	(2.4 $10^{-}5$)	(0.09)
champ simulé	0.73	-0.02	1.00	0.48	0.49	10.78
	(2.4 $10^{-}4$)	(8.9 $10^{-}5$)	(7.3 $10^{-}5$)	(4.2 $10^{-}5$)	(3.0 $10^{-}5$)	(0.05)
ICM	0.98	-0.04	1.02	0.47	0.47	10.52
	(0.02)	(4.6 $10^{-}4$)	(3.0 $10^{-}4$)	(3.9 $10^{-}4$)	(3.4 $10^{-}4$)	(0.19)

TAB. 4.3 – *moyenne et variance des estimations des paramètres d'un modèle de Potts simple et du pourcentage de pixels mal classés, pour $K = 2$ et $\beta = 0.78$.*

	β	μ_1	μ_2	μ_3
valeurs exactes	0.9	0	1	2
EM indépendant	-	0.00	0.97	1.96
	-	($9.9\ 10^{-3}$)	($3.0\ 10^{-2}$)	($1.1\ 10^{-2}$)
champ moyen	1.02	-0.01	1.00	2.01
	($4.5\ 10^{-4}$)	($2.3\ 10^{-4}$)	($2.7\ 10^{-4}$)	($3.1\ 10^{-4}$)
champ modal	0.90	-0.02	0.99	2.01
	($4.8\ 10^{-4}$)	($2.7\ 10^{-4}$)	($4.4\ 10^{-4}$)	($7.9\ 10^{-5}$)
champ simulé	0.83	-0.01	0.99	2.00
	($4.8\ 10^{-4}$)	($1.1\ 10^{-4}$)	($2.8\ 10^{-4}$)	($2.7\ 10^{-4}$)
ICM	0.84	-0.06	0.99	2.04
	($1.4\ 10^{-3}$)	($2.7\ 10^{-4}$)	($4.8\ 10^{-4}$)	($2.6\ 10^{-4}$)

	σ_1	σ_2	σ_3	taux d'erreur
valeurs exactes	0.5	0.5	0.5	-
EM indépendant	0.50	0.47	0.51)	21.46
	($2.5\ 10^{-4}$)	($2.4\ 10^{-4}$)	($1.3\ 10^{-4}$)	(0.74)
champ moyen	0.49	0.48	0.49	13.84
	($5.3\ 10^{-5}$)	($4.6\ 10^{-4}$)	($1.2\ 10^{-4}$)	(0.26)
champ modal	0.48	0.46	0.49	14.60
	($2.5\ 10^{-5}$)	($1.1\ 10^{-4}$)	($4.7\ 10^{-5}$)	(0.30)
champ simulé	0.48	0.44	0.49	15.32
	($5.8\ 10^{-5}$)	($2.2\ 10^{-4}$)	($1.1\ 10^{-4}$)	(0.17)
ICM	0.43	0.34	0.43	15.84
	($6.2\ 10^{-5}$)	($2.1\ 10^{-4}$)	($7.4\ 10^{-5}$)	(0.26)

TAB. 4.4 – *moyenne et variance des estimations des paramètres d'un modèle de Potts simple pour $K = 3$ et $\beta = 0.9$.*

	β	μ_1	μ_2	μ_3	μ_4
valeurs exactes	1	0	1	2	3
EM indépendant	-	0.04	1.01	1.94	2.95
	-	($8.9\ 10^{-4}$)	($3.0\ 10^{-3}$)	($4.0\ 10^{-3}$)	($2.0\ 10^{-3}$)
champ moyen	1.12	-0.01	1.00	2.00	3.01
	($2.0\ 10^{-4}$)	($2.8\ 10^{-4}$)	($2.8\ 10^{-4}$)	($3.6\ 10^{-4}$)	($2.5\ 10^{-5}$)
champ modal	0.98	-0.01	2.00	2.00	3.01
	($2.2\ 10^{-4}$)	($2.9\ 10^{-4}$)	($7.0\ 10^{-4}$)	($4.8\ 10^{-4}$)	($2.4\ 10^{-4}$)
champ simulé	0.93	-0.00	1.01	2.01	3.01
	($4.9\ 10^{-4}$)	($4.6\ 10^{-4}$)	($8.0\ 10^{-4}$)	($1.0\ 10^{-3}$)	($3.0\ 10^{-4}$)
ICM	0.91	-0.05	1.01	1.99	3.04
	($6.6\ 10^{-4}$)	($7.2\ 10^{-4}$)	($2.4\ 10^{-3}$)	($1.6\ 10^{-3}$)	($3.8\ 10^{-4}$)

	σ_1	σ_2	σ_3	σ_4	taux d'erreur
valeurs exactes	0.5	0.5	0.5	0.5	-
EM indépendant	0.51	0.45	0.46	0.52	24.54
	($3.3\ 10^{-4}$)	($5.9\ 10^{-4}$)	($3.9\ 10^{-4}$)	($3.6\ 10^{-4}$)	(0.33)
champ moyen	0.49	0.47	0.48	0.49	15.80
	($7.7\ 10^{-5}$)	($3.1\ 10^{-4}$)	($2.5\ 10^{-4}$)	($8.2\ 10^{-5}$)	(0.30)
champ modal	0.48	0.44	0.44	0.49	16.74
	($7.0\ 10^{-5}$)	($1.1\ 10^{-4}$)	($5.4\ 10^{-5}$)	($1.0\ 10^{-4}$)	(0.29)
champ simulé	0.49	0.44	0.43	0.49	17.61
	($1.2\ 10^{-4}$)	($3.0\ 10^{-4}$)	($2.7\ 10^{-4}$)	($1.1\ 10^{-5}$)	(0.37)
ICM	0.43	0.33	0.33	0.43	18.30
	($1.0\ 10^{-4}$)	($9.2\ 10^{-5}$)	($1.3\ 10^{-4}$)	($7.0\ 10^{-5}$)	(0.53)

TAB. 4.5 – *moyenne et variance des estimations des paramètres d'un modèle de Potts simple pour $K = 4$ et $\beta = 1$.*

	$K = 5,\ \beta = 1$	$K = 6,\ \beta = 1.1$
champ moyen	1.15	1.25
	($1.8\ 10^{-4}$)	($7.6\ 10^{-5}$)
champ modal	0.99	1.08
	($1.5\ 10^{-4}$)	($1.4\ 10^{-4}$)
champ simulé	0.92	1.03
	($3.5\ 10^{-4}$)	($2.2\ 10^{-4}$)
ICM	0.89	0.99
	($3.8\ 10^{-4}$)	($4.0\ 10^{-4}$)

TAB. 4.6 – *moyenne et variance des estimations du paramètre spatial d'un modèle de Potts simple, pour $K = 5$, $\beta = 1$ et $K = 6$, $\beta = 1.1$.*

	β	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	taux d'erreur
valeurs exactes	-	1	2	3	4	0.5	0.5	0.5	0.5	-
EM indépendant	-	0.8	1.8	3.5	4.1	0.4	0.7	0.7	0.4	27.4
champ moyen	2.0	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5	0.5
champ modal	1.8	1.0	2.1	3.1	4.0	0.5	0.5	0.5	0.5	2.8
champ simulé	2.1	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5	0.4
ICM	1.1	1.0	2.1	3.2	4.0	0.5	0.6	0.5	0.5	4.6

TAB. 4.7 – *image Damier: estimation des paramètres et taux d'erreur pour les différentes méthodes.*

4.9.2 Images avec bruit simulé

Dans cette partie nous considérons deux images en niveaux de gris que nous avons bruitées avec un bruit gaussien, selon la procédure décrite dans la partie précédente, la formule (4.24) étant remplacée par

$$\forall i \in S, \quad y_i = cl(z_i) + \epsilon_i.$$

Pour ces exemples, l'image originale n'a pas été générée selon un modèle de champ de Markov. Nous avons fait l'hypothèse d'un modèle de Potts simple avec voisinage d'ordre 2.

Image Damier : la première image traitée est l'image Damier représentée sur la figure 4.1 (a). Il s'agit d'une image de taille 128×128 composée de 4 niveaux de gris codés 1, 2, 3 et 4. Les paramètres du bruit sont donc $\theta = \{(\mu_k,\sigma_k), k = 1,\ldots,4\}$ avec $\mu_k = k$ et $\sigma_k = 0.5$ pour $k = 1,\ldots,4$. La variance est donc commune aux 4 classes. Ici encore, nous avons utilisé un modèle avec variance dépendant de la classe pour estimer les paramètres puis pour restaurer l'image. L'image bruitée, l'image utilisée pour initialiser les algorithmes et les segmentations obtenues sont montrées sur les figures 4.1 (b) à (h). Les paramètres et les taux d'erreur correspondants sont reportés dans le tableau 4.7. Pour cet exemple, le nombre d'itérations était fixé à $N = 100$. On constate que les quatre algorithmes utilisant un modèle prenant en compte l'information spatiale retrouvent le fait que les variances des classes sont égales. Les performances de ces algorithmes sont nettement supérieures à celles de EM indépendant qui fournit une segmentation encore très bruitée, conséquence d'une mauvaise estimation des paramètres du bruit. Les algorithmes en champ moyen, en champ modal et en champ simulé fournissent des estimations assez proches du paramètre β. On remarque que l'algorithme en champ moyen et l'algorithme en champ simulé ont des résultats semblables et conduisent à de meilleures estimations des paramètres du bruit et à un plus faible taux d'erreur que l'algorithme en champ modal et ICM. Sur cet exemple, les segmentations par ces deux dernières procédures sont très semblables, bien qu'avec un taux d'erreur plus grand pour ICM, ce qui rejoint la remarque faite dans la partie 4.6 sur la proximité de leurs structures.

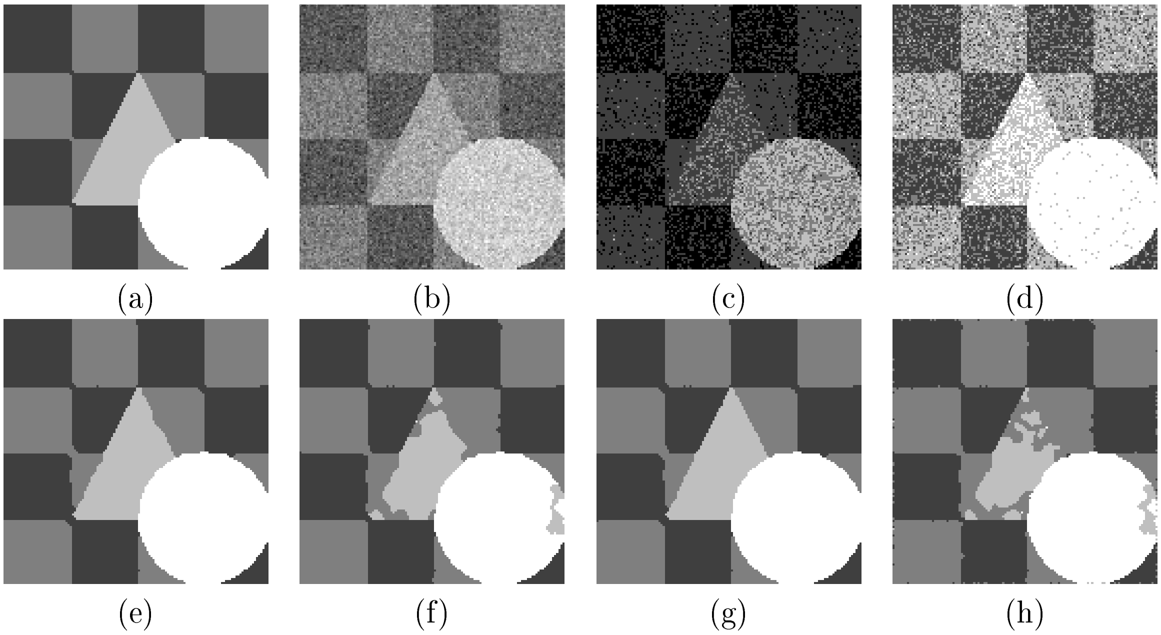


FIG. 4.1 – *image Damier dégradée par un bruit gaussien et segmentations fournies par les différentes méthodes : (a) image originale, (b) image bruitée, (c) initialisation, (d) EM indépendant, (e) algorithme en champ moyen, (f) algorithme en champ modal, (g) algorithme en champ simulé, (h) ICM.*

Image Logo : le second exemple traite le cas de l'image binaire Logo de taille 142×133 de la figure 4.2 (a). Afin d'étudier les variations des performances des algorithmes en fonction du bruit, nous avons testé deux bruits gaussiens. Dans le premier cas, les variances des lois gaussiennes dépendent de la classe et les paramètres du bruit sont $(\mu_1,\sigma_1) = (51{,}130)$ et $(\mu_2,\sigma_2) = (255{,}300)$. Les résultats des cinq algorithmes sont reportés sur la figure 4.2 et dans le tableau 4.8. Dans le second cas les variances sont communes de sorte que l'on a $(\mu_1,\sigma_1) = (51{,}300)$ et $(\mu_2,\sigma_2) = (255{,}300)$, ce qui correspond à une plus forte dégradation de l'image originale. Les résultats correspondants sont illustrés sur la figure 4.3 et dans le tableau 4.8. Nous avons dans les deux cas utilisé un modèle de bruit avec variance dépendant de la classe et un voisinage d'ordre 2 dans les algorithmes d'estimation. Cette image est plus difficile que l'image Damier car elle est formée d'une classe comprenant les pixels des régions fines dessinant les lettres i et R et d'une classe constituant le fond de l'image. On peut craindre que la classe constituant les lettres soit absorbée par la classe des pixels du fond dans le cas d'une méthode de segmentation non supervisée utilisant l'information spatiale. C'est ce que l'on constate effectivement, et cela de manière plus évidente dans le cas d'un bruit avec variance commune (figures 4.3 (e) à (h)).
Étudions d'abord le cas d'un bruit avec variances distinctes. Comme dans l'exemple précédent, d'un point de vue visuel la segmentation par EM indépendant (figure 4.2 (d)) est clairement moins satisfaisante que les segmentations par les méthodes spatiales. Ceci est confirmé par un fort taux d'erreur. Notons cependant que les paramètres du bruit sont bien estimés par EM indépendant, ce qui signifie que c'est l'étape de segmentation (4.23) qui est à incriminer. La prise en compte de l'état du voisinage permet de corriger la majeure partie des erreurs de classification, comme le montrent les figures 4.2 (e) à (h). Parmi les procédures spatiales, on observe de meilleurs résultats avec l'algorithme en champ moyen et l'algorithme en champ simulé, notamment au niveau de l'estimation des moyennes μ_1 et μ_2. On retrouve sur cet exemple le problème de biais inhérent à ICM et on observe le même effet avec l'algorithme en champ modal. Les résultats reportés pour cet exemple correspondent à $N = 200$ itérations de chaque procédure. Afin de nous assurer de la convergence des algorithmes nous avons refait les expériences avec 1000 itérations : aucune dégradation des segmentations n'est observée.
Le deuxième exemple de dégradation de l'image Logo, avec une variance commune pour les K classes, est intéressant car on observe dans ce cas des résultats inverses par rapport à une dégradation avec variances distinctes. En effet, même si globalement les performances des algorithmes sont moins bonnes puisque le bruit est plus fort, ce sont maintenant l'algorithme en champ modal et ICM qui fournissent les meilleures segmentations visuellement et aussi par rapport au nombre de pixels mal classés. L'algorithme en champ moyen conduit à une segmentation dégénérée où les lettres du logo ont complètement disparues. Cet effet de surlissage est dû au fait que le paramètre spatial est estimé à 10 avec cet algorithme. Cela signifie que la procédure d'estimation de β n'a pas convergé. Ici le comportement de l'algorithme en champ simulé se détache de celui de l'algorithme en champ moyen. Ces deux procédures fournissant des résultats assez semblables dans le cas de l'image Damier et sur le premier exemple de l'image Logo, nous avons réalisé plus d'itérations de l'algorithme en champ si-

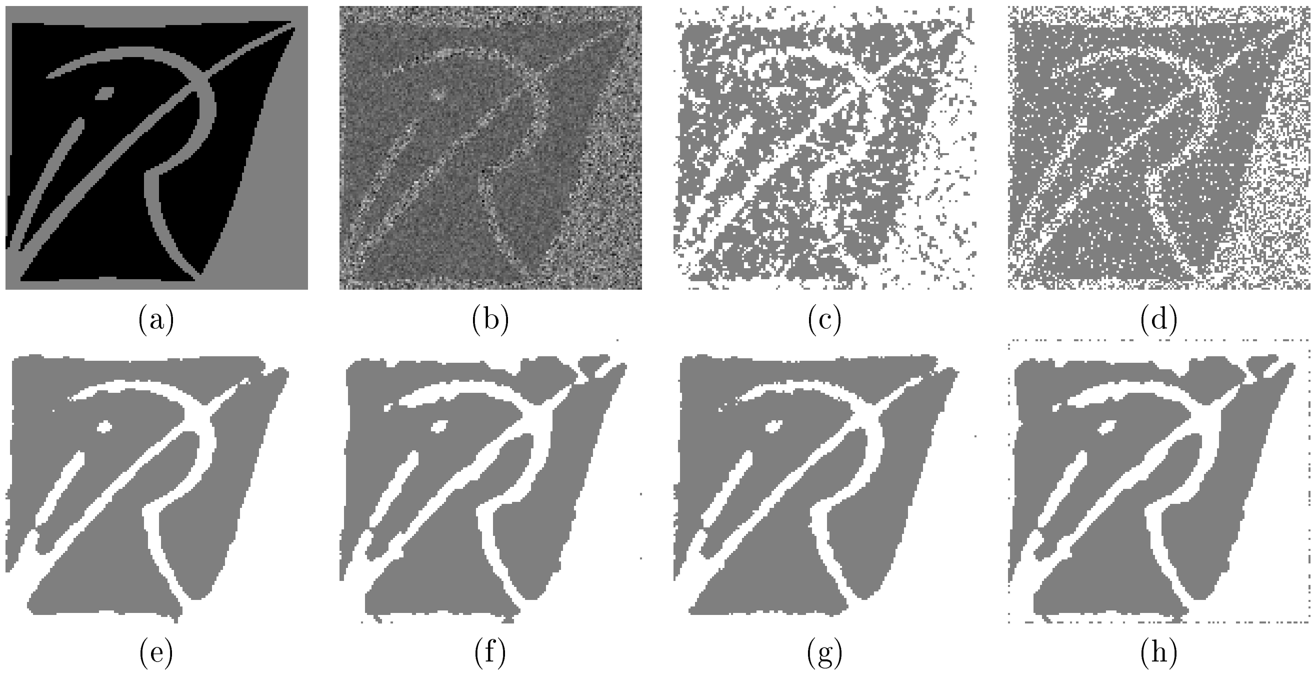
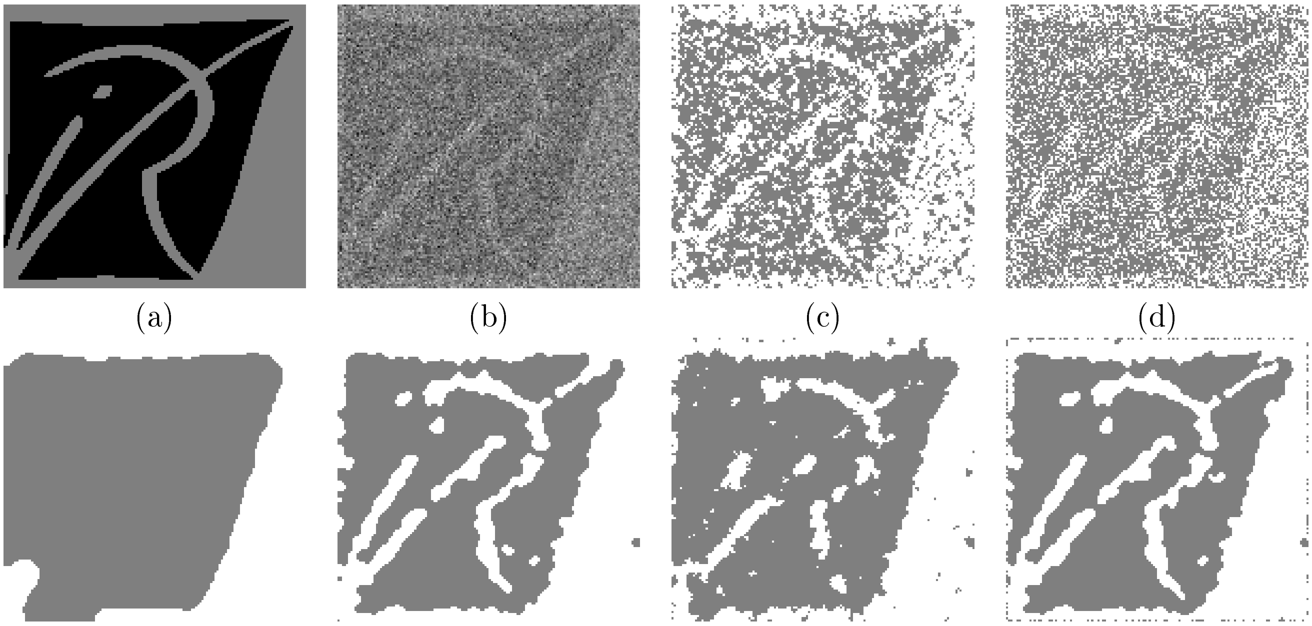


Fig. 4.2 – *image Logo dégradée par un bruit gaussien avec variance dépendant de la classe et segmentations fournies par les différentes méthodes : (a) image originale, (b) image bruitée, (c) initialisation, (d) EM indépendant, (e) algorithme en champ moyen, (f) algorithme en champ modal, (g) algorithme en champ simulé, (h) ICM.*

mulé (1500 itérations), ce qui a permis de vérifier que les différences observées ne sont pas l'effet d'une convergence lente de l'algorithme en champ simulé. Nous avons également utilisé les algorithmes en champ modal et ICM avec 1500 itérations sans observer de changement significatif dans les segmentations. Il est difficile d'interpréter les estimations du paramètre spatial, les différentes méthodes donnant des valeurs peu homogènes.

La comparaison des deux tests sur l'image Logo illustre le fait que lorsque le mélange est formé de densités mal séparées les performances de l'algorithme en champ simulé et en champ moyen se dégradent. Les algorithmes en champ modal et ICM semblent moins touchés. En effet, le principe de segmentation par MAP revient à un partitionnement de l'histogramme des données, ce qui dans le cas d'un mélange de deux densités gaussiennes conduit à sous-estimer la plus petite espérance et à surestimer la plus grande. Ce qui est plutôt décrit comme un défaut de la méthode en terme d'estimation, devient un avantage en terme de segmentation dans le cas de densités très mélangées puisque la méthode renforce la distance entre les classes et évite des résultats du type de la figure 4.3 (e) où une classe l'emporte sur l'autre.



(a) (b) (c) (d)

(e) (f) (g) (h)

FIG. 4.3 – *image Logo dégradée par un bruit gaussien avec variance commune et segmentations fournies par les différentes méthodes : (a) image originale, (b) image bruitée, (c) initialisation, (d) EM indépendant, (e) algorithme en champ moyen, (f) algorithme en champ modal, (g) algorithme en champ simulé, (h) ICM.*

	β	μ_1	μ_2	σ_1	σ_2	taux d'erreur
valeurs exactes	-	51	255	130	300	-
EM indépendant	-	52	255	128	304	22.11
champ moyen	1.26	52.4	253	129	303	3.24
champ modal	1.01	50.2	245	127	299	5.63
champ simulé	0.79	52.3	253	128	303	3.45
ICM	0.89	46.8	251	128	296	5.94

	β	μ_1	μ_2	σ_1	σ_2	taux d'erreur
valeurs exactes	-	51	255	300	300	-
EM indépendant	-	55.3	232	297	314	35.61
champ moyen	10	92	256	311	303	13.98
champ modal	3.88	49.6	259	298	302	7.7
champ simulé	0.45	56.6	281	299	298	10.30
ICM	0.86	39.2	275	295	295	8.2

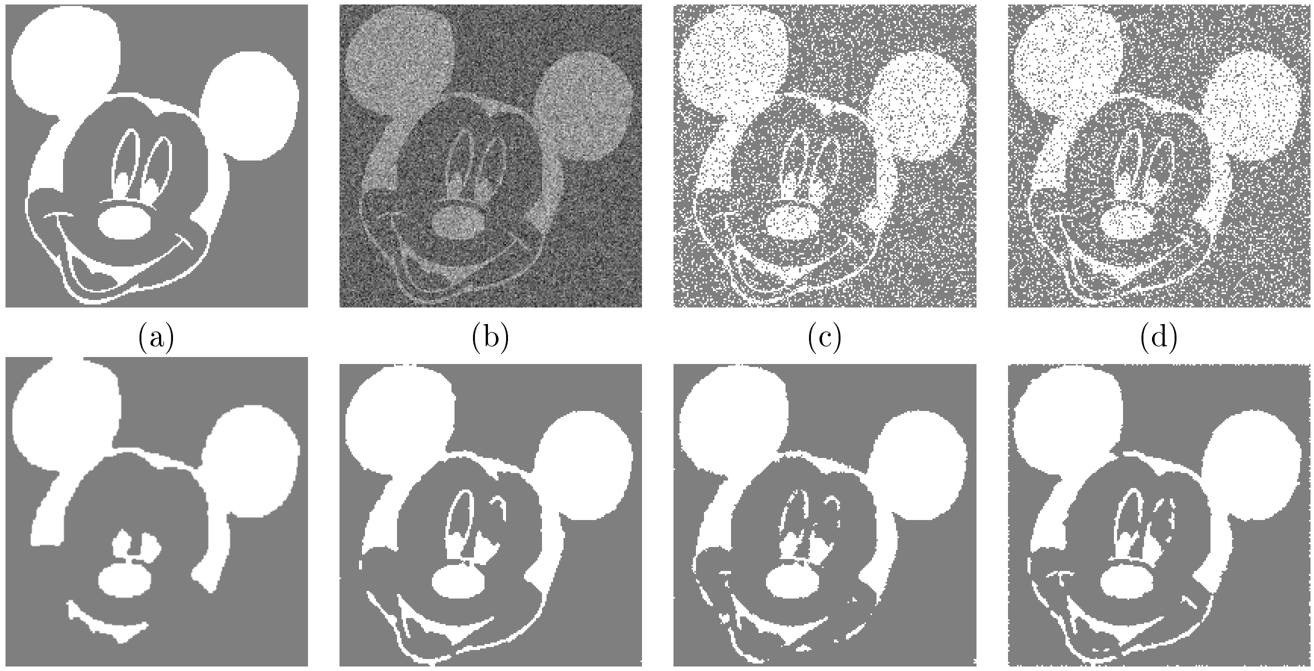
TAB. 4.8 – *image Logo : estimation des paramètres et taux d'erreur pour les différentes méthodes.*

	β	μ_1	μ_2	σ_1	σ_2	taux d'erreur
valeurs exactes	-	1	2	0.5	0.5	-
EM indépendant	-	0.91	1.81	0.47	0.56	12.99
champ moyen	2.50	1.03	1.99	0.53	0.50	0.34
champ modal	1.46	1.00	1.98	0.51	0.59	0.73
champ simulé	0.73	1.00	2.00	0.50	0.51	0.52
ICM	0.82	0.99	1.98	0.50	0.49	0.90

TAB. 4.9 – *image Mickey: estimation des paramètres et taux d'erreur pour les différentes méthodes.*

Image Mickey : la particularité de l'image Mickey (figure 4.4 (a)) est la présence de lignes fines représentant les joues, le sourire et les yeux du personnage. Les valeurs utilisées pour bruiter l'image sont $(\mu_1,\sigma_1) = (1,0.5)$ et $(\mu_2,\sigma_2) = (2,0.5)$, ce qui conduit à l'image dégradée de la figure 4.4 (b). Cet exemple permet d'illustrer la tendance de l'algorithme en champ moyen à surlisser les segmentations, tendance déjà observée avec l'image Logo (figure 4.3 (e)). L'algorithme en champ moyen donne l'estimation la plus grande du paramètre spatial (*cf.* tableau 4.9), ce qui signifie une plus grande dépendance spatiale entre les états de chaque pixel. Cela se traduit par l'absorption des lignes trop fines puisque les pixels concernés ont un voisinage presque complètement inclu dans la classe "fond". Ici encore nous avons vérifié que les meilleurs résultats, en terme de segmentation, des autres algorithmes spatiaux ne sont pas dus à une convergence lente. Les quatres méthodes spatiales fournissent de bonnes estimations du bruit (*cf.* tableau 4.9), aussi c'est essentiellement les différences dans l'estimation du paramètre β qui explique les différences au niveau des segmentations. L'algorithme EM indépendant, comme dans les exemples précédents, ne parvient pas à supprimer tout le bruit. Les algorithmes en champ simulé, en champ modal et ICM fournissent les segmentations les plus satisfaisantes visuellement, même si en terme de pixels mal classés l'algorithme en champ moyen donne de meilleurs résultats. Cette dernière remarque montre combien la qualification de "bonne segmentation" est subjective et dépend de l'objectif recherché dans la segmentation.

Remarquons également une tendance qui se dessine au fil des expériences présentées et qui se confirme dans le cas de l'image Mickey : l'estimation de β par l'algorithme en champ moyen est généralement supérieure à celles fournies par les autres méthodes, et plus précisemment, les algorithmes en champ moyen et en champ modal conduisent à une estimation du paramètre spatial plus grande que les algorithmes en champ simulé et ICM.



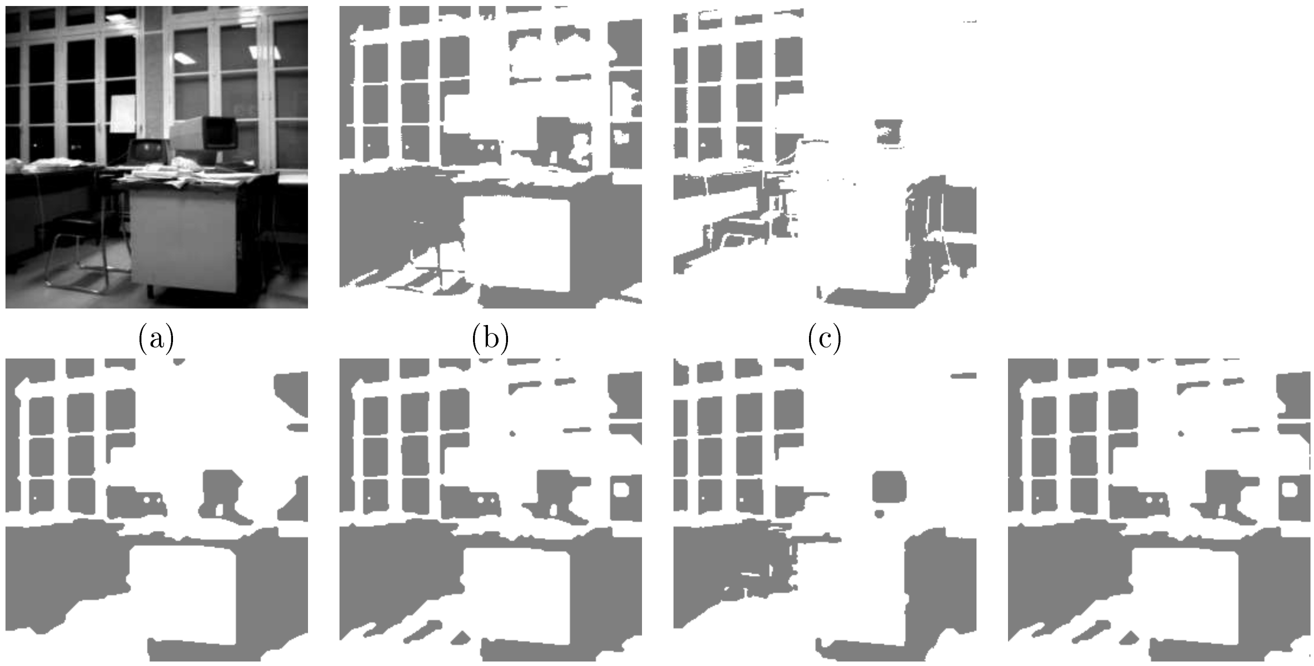
(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4.4 – *image Mickey dégradée par un bruit gaussien et segmentations fournies par les différentes méthodes : (a) image originale, (b) image bruitée, (c) initialisation, (d) EM indépendant, (e) algorithme en champ moyen, (f) algorithme en champ modal, (g) algorithme en champ simulé, (h) ICM.*

4.9.3 Image réelle

Le dernier exemple traité est une image en niveau de gris de taille 256×256 représentant une scène de bureau (figure 4.5 (a)). Il n'existe pas de valeur exacte du nombre de classes présentes dans les données. Il est évident que le choix d'un grand nombre de classes conduira à une segmentation se rapprochant plus, visuellement, de l'image réelle. Nous avons testé les algorithmes dans le cas d'une segmentation en 2 classes, l'objectif de cet exemple n'étant pas tant d'obtenir des segmentations fines que d'attirer l'attention sur un comportement de l'algorithme en champ moyen. Les segmentations obtenues à partir des cinq méthodes étudiées et pour $N = 100$ sont reportées sur les figures 4.5 (c) à (g). Nous avons utilisé un modèle de Potts simple avec un voisinage d'ordre 2 et un mélange de lois gaussiennes de moyennes et variances distinctes. On constate que toutes ne capturent pas les mêmes détails de l'image réelle, notamment au niveau du quadrillage des fenêtres dans le coin supérieur gauche et le coin supérieur droit de l'image et pour la reconstitution de l'ordinateur. On remarque par exemple que sur les segmentations obtenues par l'algorithme en champ moyen, en champ modal et ICM la structure de la fenêtre en haut à gauche n'est pas complètement retrouvée, contrairement au cas de la segmentation par l'algorithme en champ simulé. Il faut 300 itérations à l'algorithme en champ moyen pour retrouver ces détails (voir figure 4.6). La segmentation obtenue pour 1000 itérations du même algorithme est identique. Ainsi dans ce cas, l'apparente tendance de cet algorithme à surlisser l'image segmentée est en fait due à une convergence lente. Nous avons également testé les trois autres algorithmes spatiaux avec $N = 1000$ itérations, sans constater d'évolution significative par rapport aux segmentations obtenues en 100 itérations, de sorte que seul l'algorithme en champ moyen présente un tel comportement.



(a) (b) (c)

(d) (e) (f) (g)

FIG. 4.5 – *image Bureau et segmentations en deux classes fournies par les différentes méthodes : (a) image observée, (b) initialisation, (c) EM indépendant, (d) algorithme en champ moyen, (e) algorithme en champ modal, (f) algorithme en champ simulé, (g) ICM.*

FIG. 4.6 – *image Bureau : évolution de l'algorithme en champ moyen. De gauche à droite, segmentations pour* $N = 100$, 200 *et* 300 *itérations.*

4.9.4 Conclusion

Plusieurs faits ressortent de cette étude expérimentale. D'abord, l'intérêt de la prise en compte de l'information spatiale pour la classification des pixels d'une image apparaît de manière évidente dans les différents exemples traités. Comme il fallait s'y attendre, le choix d'un modèle de mélage fini indépendant conduit à des segmentations encore très bruitées et l'on observe une nette amélioration au niveau de l'homogénéité de la segmentation avec un modèle spatial. Ce résultat est rassurant car cela confirme, dans le cas de ICM, et montre, dans le cas des algorithmes EM de type champ moyen, que les approximations du modèle de champ de Markov utilisées préservent l'information markovienne.
Les expériences sur des simulations d'un modèle de Potts simple montrent que les résultats fournis par les algorithmes EM de type champ moyen, ainsi que EM indépendant et ICM sont relativement stables. Les différents exemples étudiés permettent de mettre en avant les bonnes performances des algorithmes en champ simulé et en champ modal. Ces algorithmes fournissent de bonnes estimations du paramètre spatial β, alors que l'algorithme en champ moyen tend à surestimer ce paramètre. Sur ces exemples, on ne retrouve pas avec les algorithmes EM de type champ moyen, du fait de leur structure de type EM, le problème de biais propre à ICM au niveau des paramètres du bruit.
En termes de segmentation, les performances de l'algorithme en champ simulé sont de qualité équivalente ou supérieure à celles des autres algorithmes. Notamment, il ne paraît pas présenter les mêmes problèmes de surlissage et de convergence lente observés dans certains cas pour l'algorithme en champ moyen. Les segmentations fournies par l'algorithme en champ modal et l'algorithme ICM sont très similaires, ce qui est normal du fait de la proximité de l'étape de segmentation de ICM avec l'étape (NR) de l'algorithme en champ modal. On remarque que dans le cas des images avec bruit simulé, le phénomène de biais se retrouve dans les estimations obtenues par l'algorithme en champ modal, dans une moindre mesure cependant par rapport aux résultats de ICM. Ainsi il apparaît que dans le cas d'images simulées l'algorithme en champ modal présente les meilleures performances alors que dans le cas d'images avec bruit simulé ou d'images réelles (c'est-à-dire les cas où $\mathbf{z}$ n'est pas une réalisation d'un modèle de champ de Markov caché) c'est l'algorithme en champ simulé qui conduit aux segmentations les plus satisfaisantes.
D'autre part, cette étude soulève quelques questions quant à l'interprétation à donner aux estimations du paramètre spatial. Sur les exemples de l'image Damier, Logo et Mickey, on constate que les algorithmes testés fournissent des estimations très différentes. L'algorithme en champ moyen tend à toujours fournir une estimation supérieure à celles des autres méthodes, ce qui explique l'effet de surlissage apparaissant dans les segmentations correspondantes. L'estimation du paramètre spatial par l'algorithme en champ modal est aussi le plus souvent supérieure à celles fournies par ICM et l'algorithme en champ simulé. Il est difficile d'interpréter cette tendance, surtout si l'on s'apercoit qu'une forte différence dans l'estimation de β peut conduire à des segmentations très proches (voir par exemple les figures 4.3 (f) et (h) et le tableau 4.8). Rappelons que le principe des algorithmes EM de type champ moyen consiste à se ramener à un modèle plus simple que le modèle de champ de Markov caché. Ainsi, une explication possible à de telles variations dans l'estimation du

paramètre spatial est que cela correspond à l'estimation de modèles distincts et que β n'a donc pas forcément une interprétation commune pour l'approximation en champ moyen, en champ modal et en champ simulé. En outre, les travaux de Besag (1986) et Forbes et Raftery (1998) sur ICM et les travaux de Zhang (1992) sur l'algorithme en champ moyen suggèrent qu'une estimation précise de β n'est pas une exigeance et que la solution qui consiste à fixer le paramètre dans un intervalle $[\beta_{min},\beta_{max}]$ correct peut conduire à des segmentations tout aussi satisfaisantes. Ceci peut expliquer le fait d'observer des segmentations similaires pour des estimations différentes de β. Dans ce cas, fixer ce paramètre conduirait à des algorithmes moins coûteux. Plus d'expériences sont cependant nécessaires, notamment il faut pouvoir donner une méthode fiable pour déterminer l'intervalle $[\beta_{min},\beta_{max}]$ pertinent dans lequel choisir β. Ces dernières remarques seront à considérer ultérieurement afin d'envisager une utilisation optimisée des algorithmes EM de type champ moyen.

Chapitre 5

Convergence des algorithmes de type MCEM

L'algorithme MCEM a été introduit par Wei et Tanner (1990) pour répondre à des situations où, à l'étape (E) de EM, le calcul de $Q(\Psi \mid \Psi^{(q-1)})$ n'est pas faisable. La méthode, basée sur des simulations MC, s'applique à tout problème avec données manquantes (voir partie 2.2.5.2). Dans ce chapitre, nous présentons des résultats théoriques sur l'algorithme MCEM pour l'estimation des paramètres d'un modèle de champ de Markov caché.
La première analyse sérieuse de la convergence de l'algorithme MCEM est due à Biscarat (1994). L'auteur établit, sous certaines conditions de régularité, la convergence presque sûre de toute suite dans $I\!R^q$ générée par une classe particulière d'algorithmes stochastiques. Biscarat (1994) montre ensuite que pour des modèles de mélanges finis indépendants l'algorithme MCEM appartient à cette classe et que dans ce contexte, sous certaines conditions de régularité, la suite des estimateurs générée par MCEM converge presque sûrement vers un maximum local de la vraisemblance des paramètres quelle que soit l'initialisation. Cependant, nous ne savons pas si ces résultats restent valables dans le cas d'un modèle de champ de Markov caché.
Chan et Ledolter (1995) ont étudié la convergence de MCEM pour des modèles tels que la log-vraisemblance complète des paramètres peut s'écrire sous la forme $l(\mathbf{W},\Psi)$ où $\mathbf{W}$ est une fonction mesurable de $(\mathbf{Z},\mathbf{Y})$ et l est une fonction linéraire en $\mathbf{W}$. Ces modèles sont regroupés sous le nom de *modèles exponentiels courbes*. Chan et Ledolter (1995) montrent que sous certaines conditions de régularité pour tout maximum isolé Ψ^* de $L(\Psi,\mathbf{y})$ il existe un voisinage V^* de Ψ^* tel que si l'algorithme MCEM est initialisé dans ce voisinage alors quel que soit $\epsilon > 0$ il existe $q(\epsilon)$ tel que

$$\lim_{T\to\infty} P(|\Psi_T^{(q)} - \Psi^*| < \epsilon, \text{ pour } q \leq q(\epsilon)) = 1,$$

où T est le nombre commun de réalisations de la distribution conditionnelle des données manquantes utilisées à chaque itération de la procédure MCEM. Cela signifie que si l'initialisation de la procédure et T sont correctement choisis alors avec une grande probabilité la suite des estimateurs générée par MCEM se rapproche d'un maximum de la vraisemblance

des paramètres. Dans des travaux récents, Fort (2001) établit la convergence de MCEM pour la même famille de modèles exponentiels courbes sous des conditions plus faibles que Chan et Ledolter (1995).
Les résultats de Chan et Ledolter (1995) et Fort (2001) s'appliquent au cas d'un modèle de champ de Markov caché dès lors que la fonction énergie de la distribution de Gibbs permet de vérifier l'hypothèse de linéarité de la log-vraisemblance complète des paramètres définissant les modèles exponentiels courbes. Cependant, la complexité du modèle de champ de Markov caché implique d'approximer la distribution marginale des données cachées, pour la mise en œuvre de MCEM. Ainsi, on n'est plus assuré des bonnes propriétés de convergence de la procédure.
Lorsque le paramètre spatial β est connu et que seul le paramètre de bruit θ doit être estimé, cette approximation n'est pas nécessaire. Dans ce contexte, Comer et Delp (2000) ont considéré une version stochastique de EM, l'algorithme EM/MPM, qui consiste à utiliser, dans les formules de EM, des estimations des probabilités markoviennes conditionnelles obtenues par un échantillonneur de Gibbs. Les auteurs ont établi que presque sûrement l'estimateur du paramètre θ obtenu par EM/MPM peut être rendu arbitrairement proche de l'estimateur obtenu par EM, si un nombre suffisant d'itérations de l'algorithme et de l'échantillonneur de Gibbs sont réalisées. Cela signifie que si l'estimateur obtenu par EM est proche de la vraie valeur du paramètre, alors l'estimateur obtenu par EM/MPM aura la même propriété. Ceci fournit un premier résultat théorique sur la convergence de MCEM pour champ de Markov caché, puisque nous mettons en évidence dans ce chapitre que EM/MPM est équivalent à MCEM.
Toujours dans le cas où le paramètre spatial est connu, nous présentons dans ce chapitre de nouveaux résultats plus généraux. Nous donnons des conditions suffisantes pour assurer que presque sûrement la suite des log-vraisemblances générée par MCEM ne décroît pas et converge. Puis nous adoptons un point de vue sur l'algorithme MCEM différent de celui de Comer et Delp (2000) et qui consiste à rapprocher MCEM des algorithmes EM *Généralisés* (Dempster, Laird, et Rubin 1977). Cette approche nous permet de montrer que sous certaines conditions de régularité l'estimateur obtenu par MCEM converge presque sûrement vers un maximum local de la log-vraisemblance des paramètres. Ces résultats s'étendent à toute variante de MCEM utilisant des simulations MC pour approximer les probabilités markoviennes conditionnelles, dès lors que les approximations convergent presque sûrement vers les valeurs exactes. En particulier, nous établissons la convergence de l'algorithme *en champ simulé généralisé*, dérivé de l'algorithme en champ simulé présenté dans le chapitre 4. Ces résultats sont également présentés dans Peyrard (2001).
La première partie de ce chapitre est consacrée à la description de l'algorithme MCEM et des variantes citées ci-dessus, que nous appelons *algorithmes de type* MCEM. Les résultats théoriques établis sont présentés dans la partie 5.2 et le comportement de l'algorithme MCEM et de l'algorithme en champ simulé généralisé sont illustrés dans une dernière partie.

5.1 MCEM et algorithmes de type MCEM

L'algorithme MCEM a été décrit dans la partie 2.2.5.2, dont nous ne redéfinissons pas ici les notations. Nous considérons dans ce chapitre que la chaîne de Markov générée à l'étape (MC) de l'itération q est une chaîne ergodique de mesure invariante $P_G(\mathbf{z} \mid \mathbf{y},\theta^{(q-1)})$. C'est par exemple le cas si l'étape (MC) est réalisée en utilisant l'échantillonneur de Gibbs. Nous nous plaçons dans le cas où le paramètre spatial β est connu. Dans le but d'éclairer la structure de l'algorithme MCEM et le lien entre MCEM et d'autres versions stochastiques de EM, nous comparons l'estimateur $\theta^{(q)}_{\mathrm{EM}}$ obtenu après une itération de EM avec $\theta^{(q)}_{T_q}$, l'estimateur obtenu après une itération de MCEM, lorsque les deux procédures sont initialisées avec la même valeur courante $\theta^{(q-1)}_{T_{q-1}}$. Considérons le vecteur des probabilités markoviennes conditionnelles,

$$\begin{aligned}\mathbf{p}^{(q)}_G &= \{P_G(Z_i = e_k \mid \mathbf{y},\theta^{(q-1)}_{T_{q-1}}),\ i \in S,\ k \in \{1,\ldots,K\}\}\\ &= \{p^{(q)}_{Gik},\ i \in S,\ k \in \{1,\ldots,K\}\}.\end{aligned}$$

La fonction Q_θ définie par (2.31) se réécrit en fonction de $\mathbf{p}^{(q)}_G$

$$Q_\theta(\theta \mid \theta^{(q-1)}_{T_{q-1}}) = \sum_{i\in S}\sum_{k=1}^{K} p^{(q)}_{Gik} \log g(y_i \mid Z_i = e_k,\, \theta).$$

Cette fonction dépend de $\theta^{(q-1)}_{T_{q-1}}$ uniquement à travers $\mathbf{p}^{(q)}_G$. Nous la noterons dans la suite de ce chapitre $R(\theta \mid \mathbf{p}^{(q)}_G)$. Soit $\mathcal{P}$ l'ensemble $\{\mathbf{p} \in [0,1]^{n\times K} / \forall i \in S, \sum_{k=1}^{K} p_{ik} = 1\}$, nous noterons $\mathbf{\Phi}$ la fonction de $\mathcal{P}$, à valeurs dans Θ, définie par

$$\forall \mathbf{p} \in \mathcal{P},\ \mathbf{\Phi}(\mathbf{p}) = \arg\max_\theta R(\theta \mid \mathbf{p}).$$

Ainsi, une itération de EM, lorsque la valeur courante du paramètre est $\theta^{(q-1)}_{T_{q-1}}$, est équivalente à

$$\begin{aligned}\theta^{(q)}_{\mathrm{EM}} &= \arg\max_\theta Q_\theta(\theta \mid \theta^{(q-1)}_{T_{q-1}})\\ &= \arg\max_\theta R(\theta \mid \mathbf{p}^{(q)}_G)\\ &= \mathbf{\Phi}(\mathbf{p}^{(q)}_G).\end{aligned} \tag{5.1}$$

Soit $\mathbf{d}$ la fonction indicatrice sur $\{0,1\}^{n\times K}$ d'élément générique

$$d_{ik}(\mathbf{z}) = \begin{cases} 1, \text{ si } z_i = e_k \\ 0, \text{ si } z_i \neq e_k \end{cases}.$$

Le vecteur des fréquences markoviennes conditionnelles empiriques est

$$\mathbf{f}^{(q)} = \{f^{(q)}_{ik},\ i \in S,\ k \in \{1,\ldots,K\}\},$$

avec

$$f_{ik}^{(q)} = \frac{1}{T_q} \sum_{t=1}^{T_q} d_{ik}(\mathbf{z}(q,t)). \tag{5.2}$$

Pour l'algorithme MCEM, la fonction à maximiser à l'étape (M) est

$$Q_{T_q}(\theta \mid \theta_{T_{q-1}}^{(q-1)}) = \frac{1}{T_q} \sum_{t=1}^{T_q} \log g(\mathbf{y} \mid \mathbf{z}(q,t),\, \theta),$$

soit encore

$$\begin{aligned} Q_{T_q}(\theta \mid \theta_{T_{q-1}}^{(q-1)}) &= \frac{1}{T_q} \sum_{t=1}^{T_q} \sum_{i \in S} \sum_{k=1}^{K} d_{ik}(\mathbf{z}(q,t)) \log g(y_i \mid Z_i = e_k,\, \theta) \\ &= \sum_{i \in S} \sum_{k=1}^{K} f_{ik}^{(q)} \log g(y_i \mid Z_i = e_k,\, \theta) \\ &= R(\theta \mid \mathbf{f}^{(q)}). \end{aligned}$$

Finalement,

$$\begin{aligned} \theta_{T_q}^{(q)} &= \arg\max_{\theta} Q_{T_q}(\theta \mid \theta_{T_{q-1}}^{(q-1)}) \\ &= \mathbf{\Phi}(\mathbf{f}^{(q)}). \end{aligned} \tag{5.3}$$

Ainsi, la mise à jour du paramètre θ par MCEM utilise les mêmes formules que celles de l'étape (E) de EM, les probabilités markoviennes conditionnelles étant remplacées par les fréquences markoviennes empiriques. Comer et Delp (2000) ont proposé l'algorithme EM/MPM qui consiste dans un premier temps à générer, par l'échantillonneur de Gibbs, une suite de réalisations de la distribution markovienne conditionnelle pour la valeur courante des paramètres, puis à approximer $\mathbf{p}_G^{(q)}$ d'après la formule (5.2). Ensuite les fréquences empiriques $\mathbf{f}^{(q)}$ sont substituées à $\mathbf{p}_G^{(q)}$ dans l'équation de mise à jour du paramètre θ de l'algorithme EM. L'algorithme MCEM, dans le cas où le paramètre spatial est connu et si l'étape (MC) est réalisée en utilisant l'échantillonneur de Gibbs, est donc équivalent à l'algorithme EM/MPM de Comer et Delp (2000).
L'idée directrice de l'algorithme MCEM pour champ de Markov caché est d'utiliser des simulations MC pour obtenir une approximation $\hat{\mathbf{p}}^{(q)}$ des probabilités conditionnelles $\mathbf{p}_G^{(q)}$, non calculables de manière directe, et d'utiliser ensuite ces approximations dans les formules de EM. Pour cet algorithme, il s'agit des fréquences conditionnelles empiriques, mais d'autres choix peuvent être considérés et la méthode se généralise. Nous désignerons par *algorithme de type* MCEM une version de EM obtenue par substitution de $\mathbf{p}_G^{(q)}$ à partir de simulations MC. Par exemple, le vecteur $\mathbf{p}_G^{(q)}$ peut être approximé dans l'étape (MC) par $\hat{\mathbf{p}}^{(q)}$ donné par

$$\forall\, i \in S, \forall\, k \in \{1,\ldots,K\},\ \hat{p}_{ik}^{(q)} = \frac{1}{T_q} \sum_{t=1}^{T_q} P_G(Z_i = e_k \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}(q,t), \mathbf{y}, \theta_{T_{q-1}}^{(q-1)}). \tag{5.4}$$

Ce choix est d'une nature différente de l'approximation (5.2) utilisée dans MCEM. Si $T_q = 1$, $\mathbf{f}^{(q)}$ se réduit à une distribution de Dirac, donnant probabilité 1 à la réalisation $\mathbf{z}(q,1)$, ce qui n'est pas le cas de $\hat{\mathbf{p}}^{(q)}$. Plus précisement, l'approximation de $p_{G\,ik}^{(q)}$ définie par (5.4) est obtenue en approximant l'influence du voisinage du pixel i par la valeur simulée $\mathbf{z}_{N(i)}(q,1)$. Nous retrouvons ici l'idée conductrice des approximations de type champ moyen pour EM décrites dans le chapitre 4. Si l'étape (MC) utilise l'échantillonneur de Gibbs pour obtenir l'approximation (5.4), le cas $T_q = 1$ correspond à l'algorithme en champ simulé décrit au chapitre 4, lorsque le paramètre spatial n'est pas estimé. L'extension de cette procédure au cas $T_q \geq 1$ sera désignée par *algorithme en champ simulé généralisé*.
Des résultats de convergence sur l'algorithme MCEM et l'algorithme en champ simulé généralisé sont présentés dans la partie suivante.

5.2 Résultats théoriques sur les algorithmes de type MCEM

Dans cette partie, deux aspects de la convergence des algorithmes de type MCEM sont étudiés. Dans un premier temps, nous considérons la perturbation aléatoire $\theta_{T_q}^{(q)} - \theta_{\mathrm{EM}}^{(q)}$ ajoutée à chaque itération q à l'estimateur EM par la procédure MCEM, et due à l'approximation stochastique de la fonction Q. Il s'agit d'une erreur locale au sens où elle décrit l'erreur après une seule itération. Si cette différence est petite, une itération de MCEM se comporte comme une itération de EM. Nous donnons des conditions suffisantes pour que la log-vraisemblance des paramètres du modèle, $LL(\theta,\mathbf{y}) = \log P(\mathbf{y} \mid \theta)$ augmente presque sûrement à chaque itération. Cette propriété de croissance de la log-vraisemblance est une propriété qui caractérise les procédures EM *Généralisées* (GEM) (Dempster, Laird, et Rubin 1977) et qui est la clé de leurs bonnes propriétés théoriques. Nous utilisons alors la théorie sur les algorithmes GEM pour établir des résultats de convergence sur l'estimateur MCEM, et plus généralement pour l'estimateur d'un algorithme de type MCEM.

5.2.1 Convergence de la vraisemblance

Théorème 1 *Si les propriétés suivantes sont satisfaites*

(i) $R(. \mid \mathbf{p})$ *est une fonction unimodale pour tout* $\mathbf{p}$ *dans* $\mathcal{P}$ *;*

(ii) $Q(. \mid \theta')$ *est une fonction continue sur* Θ*, pour tout* $\theta' \in \Theta$ *;*

(iii) $\mathbf{\Phi}(.)$ *est une fonction continue sur* $\mathcal{P}$ *;*

(iv) *si* $P_G(\mathbf{z} \mid \mathbf{y},\theta) = P_G(\mathbf{z} \mid \mathbf{y},\theta')$ *presque partout, alors* $\theta = \theta'$ *;*

alors,

$$P(\forall q \in I\!N, \exists T^* \in I\!N /\ \forall T_q > T^*, \quad LL(\theta_{T_q}^{(q)},\mathbf{y}) - LL(\theta_{T_{q-1}}^{(q-1)},\mathbf{y}) \geq 0) = 1.$$

Si de plus,

(v) *la log-vraisemblance* $LL(\theta,\mathbf{y})$ *est majorée sur* Θ *;*

alors avec probabilité 1 il existe une suite $\{T_q\}_{q \in I\!N}$ *telle que la suite des log-vraisemblances* $\{LL(\theta^{(q)}_{T_q},\mathbf{y})\}_{q \in I\!N}$ *générée par l'algorithme MCEM converge.*

Démonstration : la propriété (i) implique que la fonction $\mathbf{\Phi}(.)$ est bien définie, ainsi que les étapes (M) de EM et de MCEM. L'ergodicité de la chaîne $\{\mathbf{z}(q,t)\}_{t \in I\!N}$ générée à l'itération q entraîne que pour toute fonction h dans V^n,

$$P\left(\lim_{T_q \to \infty} \frac{1}{T_q} \sum_{t=1}^{T_q} h(\mathbf{z}(q,t)) = I\!E_{P_G}[h(\mathbf{Z}) \mid \mathbf{y},\theta^{(q-1)}_{T_{q-1}}]\right) = 1. \tag{5.5}$$

(voir par exemple le théorème C de Geman et Geman (1984) pour la démonstration dans le cas d'une chaîne générée par l'échantillonneur de Gibbs.) La propriété (5.5) appliquée à $h = \mathbf{d}$ conduit à

$$P(\lim_{T_q \to \infty} \mathbf{f}^{(q)} = I\!E_{P_G}[\mathbf{d}(\mathbf{Z}) \mid \mathbf{y},\theta^{(q-1)}_{T_{q-1}}]) = 1.$$

Comme

$$I\!E_{P_G}[d_{ik}(\mathbf{Z}) \mid \mathbf{y},\theta^{(q-1)}_{T_{q-1}}] = p^{(q)}_{G\,ik}\,,$$

nous avons

$$P(\lim_{T_q \to \infty} \mathbf{f}^{(q)} = \mathbf{p}^{(q)}_G) = 1, \tag{5.6}$$

et puisque $\mathbf{\Phi}$ est une fonction continue,

$$P(\lim_{T_q \to \infty} \mathbf{\Phi}(\mathbf{f}^{(q)}) = \mathbf{\Phi}(\mathbf{p}^{(q)}_G)) = 1.$$

D'après les égalités (5.1) et (5.3), cela signifie qu'avec probabilité 1 l'estimateur MCEM converge vers l'estimateur résultant d'une itération de EM initialisée par $\theta^{(q-1)} = \theta^{(q-1)}_{T_{q-1}}$,

$$P(\lim_{T_q \to \infty} \theta^{(q)}_{T_q} = \theta^{(q)}_{\mathrm{EM}}) = 1. \tag{5.7}$$

Considérons maintenant l'évolution de $LL(\theta^{(q)}_{T_q},\mathbf{y})$ après une itération de MCEM. La propriété de continuité de $Q(.\mid \theta^{(q-1)}_{T_{q-1}})$ implique

$$P(\lim_{T_q \to \infty} Q(\theta^{(q)}_{T_q} \mid \theta^{(q-1)}_{T_{q-1}}) = Q(\theta^{(q)}_{\mathrm{EM}} \mid \theta^{(q-1)}_{T_{q-1}})) = 1. \tag{5.8}$$

Considérons l'évènement

$$A_q = \{\forall \epsilon > 0\ \exists T^* \,/\, \forall T_q > T^*,\ |Q(\theta^{(q)}_{T_q} \mid \theta^{(q-1)}_{T_{q-1}}) - Q(\theta^{(q)}_{\mathrm{EM}} \mid \theta^{(q-1)}_{T_{q-1}})| < \epsilon\}\,.$$

D'après la définition de $\theta^{(q)}_{\mathrm{EM}}$, la différence $\alpha = Q(\theta^{(q)}_{\mathrm{EM}} \mid \theta^{(q-1)}_{T_{q-1}}) - Q(\theta^{(q-1)}_{T_{q-1}} \mid \theta^{(q-1)}_{T_{q-1}})$ est positive. Les deux cas $\alpha > 0$ et $\alpha = 0$ doivent être traités séparément. Dans un premier temps, si $\alpha > 0$ il existe $\epsilon_0 > 0$ tel que $\alpha - \epsilon_0 \geq 0$. Si l'évènement A_q est réalisé alors,

$$\exists T^0 \,/\, \forall T_q > T^0,\ Q(\theta^{(q)}_{T_q} \mid \theta^{(q-1)}_{T_{q-1}}) - Q(\theta^{(q)}_{\mathrm{EM}} \mid \theta^{(q-1)}_{T_{q-1}}) > -\epsilon_0,$$

et pour cette valeur T^0, nous avons

$$\forall T_q > T^0,\ Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{\mathrm{EM}}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) + \alpha > \alpha - \epsilon_0.$$

Alors,

$$\forall T_q > T^0,\ Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{T_{q-1}}^{(q-1)} \mid \theta_{T_{q-1}}^{(q-1)}) \geq 0.$$

La log-vraisemblance du modèle peut être réécrite (voir partie 2.2.1),

$$LL(\theta,\mathbf{y}) = Q(\theta \mid \theta_{T_{q-1}}^{(q-1)}) - \mathcal{H}(\theta \mid \theta_{T_{q-1}}^{(q-1)}),$$

où

$$\mathcal{H}(\theta \mid \theta_{T_{q-1}}^{(q-1)}) = I\!\!E_{P_G}[\log P_G(\mathbf{Z} \mid \mathbf{y},\theta)|\mathbf{y},\theta_{T_{q-1}}^{(q-1)}].$$

Puisque $\mathcal{H}$ est une fonction de θ qui admet un maximum global pour $\theta = \theta_{T_{q-1}}^{(q-1)}$ (voir Dempster, Laird, et Rubin 1977), il vient

$$\forall T_q > T^0,\ LL(\theta_{T_q}^{(q)},\mathbf{y}) - LL(\theta_{T_{q-1}}^{(q-1)},\mathbf{y}) \geq 0. \tag{5.9}$$

Considérons maintenant le cas $\alpha = 0$. La condition (i) implique que $Q(. \mid \theta_{T_{q-1}}^{(q-1)})$ est une fonction unimodale. Donc, si $\alpha = 0$, les deux estimateurs $\theta_{T_{q-1}}^{(q-1)}$ et $\theta_{\mathrm{EM}}^{(q)}$ sont égaux. Considérons la différence $\delta \geq 0$ définie par

$$\delta = \mathcal{H}(\theta_{T_{q-1}}^{(q-1)} \mid \theta_{T_{q-1}}^{(q-1)}) - \mathcal{H}(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}).$$

Si $\delta > 0$ alors, pour un $\epsilon_1 > 0$ donné tel que $\delta - \epsilon_1 \geq 0$, A_q entraîne que

$$\exists T^1 \,/\, \forall T_q > T^1,\ Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{\mathrm{EM}}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) > -\epsilon_1,$$

et puisque $\theta_{T_{q-1}}^{(q-1)} = \theta_{\mathrm{EM}}^{(q)}$,

$$\forall T_q > T^1,\ Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{T_{q-1}}^{(q-1)} \mid \theta_{T_{q-1}}^{(q-1)}) > -\epsilon_1.$$

Ainsi,

$$\forall T_q > T^1,\ LL(\theta_{T_q}^{(q)},\mathbf{y}) - LL(\theta_{T_{q-1}}^{(q-1)},\mathbf{y}) > -\epsilon_1 + \delta \geq 0. \tag{5.10}$$

Dempster, Laird, et Rubin (1977) ont démontré que la différence δ est égale à zero si et seulement si $P_G(\mathbf{z} \mid \mathbf{y},\theta_{T_q}^{(q)}) = P_G(\mathbf{z} \mid \mathbf{y},\theta_{T_{q-1}}^{(q-1)})$ presque partout. Nous déduisons de l'hypothèse (iv) que dans ce cas $\theta_{T_q}^{(q)} = \theta_{T_{q-1}}^{(q-1)}$, et donc $LL(\theta_{T_q}^{(q)},\mathbf{y}) - LL(\theta_{T_{q-1}}^{(q-1)},\mathbf{y}) = 0$.
Finalement, en combinant les inégalités (5.9) et (5.10), quel que soit $\alpha \geq 0$,

$$\exists T^* = \sup(T^0,T^1) \,/\, \forall T_q > T^*, \quad LL(\theta_{T_q}^{(q)},\mathbf{y}) - LL(\theta_{T_{q-1}}^{(q-1)},\mathbf{y}) \geq 0, \tag{5.11}$$

avec T^* fini, comme T^0 et T^1. L'évènement A_q est inclus dans l'évènement B_q défini par (5.11), et $P(A_q) = 1$ d'après (5.8), de sorte que,

$$\forall q \in \mathbb{N},\ P(B_q) \ = \ 1,$$

et puisque l'intersection est dénombrable,

$$P(\bigcap_q B_q) \ = \ 1.$$

Cela signifie que

$$P(\forall q \in \mathbb{N}, \exists T^* \in \mathbb{N} /\ \forall T_q > T^*, \quad LL(\theta_{T_q}^{(q)},\mathbf{y}) - LL(\theta_{T_{q-1}}^{(q-1)},\mathbf{y}) \geq 0) = 1.$$

Si de plus la log-vraisemblance $LL(\theta,\mathbf{y})$ est majorée, ce résultat entraîne que presque sûrement il existe une suite $\{T_q\}_{q\in\mathbb{N}}$ telle que la suite $\{LL(\theta_{T_q}^{(q)},\mathbf{y})\}_{q\in\mathbb{N}}$ converge.

□

Remarques : notons d'abord que le résultat se généralise à tout algorithme de type MCEM, dès lors que l'approximation $\hat{\mathbf{p}}^{(q)}$ satisfait

$$P(\lim_{T_q\to\infty} \hat{\mathbf{p}}^{(q)} = \mathbf{p}_G^{(q)}) = 1. \tag{5.12}$$

Pour une telle procédure, la preuve du théorème 1 se déroule de manière identique, la propriété (5.6) étant remplacée par (5.12). En particulier, la propriété (5.12) est satisfaite par l'algorithme en champ simulé généralisé : la distribution $\hat{\mathbf{p}}^{(q)}$ définie par l'égalité (5.4) converge vers

$$\mathbb{E}_{P_G}[P_G(Z_i = e_k \mid \mathbf{Z}_{N(i)},\mathbf{y},\theta_{T_{(q-1)}}^{(q-1)}) \mid \mathbf{y},\theta_{T_{(q-1)}}^{(q-1)}] \ = \ P_G(Z_i = e_k \mid \mathbf{y},\theta_{T_{(q-1)}}^{(q-1)}),$$

d'après la propriété d'ergodicité (5.5).

Dans le théorème 1, l'hypothèse (v) peut être remplacée par l'hypothèse plus forte :
(v') *pour tout i dans S et pour tout k dans $\{1,\ldots,K\}$, la distribution $g(y_i \mid Z_i = e_k,\theta)$ est strictement positive et bornée uniformément en θ.*
Dans ce cas, les hypothèses (i), (ii), (iv) et (v') sont suffisantes pour assurer que presque sûrement la suite des log-vraisemblances ne décroît pas et converge. La condition (iii) n'est plus nécessaire. En effet, nous avons le lemme suivant :

Lemme 1 *Si pour tout $i \in S$ et pour tout $K \in \{1,\ldots,K\}$, la distribution $g(y_i \mid Z_i = e_k,\theta)$ est strictement positive et si elle est bornée uniformément en $\theta \in \Theta$, alors*

$$P(\lim_{T_q\to\infty} Q_{T_q}(\theta \mid \theta_{T_{q-1}}^{(q-1)}) = Q(\theta \mid \theta_{T_{q-1}}^{(q-1)}), \text{ uniformément en } \theta) = 1.$$

Ceci assure que la propriété (5.7) est encore vraie, puisque la convergence uniforme d'une suite $\{a_n\}_{n\in I\!N}$ vers une fonction a implique la convergence de $\{\arg\max_x a_n(x)\}_{n\in I\!N}$ vers $\arg\max_x a(x)$. La suite de la démonstration du théorème 1 reste identique.

Démonstration du lemme 1 : l'hypothèse (v') implique qu'il existe une constante strictement positive M, telle que,

$$\forall i \in S,\ \forall k \in \{1,\ldots,K\}, \forall \theta \in \Theta, \quad \mid \log g(y_i \mid Z_i = e_k,\, \theta) \mid \leq M.$$

Considérons la différence $D^{(q)}(\theta) = \mid Q_{T_q}(\theta \mid \theta^{(q-1)}_{T_{q-1}}) - Q(\theta \mid \theta^{(q-1)}_{T_{q-1}}) \mid$.

$$D^{(q)}(\theta) \;=\; \mid \frac{1}{T_q}\sum_{t=1}^{T_q} \log P_G(\mathbf{y} \mid \mathbf{Z} = \mathbf{z}(q,t),\theta) - I\!E_{P_G}[\log P_G(\mathbf{y} \mid \mathbf{Z},\theta) \mid \mathbf{y},\theta^{(q-1)}_{T_{q-1}}] \mid.$$

Le terme $D^{(q)}(\theta)$ peut être majoré uniformément en θ. En effet, si l'on note $I\!E_{\mathbf{f}^{(q)}}$ l'espérance selon la distribution empirique définie par $\mathbf{f}^{(q)}$,

$$\begin{aligned}
\forall \theta \in \Theta, \quad D^{(q)}(\theta) &= \mid I\!E_{\mathbf{f}^{(q)}}[\log P_G(\mathbf{y} \mid \mathbf{Z},\theta)] - I\!E_{P_G}[\log P_G(\mathbf{y} \mid \mathbf{Z},\theta) \mid \mathbf{y},\theta^{(q-1)}_{T_{q-1}}] \\
&\leq M \sum_{i\in S}\sum_{k=1}^{K} \mid f^{(q)}_{ik} - p^{(q)}_{G\,ik} \mid.
\end{aligned}$$

Puisque d'après (5.6) les fréquences empiriques $\mathbf{f}^{(q)}$ convergent presque sûrement vers les fréquences théoriques $\mathbf{p}^{(q)}_G$,

$$P(\lim_{T_q\to\infty} \sup_\theta D^{(q)}(\theta) = 0) \;\geq\; P(\forall i \in S,\ \forall k \in \{1,\ldots,K\},\ \lim_{T_q\to\infty} \mid p^{(q)}_{G\,ik} - f^{(q)}_{ik} \mid = 0) = 1,$$

ce qui démontre la convergence presque sûre annoncée.

□

En analyse d'images, les données observées y_i sont souvent supposées issues d'une densité gaussienne conditionnellement à z_i. Dans le cas d'observations unidimensionnelles, si l'une des données y_i est égale à la moyenne m_k de l'une des composantes, et si la variance σ_k^2 tend vers zéro, alors la densité $g(y_i \mid Z_i = e_k,\, \theta)$ est infinie. En contraignant l'espace d'états Θ des paramètres, par exemple en supposant pour tout k dans $\{1,\ldots,K\}$ $\sigma_k \in [\underline{\sigma},\overline{\sigma}]$ avec $\underline{\sigma} > 0$, la situation décrite est évitée et l'hypothèse (v') est vérifiée.

5.2.2 Convergence de l'estimateur

La partie précédente établit des conditions suffisantes pour la convergence presque sûre de la suite des log-vraisemblances vers une limite L^*, pour les procédures de type MCEM.

Nous considérons maintenant la question de la nature de cette limite et de la convergence de la suite $\{\theta_{T_q}^{(q)}\}_{q\in I\!N}$ vers un point stationnaire ou un maximum local de $LL(\theta,\mathbf{y})$. Les résultats que nous obtenons utilisent l'étude des propriétés de convergence des algorithmes GEM par Wu (1983). Dempster, Laird, et Rubin (1977) définissent un algorithme GEM comme un algorithme moins contraignant que EM, pour lequel l'étape (M) doit seulement vérifier

$$Q(\theta^{(q)} \mid \theta^{(q-1)}) \quad \geq \quad Q(\theta^{(q-1)} \mid \theta^{(q-1)}).$$

Pour toute suite $\theta^{(q)}$ d'un algorithme GEM, nous avons

$$LL(\theta^{(q)},\mathbf{y}) \geq LL(\theta^{(q-1)},\mathbf{y}),$$

ce qui assure que l'algorithme hérite de la plupart des propriétés de convergence de EM. Dans la suite nous adoptons les notations de Wu (1983). Soit $\mathcal{S}$ l'ensemble des points stationnaires et $\mathcal{M}$ l'ensemble des maxima locaux de $LL(\theta,\mathbf{y})$. Une application directe du théorème 1 de Wu (1983) conduit au théorème suivant :

Théorème 2 *Supposons que les hypothèses suivantes sont vraies presque sûrement,*

(i) $\forall q \in I\!N, \exists T^* \,/\, \forall T_q > T^*,\ LL(\theta_{T_q}^{(q)},\mathbf{y}) - LL(\theta_{T_{q-1}}^{(q-1)},\mathbf{y}) \geq 0$ *;*

(ii) $\forall q \in I\!N, \forall \theta_{T_{q-1}}^{(q-1)} \notin \mathcal{S}$ *(resp.* $\mathcal{M}$*),* $\exists T^{**} \,/\, \forall T_q > T^{**},\ LL(\theta_{T_q}^{(q)},\mathbf{y}) - LL(\theta_{T_{q-1}}^{(q-1)},\mathbf{y}) > 0$ *;*

(iii) la fonction $\arg\max_{\theta} Q_{T_q}(\theta \mid \theta')$ *est fermée sur le complémentaire de* $\mathcal{S}$ (resp. $\mathcal{M}$) *;*

(iv) la suite $\{LL(\theta_{T_q}^{(q)},\mathbf{y})\}_{q\in I\!N}$ *est majorée pour tout* $\theta_{T_0}^{(0)}$ *dans* Θ.

Si à chaque itération d'un algorithme de type MCEM, le nombre T_q *est choisi de sorte que* $T_q > \max(T^*,T^{**})$*, alors presque sûrement tous les points limites de* $\{\theta_{T_q}^{(q)}\}_{q\in I\!N}$ *sont des points stationnaires, (resp. des maxima locaux) de* $LL(\theta,\mathbf{y})$*, et presque sûrement la suite* $\{LL(\theta_{T_q}^{(q)},\mathbf{y})\}_{q\in I\!N}$ *est monotone et converge vers* $L^* = LL(\theta^*,\mathbf{y})$ *pour un* θ^* *dans* $\mathcal{S}$ *(resp.* $\mathcal{M}$*).*

Nous avons donné avec le théorème 1 des conditions qui assurent que l'hypothèse (i) est satisfaite presque sûrement. Cette hypothèse signifie qu'avec probabilité 1 la suite des log-vraisemblances générée par un algorithme de type MCEM se comporte comme au cours d'un algorithme GEM. Ceci est d'un grand intérêt puisque la croissance de la log-vraisemblance est l'hypothèse principale pour obtenir les résultats de convergence de l'estimateur obtenu par un algorithme GEM. Ainsi, dès que la condition (i) du théorème 2 est vérifiée presque sûrement, l'estimateur fourni par un algorithme de type MCEM possède avec probabilité 1 les mêmes propriétés que l'estimateur obtenu par un algorithme GEM. Le théorème 2 en donne un exemple. Des conditions pour la convergence ponctuelle presque sûre peuvent aussi se déduire des résultats théoriques sur les procédures GEM (voir Wu 1983 pour les propriétés de convergence de l'estimateur fournit par un algorithme GEM). Nous ne détaillons pas les propriétés correspondantes, il s'agirait d'une réécriture des conditions nécessaires établies pour les algorithmes GEM, celles-ci devant être vérifiées presque sûrement pour un algorithme de type MCEM.

Ces résultats viennent compléter les résultats établis par Comer et Delp (2000) sur la convergence de l'algorithme MCEM pour modèle de champ de Markov caché dans le cas où le paramètre spatial est connu. Dans leur étude, Comer et Delp (2000) ne remarquent pas le fait que MCEM peut être rapproché des procédures GEM et en conséquence ils ne s'appuient pas sur les résultats théoriques existants pour ces algorithmes. Pour plus de clarté, notons $\mathbf{\Phi}_q$ la fonction correspondant à q compositions de la fonction $\mathbf{\Phi}$. Leur approche consiste à comparer l'estimateur résultant de q itérations de MCEM, $\theta_{T_q}^{(q)}$, avec l'estimateur obtenu après q itérations de EM, $\mathbf{\Phi}_q(\mathbf{p}_G^{(0)})$, lorsque les deux procédures sont initialisées avec la même valeur. Leurs résultats garantissent l'existence d'un entier N et d'une suite $\{T_q\}_{1\leq q\leq N}$ tels que $\theta_{T_q}^{(q)}$ peut être rendu arbitrairement proche de l'estimateur EM de θ avec probabilité 1, lorsque les deux procédures sont initialisées avec la même valeur. Ceci est en fait un résultat asymptotique puisque le nombre T_q d'itérations nécessaires peut être infini pour tout $1 \leq q \leq N$. Le résultat laisse donc espérer au mieux une convergence en probabilité de l'estimateur MCEM, alors que la combinaison des théorèmes 1 et 2 donne des conditions suffisantes pour la convergence presque sûre. La principale différence entre notre approche et celle de Comer et Delp (2000) réside dans le niveau de comparaison avec l'algorithme EM. La preuve du théorème 1 est fondée sur le contrôle de l'erreur locale $\mathbf{\Phi}(\mathbf{f}^{(q)})-\mathbf{\Phi}(\mathbf{p}_G^{(q)}) = \theta_{T_q}^{(q)}-\theta_{\mathrm{EM}}^{(q)}$, et non pas sur l'étude de l'erreur globale $\theta_{T_q}^{(q)} - \mathbf{\Phi}_q(\mathbf{p}_G^{(0)})$ comme dans Comer et Delp (2000).

5.3 Illustrations

Dans cette partie le comportement des algorithmes MCEM et en champ simulé généralisé est illustré sur des images dégradées par un bruit blanc gaussien comme décrit dans la partie 4.9. Le modèle supposé pour les données cachées $\mathbf{Z}$ est un modèle de Potts simple dont nous rappelons la définition

$$P_G(\mathbf{z} \mid \beta) \quad = \quad W(\beta)^{-1} \exp(\beta \sum_{i\sim j} z_i^t z_j).$$

L'étape (MC) est réalisée par l'échantillonneur de Gibbs dans les deux algorithmes. La fonction $R(\theta \mid \mathbf{p}_G^{(q)})$ à maximiser à l'étape (M) pour mettre à jour l'estimation du paramètre $\theta = \{(m_k,\sigma_k),k = 1,\ldots,K\}$ est donnée par

$$R(\theta \mid \mathbf{p}_G^{(q)}) \quad = \quad \sum_{i\in S}\sum_{k=1}^{K} p_{G\,ik}^{(q)} \log\left[\frac{1}{\sigma_k\sqrt{2\pi}} \exp\{-\frac{(y_i-m_k)^2}{2\sigma_k^2}\}\right].$$

Le paramètre spatial β n'est pas estimé, il est donc fixé à une valeur constante au cours de l'algorithme. Les performances des algorithmes MCEM et en champ simulé généralisé sont étudiées en fonction du nombre T_q d'itérations réalisées au cours de l'échantillonneur de Gibbs. Nous avons utilisé les valeurs suivantes : $T_q = 1$, 10 et 100. Chaque algorithme est initialisé avec la même segmentation obtenue par seuillage des données observées. Le nombre d'itérations est fixé à $N = 100$. Nous n'avons pas observé d'amélioration significative des résultats pour plus d'itérations. Une segmentation est ensuite obtenue par maximisation des fréquences empiriques, c'est à dire pour l'algorithme MCEM

$$\forall i \in S, \quad \hat{z}_i = e_k,$$
$$\text{avec } e_k = \arg\max_{1 \leq l \leq K} \hat{f}_{il}^{(N)}.$$

et pour l'algorithme en champ simulé généralisé

$$\forall i \in S, \quad \hat{z}_i = e_k,$$
$$\text{avec } e_k = \arg\max_{1 \leq l \leq K} \hat{p}_{il}^{(N)}.$$

Les algorithmes sont illustrés sur l'image Damier, l'image Logo et l'image Mickey, déjà présentées dans la partie 4.9. Pour les deux premières images, le paramètre spatial a été fixé à 1. Pour l'image Mickey, les résultats ont été obtenus pour $\beta = 0.7$, le choix d'une valeur plus grande conduisant à de moins bonnes restaurations. Les segmentations obtenues par les deux algorithmes pour les différentes valeurs de T_q sont illustrées sur les figures 5.1, 5.2 et 5.3. Les estimations des paramètres sont reportés dans les tableaux 5.1, 5.2 et 5.3.
Si un trop petit nombre d'itérations de l'échantillonneur de Gibbs est utilisé, les algorithmes n'ont pas toujours convergé, comme le montrent les figures 5.1 (d) et (g) et les figures 5.3 (d) et (g). Les résultats obtenus pour $T_q = 10$ sont satisfaisants pour les deux algorithmes qui présentent des performances semblables en termes d'estimation du paramètre θ et en termes de segmentation. Ainsi, la convergence établie par les résultats théoriques semble pouvoir être atteinte pour un nombre raisonnable d'itérations de l'échantillonneur de Gibbs.

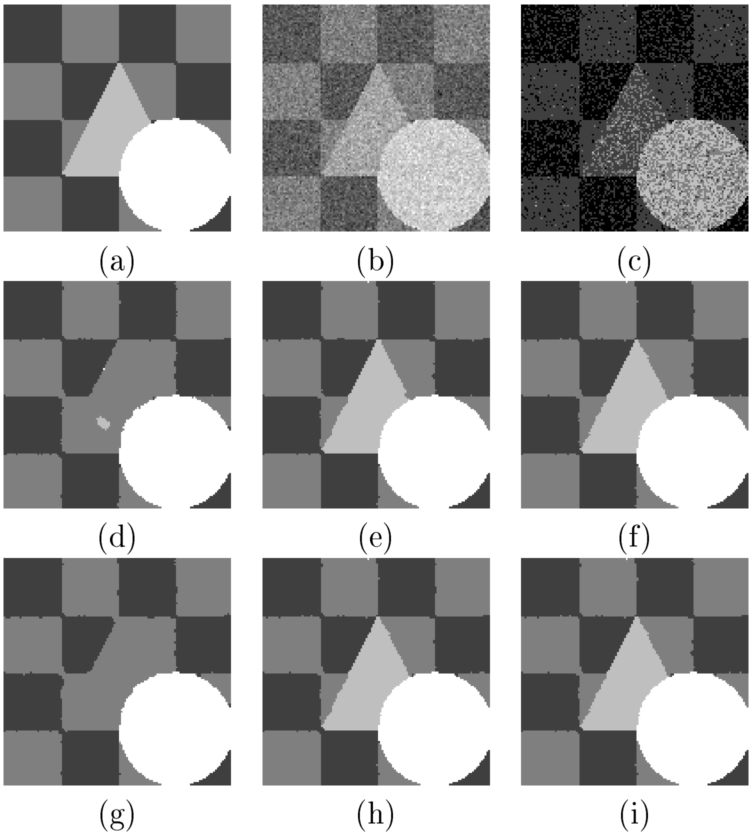


FIG. 5.1 – *image Damier dégradée avec un bruit gaussien : (a) image originale, (b) image bruitée, (c) segmentation initiale, (d), (e), (f) segmentations MCEM pour* $T_q = 1, 10, 100$, *(g), (h), (i) segmentations par l'algorithme en champ simulé généralisé pour* $T_q = 1, 10, 100$.

	T_q	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4
valeurs exactes	-	1	2	3	4	0.5	0.5	0.5	0.5
MCEM	1	1.0	2.2	3.5	4.0	0.5	0.6	0.6	0.5
	10	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.6
	100	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5
algorithme en champ simulé généralisé	1	1.0	2.0	2.2	4.0	0.5	0.3	0.6	0.5
	10	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5
	100	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5

TAB. 5.1 – *estimation des paramètres par l'algorithme MCEM et l'algorithme en champ simulé généralisé pour l'image Damier.*

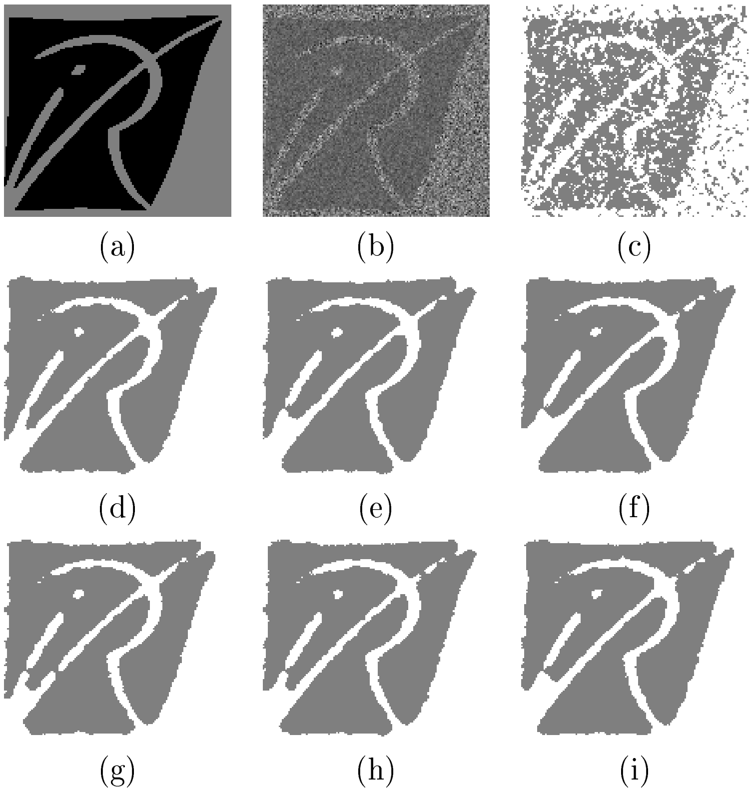


FIG. 5.2 – *image Logo dégradée avec un bruit gaussien : (a) image originale, (b) image bruitée, (c) segmentation initiale, (d), (e), (f) segmentations MCEM pour* $T_q = 1,\ 10,\ 100$, *(g), (h), (i) segmentations par l'algorithme en champ simulé généralisé pour* $T_q = 1,\ 10,\ 100$.

	T_q	μ_1	μ_2	σ_1	σ_2
valeurs exactes	-	51	255	130	300
	1	52.3	251	129	303
MCEM	10	52.8	251	129	303
	100	52.6	250	129	302
algorithme en	1	52.8	252	129	303
champ simulé	10	52.5	250	129	303
généralisé	100	52.5	249	129	302

TAB. 5.2 – *estimation des paramètres par l'algorithme MCEM et l'algorithme en champ simulé généralisé pour l'image Logo.*

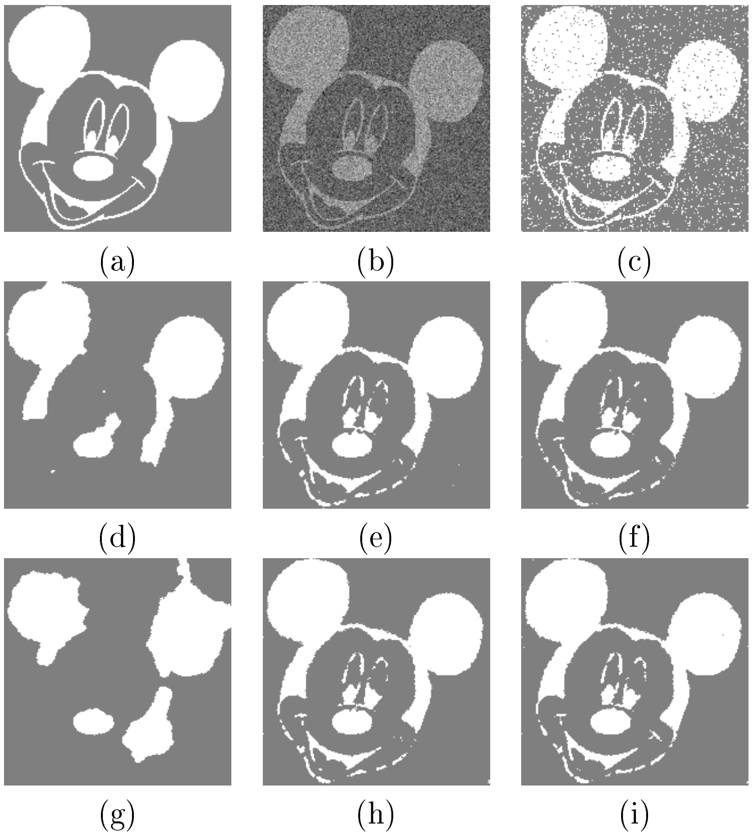


FIG. 5.3 – *image Mickey dégradée avec un bruit gaussien : (a) image originale, (b) image bruitée, (c) segmentation initiale, (d), (e), (f) segmentations MCEM pour* $T_q = 1,\ 10,\ 100$, *(g), (h), (i) segmentations par l'algorithme en champ simulé généralisé pour* $T_q = 1,\ 10,\ 100$.

	T_q	μ_1	μ_2	σ_1	σ_2
valeurs exactes	-	1	2	25	25
MCEM	1	1.2	1.7	0.6	0.7
	10	1.0	2.0	0.5	0.5
	100	1.0	2.0	0.5	0.5
algorithme en champ simulé généralisé	1	1.2	1.6	0.7	0.7
	10	1.0	2.0	0.5	0.5
	100	1.0	2.0	0.5	0.5

TAB. 5.3 – *estimation des paramètres par l'algorithme MCEM et l'algorithme en champ simulé généralisé pour l'image Mickey.*

Chapitre 6

Choix du nombre de classes

Dans les chapitres précédents, il a toujours été supposé que le nombre K de classes (ou étiquettes) dans l'image segmentée était connu. En pratique c'est rarement le cas et l'on se trouve confronté au problème suivant : *Le nombre de ce que l'on ne connaît pas fait partie de ce que l'on ne connaît pas* ! En effet, dans la plupart des domaines d'application des techniques de segmentation d'images, les experts n'ont souvent qu'une idée approximative du nombre de classes présentes dans les données. Seul un intervalle de valeurs possibles a priori est connu en général. Déterminer le nombre exact de classes dans l'image est un exercice subjectif. Il dépend fortement du critère utilisé pour faire le choix, qui n'est pas forcément le même pour tous les utilisateurs de l'image. À côté de ces critères d'experts ou ad hoc existent des critères basés sur les caractéristiques de l'image observée. Ces derniers ne conduisent pas forcément aux mêmes choix que les premiers, car ils ne prennent pas en compte l'objectif de l'utilisateur des données. Ils présentent par contre l'avantage d'être objectifs et de conduire à des techniques automatiques pour la sélection du nombre de classes. Dans tous les cas, un principe de parcimonie est souvent adopté car l'expérience a montrée que le choix d'un petit nombre de classes est souvent plus facilement interprétable et exploitable.
Dans un cadre d'analyse statistique d'images, sélectionner K revient à sélectionner le modèle probabiliste le plus approprié aux données observées. Récemment, la technique des chaînes de Markov de Monte-Carlo à sauts réversibles (Green 1995) a été utilisée afin d'estimer simultanément le nombre de classes et les paramètres du modèle par Barker et Rayner (1997) pour des images en niveaux de gris et par (Kato 1999) pour des images en couleurs. Cette méthode bayésienne est assez lourde à mettre en œuvre (choix des lois a priori, problème de permutation des classes). Une autre approche bayésienne pour la sélection consiste à prendre le modèle pour lequel la probabilité des observations (*i.e.* la vraisemblance des paramètres intégrée) est la plus grande. Cette quantité n'est en général pas accessible directement et peut être approchée par des méthodes par *Chaînes de Markov de Monte-Carlo* (MCMC) (*e.g.* Kass et Raftery 1995), relativement coûteuses. Une autre méthode consiste à approcher la log-vraisemblance intégrée par le maximum de la log-vraisemblance pénalisée. Dans le cas du critère *Akaike Information Criterium* (AIC, Akaike 1973), la pénalité est égale à deux fois le nombre de paramètres indépendants du modèle et l'on observe que ce critère ne

pénalise pas assez la complexité du modèle et surestime le nombre de paramètres nécessaires. Le *Bayesian Information Criterion* (BIC) de Schwarz (1978), ou de manière équivalente le *Minimum Description Length* (MDL, Rissanen 1978), tend à sélectionner des modèles moins complexes avec un terme de pénalité égal au nombre de paramètres indépendants multiplié par le logarithme de la taille des données. La maximisation de la vraisemblance complète intégrée est une autre approche qui permet de prendre en compte l'objectif de segmentation dans la sélection du modèle. Cela conduit au critère *Integrated Completed Likelihood* (ICL), plus spécifique au problème de la sélection du nombre de classes que BIC, et proposé par Biernacki, Celeux, et Govaert (2000) pour des modèles de mélanges finis indépendants. Pour ces modèles à structure simple, des méthodes efficaces existent pour estimer les paramètres (*e.g.* algorithme EM, voir Dempster, Laird, et Rubin 1977 et la partie 2.2.2), et le calcul de la vraisemblance ou de la vraisemblance complète est direct. Pour les modèles plus complexes de champs de Markov cachés, le calcul direct des critères n'est plus possible. Nous proposons des approximations de BIC et ICL (présentées dans Forbes et Peyrard 2001) basées sur les approximations de type champ moyen d'un champ de Markov (voir chapitre 3 et la partie 4.2), dont le calcul fait appel à des procédures peu coûteuses. Nous testons ces nouveaux critères pour la sélection du nombre de classes sur des images de différentes natures afin d'une part d'évaluer les potentialités de ces critères et d'autre part de juger de l'intérêt de la prise en compte de l'information spatiale pour traiter ce problème. Les critères proposés, comme BIC et ICL, peuvent être utilisés dans un contexte plus général de sélection d'un modèle de champ de Markov.

6.1 Critères pour la sélection de modèle

6.1.1 Critère BIC

Le choix du nombre de classes dans un jeu de données (ici une image) étant abordé comme un problème de comparaison de modèles probabilistes, les outils de la théorie bayésienne sont disponibles. L'approche bayésienne pour la sélection de modèle est basée sur la comparaison des probabilités a posteriori des différents modèles envisageables. Soit M_K le modèle pour les données complètes $(\mathbf{z},\mathbf{y})$ défini par (1.3) et (1.8), lorsque le nombre d'étiquettes est égal à K, pour K appartenant à $\{K_{min},\ldots,K_{max}\}$. La vraisemblance des paramètres intégrée, pour le modèle M_K, est définie par

$$g_G(\mathbf{y}|M_K) = \int g_G(\mathbf{y}|\Psi_K,M_K)P(\Psi_K)d\Psi_K, \tag{6.1}$$

où Ψ_K est le paramètre du modèle M_K, $g_G(\mathbf{y}|\Psi_K,M_K)$ est la vraisemblance associée et $P(\Psi_K)$ est la distribution a priori du paramètre Ψ_K pour le même modèle. Si $P(M_K)$ désigne la probabilité a priori du modèle M_K alors la probabilité a posteriori de M_K sachant les observations $\mathbf{y}$ est

$$P(M_K|\mathbf{y}) \;=\; \frac{g_G(\mathbf{y}|M_K)P(M_K)}{\sum_{k=K_{min}}^{K_{max}} g_G(\mathbf{y}|M_k)P(M_k)}.$$

Supposons que les modèles sont équiprobables a priori. Le rapport des probabilités a posteriori des modèles M_K et M_L, $K \neq L$, encore nommé *facteur de Bayes* du modèle M_K contre le modèle M_L, est égal à,

$$\frac{P(M_K|\mathbf{y})}{P(M_L|\mathbf{y})} = \frac{g_G(\mathbf{y}|M_K)}{g_G(\mathbf{y}|M_L)}. \tag{6.2}$$

Ainsi, l'approche bayésienne revient à choisir le modèle qui maximise la vraisemblance des paramètres intégrée $g_G(\mathbf{y}|M_K)$. Évaluer l'intégrale (6.1) est en général difficile. Le critère BIC (voir par exemple Schwarz 1978, Kass et Raftery 1995) est dérivé de l'approximation de Laplace de la vraisemblance des paramètres intégrée. Il est donné par la formule suivante :

$$2\log g_G(\mathbf{y}|M_K) \approx \mathrm{BIC}(M_K) = 2\log g_G(\mathbf{y} \mid \Psi_K^{\mathrm{MV}}) - d_K \log(n), \tag{6.3}$$

où Ψ_K^{MV} est l'estimateur du maximum de vraisemblance (*cf.* paragraphe 2.1.2.1)

$$\Psi_K^{\mathrm{MV}} \;=\; \arg\max_{\Psi_K} g_G(\mathbf{y} \mid \Psi_K),$$

et où d_K est le nombre de paramètres indépendants du modèle M_K. Ce critère présente l'avantage de ne pas nécessiter le choix d'une distribution a priori des paramètres. Il s'écrit comme la log-vraisemblance des paramètres pénalisée par le terme $d_K \log(n)$, venant désavantager les modèles complexes. L'erreur due à l'approximation ne s'annule pas, même asymptotiquement, cependant dans le cas d'un échantillon de grande taille le rapport $\mathrm{BIC}(M_K)/\mathrm{BIC}(M_L)$ permet de donner un ordre de grandeur du facteur de Bayes et suffit pour indiquer l'attraction d'un modèle par rapport à un autre. Dans le cas de l'estimation du nombre de classes d'un modèle de mélange fini indépendant, les conditions de régularité assurant la justification théorique de BIC ne sont pas vérifiées, car les paramètres ne sont pas identifiables (voir Aitkin et Rubin 1985). Cependant, des expériences pour ces mêmes modèles (Roeder et Wasserman 1997, Fraley et Raftery 1998) montrent l'intérêt en pratique de ce critère. En outre, Leroux (1992) a établi, sous des conditions raisonnables sur le modèle, qu'asymptotiquement BIC ne surestime pas le vrai nombre de composants d'un modèle de mélange fini indépendant. On observe cependant qu'il a tendance à surestimer le nombre de classes lorsque le vrai modèle ne fait pas partie de la famille des modèles considérés, et ce au détriment de la séparabilité des classes. Ce phénomène est observé notamment dans le cas de données réelles, BIC ne conduisant pas toujours à une classification pertinente (Biernacki, Celeux, et Govaert 2000). Si le cas des modèles de mélanges finis indépendants a été largement étudié, il existe peu de travaux sur le critère BIC pour la sélection de modèles de champs de Markov discrets. Seymour et Ji (1996) et Ji et Seymour (1996) ont étudié BIC dans le cas où le champ $\mathbf{z}$ est observé, et ils proposent des approximations par Chaîne de Markov de Monte-Carlo ou par la pseudo-vraisemblance pour calculer BIC, afin de traiter la complexité de ces modèles.

6.1.2 Critère ICL

Dans le but de prendre en compte la pertinence de la classification, Biernacki, Celeux, et Govaert (2000) ont proposé de choisir le nombre de composantes d'un modèle de mélange fini indépendant en maximisant la vraisemblance complète intégrée $P_G(\mathbf{z},\mathbf{y} \mid M_K)$. L'évaluation de cette dernière implique d'intégrer suivant la loi a priori des paramètres. Biernacki, Celeux, et Govaert (2000) considèrent alors l'approximation *à la* BIC suivante

$$2 \log P_G(\mathbf{z},\mathbf{y} \mid M_K) \quad \approx \quad 2 \log P_G(\mathbf{z},\mathbf{y} \mid \Psi_K^{\mathrm{MV}}) - d_K \log(n).$$

En général, le champ $\mathbf{z}$ est inconnu. Il est alors estimé au moyen d'un opérateur du maximum a posteriori (*cf.* paragraphe 2.1.1.1) :

$$\mathbf{z}^{\mathrm{MAP}} = \arg\max_{\mathbf{z}} P_G(\mathbf{z} \mid \mathbf{y},\Psi_K^{\mathrm{MV}}),$$

ce qui conduit au critère ICL défini ainsi :

$$\begin{aligned} \mathrm{ICL}(M_K) &= 2 \log P_G(\mathbf{z}^{\mathrm{MAP}},\mathbf{y} \mid \Psi_K^{\mathrm{MV}}) - d_K \log(n) \\ &= \mathrm{BIC}(M_K) + 2 \log P_G(\mathbf{z}^{\mathrm{MAP}} \mid \mathbf{y},\Psi_K^{\mathrm{MV}}). \end{aligned} \tag{6.4}$$

Ce critère se comporte essentiellement comme BIC pénalisé par un terme d'entropie qui favorise des classes bien séparées. Ceci apparaît plus nettement si l'on considère l'espérance conditionelle suivante

$$\begin{aligned} \mathrm{EICL}(M_K) &= 2\mathbb{E}[\log P_G(\mathbf{Z},\mathbf{y} \mid \Psi_K^{\mathrm{MV}}) \mid \mathbf{y},\Psi_K^{\mathrm{MV}}] - d_K \log(n) \\ &= \mathrm{BIC}(M_K) - 2\mathbf{S}(P_G(. \mid \mathbf{y},\Psi_K^{\mathrm{MV}})). \end{aligned}$$

où l'on rappelle que $\mathbf{S}(P) = -\sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z})$ est l'entropie associée à une distribution P donnée. Le terme $\mathbf{S}(P)$ est d'autant plus grand que la distribution P est proche d'une distribution uniforme. La pénalité $2\mathbf{S}(P_G(. \mid \mathbf{y},\Psi_K^{\mathrm{MV}}))$ vient ainsi récompenser des classes bien séparées. En pratique, dans le cas d'un modèle de mélange fini indépendant, les expériences réalisées par Biernacki, Celeux, et Govaert (2000) suggèrent que si BIC est à préférer à ICL en ce qui concerne l'estimation de densité, par contre ICL devient plus performant pour répondre au problème de classification.

Remarquons que les critères BIC, ICL et EICL sont liés par les relations suivantes :

$$\mathrm{BIC}(M_K) > \mathrm{ICL}(M_K) > \mathrm{EICL}(M_K).$$

La première inégalité se déduit directement de (6.4). À partir de la définition de $\mathbf{z}^{\mathrm{MAP}}$ on obtient $\log P_G(\mathbf{z}^{\mathrm{MAP}} \mid \Psi_K^{\mathrm{MV}}) \geq \mathbb{E}[\log P_G(\mathbf{Z} \mid \Psi_K^{\mathrm{MV}}) \mid \mathbf{y},\Psi_K^{\mathrm{MV}}]$, dont découle la deuxième inégalité.

6.2 Approximations de type champ moyen des critères BIC et ICL

Dans ce chapitre, nous éludons la question de savoir si les critères BIC et ICL sont encore de bonnes approximations des vraisemblances intégrées dans le cas des champs de Markov cachés. Notre objectif est de proposer des méthodes d'approximation de ces critères qui conduisent à des critères de sélection d'un champ de Markov qui restent satisfaisants en pratique mais plus simples à calculer.
Pour un modèle de champ de Markov caché, la quantité $P_G(\mathbf{y} \mid \Psi_K^{\mathrm{MV}})$ n'est pas accessible directement et des approximations sont nécessaires pour calculer BIC (*e.g.* approximation par la pseudo-vraisemblance, Stanford et Raftery 2001). Pour les mêmes raisons, le critère ICL n'est pas calculable directement pour ce modèle. Nous avons vu dans le chapitre 3 comment le principe du champ moyen permet d'approximer la distribution d'un modèle de champ de Markov caché. La généralisation de la méthode conduit aux approximations de type champ moyen, qui permettent de se ramener au cas simple d'un modèle de mélange fini indépendant (voir partie 4.2). Dans cette partie, nous appliquons cette idée pour approcher la vraisemblance $P_G(\mathbf{y} \mid \Psi_K)$ et la vraisemblance complète $P_G(\mathbf{z},\mathbf{y} \mid \Psi_K)$ par des quantités plus simples à évaluer. Nous considérons dans un premier temps des critères obtenus par approximation de type champ moyen de la distribution marginale $P_G(\mathbf{z} \mid \beta_K)$. Dans une deuxième partie nous montrons qu'il n'est en fait pas nécessaire d'approximer entièrement cette distribution et nous présentons des critères approximant BIC et ICL à partir de l'approximation en champ moyen à l'ordre 1 de la fonction de partition d'un modèle de champ de Markov.

6.2.1 Approximation de la distribution de Gibbs

Nous proposons d'utiliser une approximation de type champ moyen de la distribution markovienne marginale $P_G(\mathbf{z} \mid \beta_K)$ d'un champ de Markov à K classes :

$$P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta_K) \;=\; \prod_{i \in S} P_G(z_i \mid \mathbf{Z}_{N(i)} = \tilde{\mathbf{z}}_{N(i)}, \beta_K), \qquad (6.5)$$

où $\tilde{\mathbf{z}}$ est un champ constant dans $\mathcal{V}^n$ à définir. Ceci conduit aux approximations suivantes de la vraisemblance complète et de la vraisemblance des paramètres (voir partie 4.3)

$$\begin{aligned} P_{\tilde{\mathbf{z}}}(\mathbf{z},\mathbf{y} \mid \Psi_K) &= \prod_{i \in S} g(y_i \mid z_i, \theta_K)\; P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}, \beta_K). \\ P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_K) &= \prod_{i \in S} \sum_{z_i} g(y_i \mid z_i, \theta_K)\; P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}, \beta_K), \end{aligned}$$

En se ramenant à un système indépendant de type (6.5), les difficultés de calcul sont supprimées. Ainsi, soit $\hat{\Psi}_K$ une approximation de Ψ_K^{MV} et $\hat{\mathbf{z}}$ une approximation de $\mathbf{z}^{\mathrm{MAP}}$ (ceux-ci n'étant pas disponibles pour les modèles de champ de Markov cachés), les critères BIC et ICL sont alors approximés par

$$\begin{aligned} \mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_K) &= 2\log P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \hat{\Psi}_K) - d_K \log(n), & (6.6)\\ \mathrm{ICL}^{\tilde{\mathbf{z}}}(\hat{\mathbf{z}},\hat{\Psi}_K) &= 2\log P_{\tilde{\mathbf{z}}}(\hat{\mathbf{z}},\mathbf{y} \mid \hat{\Psi}_K) - d_K \log(n). & (6.7)\end{aligned}$$

En ce qui concerne l'approximation $\hat{\Psi}_K$, nous proposons d'utiliser celle fournie par l'une des trois procédures de type champ moyen décrites dans le chapitre 4 (algorithmes en champ moyen, en champ modal et en champ simulé). La segmentation $\hat{\mathbf{z}}$ est ensuite obtenue suivant la règle (4.22). Notons que l'étude des performances de ces trois algorithmes réalisée dans la partie 4.9 conduit à choisir de préférence l'algorithme en champ simulé. Nous avons comparé deux méthodes pour le choix des voisins $\tilde{\mathbf{z}}$. La première consiste à choisir la valeur $\tilde{\mathbf{z}}$ fournie par l'algorithme de type champ moyen utilisé pour obtenir $\hat{\Psi}_K$ et $\hat{\mathbf{z}}_K$. Avec la deuxième méthode, $\tilde{\mathbf{z}}$ est l'approximation en champ moyen $\mathbf{z}^{mf}$ de l'espérance de la distribution de Gibbs marginale $P_G(\mathbf{z} \mid \hat{\beta}_K)$. Cette solution est motivée par le fait que $P_{\mathbf{z}^{mf}}(\mathbf{z} \mid \hat{\beta}_K)$ est la meilleure approximation, par une distribution factorisable, de la distribution du champ de Markov marginal $P_G(\mathbf{z} \mid \hat{\beta}_K)$ au sens de la divergence de Kullback-Leibler (voir partie 3.2). Cette dernière méthode conduisant en pratique à des résultats peu satisfaisants sur des données simulées (sous-estimation de K), tout en nécessitant le calcul supplémentaire de $\mathbf{z}^{mf}$, nous n'avons finalement pas retenu cette solution.

L'approximation $\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_K)$ du critère BIC pour champ de Markov caché est à rapprocher du critère *Pseudo-Likelihood Information Criterium* (PLIC) proposée par Stanford et Raftery (2001). L'idée est de ne considérer que les configurations proches de la segmentation obtenue par ICM dans le calcul de la vraisemblance des paramètres intégrée et de supposer que les observations $\mathbf{Y}$ sont indépendantes. Cela conduit à la définition suivante de PLIC : soit Ψ_K^{ICM} et $\mathbf{z}^{\mathrm{ICM}}$ respectivement l'estimateur de Ψ_K et la segmentation en K classes obtenus par la procédure ICM non supervisée,

$$\begin{aligned} \mathrm{PLIC}(M_K) &= 2\log\left(\prod_{i\in S}\sum_{z_i} g(y_i \mid z_i,\theta_K^{\mathrm{ICM}}) P_G(z_i \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}^{\mathrm{ICM}},\beta_K^{\mathrm{ICM}})\right) - d_K \log(n)\\ &= 2\log(\prod_{i\in S} g_G(y_i \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}^{\mathrm{ICM}},\Psi_K^{\mathrm{ICM}})) - d_K\log(n). & (6.8)\end{aligned}$$

L'approximation par la pseudo-vraisemblance n'apparaît pas directement dans l'expression de PLIC. Elle intervient dans l'algorithme ICM pour le calcul de Ψ_K^{ICM} et $\mathbf{z}^{\mathrm{ICM}}$. On remarque que la vraisemblance $\prod_{i\in S} g_G(y_i \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}^{\mathrm{ICM}},\Psi_K^{\mathrm{ICM}})$ intervenant dans l'équation (6.8) est égale à la densité des observations lorsque la distribution du champ caché est une distribution de type champ moyen $P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta_K)$ avec $\tilde{\mathbf{z}} = \mathbf{z}^{\mathrm{ICM}}$ et $\Psi = \Psi_K^{\mathrm{ICM}}$ (voir partie 4.2 et équation 4.12). Ainsi, PLIC se réécrit

$$\begin{aligned} \mathrm{PLIC}(M_K) &= 2\log(g_{\mathbf{z}^{\mathrm{ICM}}}(\mathbf{y} \mid \Psi^{\mathrm{ICM}})) - d_K\log(n)\\ &= \mathrm{BIC}^{\mathbf{z}^{\mathrm{ICM}}}(\Psi^{\mathrm{ICM}}).\end{aligned}$$

Le critère PLIC apparaît donc comme un cas particulier de la famille d'approximation $\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_K)$, et offre une variante aux méthodes de type champ moyen pour le choix de $\hat{\Psi}_K$, $\hat{\mathbf{z}}_K$ et $\tilde{\mathbf{z}}$.

6.2.2 Approximation de la constante de normalisation

Les approximations de type champ moyen des critères BIC et ICL qui viennent d'être définies se basent sur les écritures (6.3) et (6.4) de ces deux quantités. Dans cette partie nous réécrivons les critères en fonction des constantes de normalisation marginale et conditionnelle du modèle de champ de Markov caché. Cette reformulation permet de mettre en avant le fait qu'il n'est pas nécessaire d'approximer entièrement les distributions de Gibbs intervenant dans le calcul de BIC et ICL, et suggère une meilleure approximation de ces critères. Soit $W(\mathbf{y},\Psi_K)$ et $W(\beta_K)$ les fonctions de partition des distributions markoviennes marginale et conditionnelle, pour un modèle à K classes

$$\begin{aligned} W(\beta_K) &= \sum_{\mathbf{z}} \exp(-H(\mathbf{z} \mid \beta_K)), \\ W(\mathbf{y},\Psi_K) &= \sum_{\mathbf{z}} \exp(-H(\mathbf{z} \mid \mathbf{y},\Psi_K)). \end{aligned}$$

La loi des observations sous le modèle de champ de Markov caché est donnée par

$$g_G(\mathbf{y} \mid \Psi_K) = \frac{P_G(\mathbf{z},\mathbf{y} \mid \Psi_K)}{P_G(\mathbf{z} \mid \mathbf{y},\Psi_K)} = \frac{g(\mathbf{y} \mid \mathbf{z},\theta_K)\ P_G(\mathbf{z} \mid \beta_K)}{P_G(\mathbf{z} \mid \mathbf{y},\Psi_K)},$$

de sorte que

$$g_G(\mathbf{y} \mid \Psi_K) = \frac{g(\mathbf{y} \mid \mathbf{z},\theta_K)\ \exp(-H(\mathbf{z} \mid \beta_K))}{\exp(-H(\mathbf{z} \mid \mathbf{y},\Psi_K))} \frac{W(\mathbf{y},\Psi_K)}{W(\beta_K)},$$

ce qui d'après (1.9) se simplifie encore

$$g_G(\mathbf{y} \mid \Psi_K) = \frac{W(\mathbf{y},\Psi_K)}{W(\beta_K)}.$$

Ainsi l'expression (6.3) de BIC est équivalente à

$$\mathrm{BIC}(M_K) = 2\log W(\mathbf{y},\Psi_K^{\mathrm{MV}}) - 2\log W(\beta_K^{\mathrm{MV}}) - d_K \log(n), \tag{6.9}$$

qui fait uniquement appel aux fonctions de partition. Le critère ICL (6.4) peut également s'exprimer comme une somme de termes calculables directement, plus un terme dépendant de $W(\beta)$

$$\begin{aligned} \mathrm{ICL}(M_K) &= 2\log g(\mathbf{y} \mid \mathbf{z}^{\mathrm{MAP}},\Psi_K^{\mathrm{MV}}) + 2\log P_G(\mathbf{z}^{\mathrm{MAP}} \mid \beta_K^{\mathrm{MV}}) - d_K \log(n) \\ &= 2\log g(\mathbf{y} \mid \mathbf{z}^{\mathrm{MAP}},\Psi_K^{\mathrm{MV}}) \\ &\quad -2H(\mathbf{z}^{\mathrm{MAP}} \mid \beta_K^{\mathrm{MV}}) - 2\log W(\beta_K^{\mathrm{MV}}) - d_K \log(n). \end{aligned} \tag{6.10}$$

Soit $H^{mf}(\mathbf{z} \mid \beta_K)$ et $H^{mf}(\mathbf{z} \mid \mathbf{y},\Psi_K)$ les approximations en champ moyen des énergies $H(\mathbf{z} \mid \beta_K)$ et $H(\mathbf{z} \mid \mathbf{y},\Psi_K)$ des champs marginal et conditionnel, obtenues à partir de la formule (3.15). Notons $W^{mf}(\beta_K)$ et $W^{mf}(\mathbf{y},\Psi_K)$ les approximations correspondantes des fonctions de partition (voir équation (3.16)) et notons $\mathbb{E}^{mf}[.]$ (resp. $\mathbb{E}^{mf}[.|\mathbf{y}]$) l'espérance selon la distribution définissant l'approximation en champ moyen du champ de Markov marginal (resp. conditionnel). Nous proposons alors d'utiliser l'approximation en champ moyen à l'ordre 1 des constantes de normalisation marginale et conditionnelle, qui fournissent également un maximum de la borne de Gibbs-Bogoliubov-Feynman (*cf.* partie 3.2) :

$$\begin{aligned} W(\beta_K) &\approx W^{\text{GBF}}(\beta_K) = W^{mf}(\beta_K) \exp\{-\mathbb{E}^{mf}[H(\mathbf{Z} \mid \beta_K) - H^{mf}(\mathbf{Z} \mid \beta_K)]\}, \\ W(\mathbf{y},\Psi_K) &\approx W^{\text{GBF}}(\mathbf{y},\Psi_K) = W^{mf}(\mathbf{y},\Psi_K) \exp\{-\mathbb{E}^{mf}[H(\mathbf{Z} \mid \mathbf{y},\Psi_K) - H^{mf}(\mathbf{Z} \mid \mathbf{y},\Psi_K)]\}. \end{aligned}$$

Il pourrait paraître naturel d'approximer les fonctions de partition marginale et conditionnelle par $W^{mf}(\beta_K)$ et $W^{mf}(\mathbf{y},\Psi_K)$, puisque ce sont les fonctions de partition des distributions découlant de l'approximation en champ moyen des distributions markoviennes marginale et conditionnelle. Cependant, l'étude effectuée dans le cas particulier du modèle de Potts simple binaire (*cf.* partie 3.3) montre de manière nette sur cet exemple que l'approximation à l'ordre 1 est plus fine. Ces résultats ont été confirmés en pratique : nous avons comparé le comportement du critère ICL selon que l'approximation à l'ordre zéro ou à l'ordre 1 des fonctions de partition est utilisée et nous avons constaté de mauvaises performances pour la sélection du nombre de classes avec le critère utilisant l'approximation à l'ordre zéro (sous-estimation de K).
Les approximations de BIC et ICL basées sur l'approximation à l'ordre 1 des fonctions de partition sont les suivantes :

$$\begin{aligned} \text{BIC}^{W^{\text{GBF}}}(\hat{\Psi}_K) &= 2\log W^{mf}(\mathbf{y},\hat{\Psi}_K) - 2\mathbb{E}^{mf}[H(\mathbf{Z} \mid \mathbf{y},\hat{\Psi}_K) - H^{mf}(\mathbf{Z} \mid \mathbf{y},\hat{\Psi}_K) \mid \mathbf{y}] \\ &\quad -2\log W^{mf}(\hat{\beta}_K) + 2\mathbb{E}^{mf}[H(\mathbf{Z} \mid \hat{\beta}_K) - H^{mf}(\mathbf{Z} \mid \hat{\beta}_K)] \\ &\quad -d_K \log(n), \end{aligned} \tag{6.11}$$

$$\begin{aligned} \text{ICL}^{W^{\text{GBF}}}(\hat{\mathbf{z}},\hat{\Psi}_K) &= 2\log P(\mathbf{y} \mid \hat{\mathbf{z}},\hat{\Psi}_K) - 2H(\hat{\mathbf{z}}|\hat{\beta}_K) \\ &\quad -2\log W^{mf}(\hat{\beta}_K) + 2\mathbb{E}^{mf}[H(\mathbf{Z} \mid \hat{\beta}_K) - H^{mf}(\mathbf{Z} \mid \hat{\beta}_K)] \\ &\quad -d_K \log(n). \end{aligned} \tag{6.12}$$

Notons que les deux approches en champ moyen qui viennent d'être décrites pour l'approximation de BIC et ICL peuvent aussi s'appliquer au critère EICL. Cependant, nous avons observé qu'en pratique les critères ainsi obtenus ne sont pas plus performants que les approximations de type champ moyen de ICL. Aussi, nous ne détaillons pas plus ici les approximations de EICL correspondantes.

6.3 Résultats expérimentaux

Les différentes approximations de type champ moyen des critères BIC et ICL pour un modèle de champ de Markov caché sont utilisées pour sélectionner le nombre de classes

présentes dans des jeux de données de différentes natures. Lorsque le nombre exact de classes formant l'image segmentée est connu, celui-ci est désigné par K_T. Le modèle choisi pour réaliser les simulations et celui supposé pour segmenter les observations est identique au modèle présenté dans la partie 4.9 : la distribution du champ caché est la distribution d'un modèle de Potts simple (*cf.* équation (3.10)) et la densité des observations est gaussienne conditionnellement aux étiquettes des pixels (*cf.* équation (4.21)). Pour chaque exemple, les critères $\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_K)$, $\mathrm{BIC}^{W_{\mathrm{GBF}}}(\hat{\Psi}_K)$, $\mathrm{ICL}^{\tilde{\mathbf{z}}}(\hat{\mathbf{z}},\hat{\Psi}_K)$ et $\mathrm{ICL}^{W_{\mathrm{GBF}}}(\hat{\mathbf{z}},\hat{\Psi}_K)$ sont calculés pour $K_{min} \leq K \leq K_{max}$. Afin d'alléger les notations, par la suite les quatres critères seront notés respectivement BIC_p, BIC_w, ICL_p et ICL_w, l'indice indiquant si le critère est obtenu par approximation de la distribution de Gibbs marginale (p) ou des fonctions de partition (w). Le paramètre $\hat{\Psi}_K$ ainsi que les configurations $\tilde{\mathbf{z}}$ et $\hat{\mathbf{z}}$ sont obtenues pour 100 itérations de l'algorithme en champ simulé. Nous avons également utilisé les algorithmes en champ moyen et en champ modal, sans noter d'amélioration significative des résultats. Ces critères obtenus par approximation de type champ moyen sont comparés avec le critère PLIC, basé sur les estimations fournies par l'algorithme ICM non supervisé. Afin de juger de l'importance de la prise en compte de la dépendance spatiale existante dans l'image pour la sélection de K, les résultats obtenus par les critères BIC et ICL pour un modèle de mélange fini indépendant sont également reportés. Nous noterons ces deux derniers critères $\mathrm{BIC}^{\mathrm{EM}}$ et $\mathrm{ICL}^{\mathrm{EM}}$
La première série d'expériences réalisées concerne des images générées selon la distribution d'un modèle de Potts simple. Pour plusieurs valeurs de K_T et de β, les résultats moyens obtenus sur un ensemble de 100 simulations sont présentés dans la partie 6.3.1. Nous avons ensuite comparé les différents critères sur des images avec bruit simulé et des images réelles (parties 6.3.2 et 6.3.3). Pour ces deux derniers cas, lorsque plusieurs critères, parmi les quatres critères tenant compte de l'information spatiale, sélectionnent le nombre exact de classes (cas des images avec bruit simulé) ou un même nombre pertinent de classes (cas des images réelles), nous avons comparé les courbes d'évolution des critères en fonction de K afin de déterminer une possible tendance de l'un d'entre eux à marquer plus fortement ce choix. Dans le même ordre d'idée, lorsque la valeur du critère considéré pour $K = K'$ est proche du maximum local du critère, la valeur K' est également reportée entre parenthèse dans les tableaux, à côté de la valeur $\hat{K}$ correspondant au maximum local.

6.3.1 Simulations d'un modèle de Potts

Cette première série d'expériences concerne des images simulées suivant un modèle de champ de Markov caché. Pour un modèle de Potts simple donné, nous avons simulé 100 réalisations de taille 100×100 de ce modèle, que nous avons ensuite bruitées avec un bruit gaussien de sorte que la densité des observations est donnée par (4.21) avec $\mu_k = k$ et $\sigma_k = 0.5$ pour tout k dans $\{1,\ldots,K_T\}$ (voir partie 4.9.1 pour plus de détails). Pour chacune des simulations, le nombre de classes est estimé par maximisation des critères étudiés. Ce schéma a été reproduit pour différentes valeurs de K_T (K_T = 2, 3, 4, 5 et 6) et pour différentes valeurs du paramètre spatial β, soit au total six exemples. Les figures 6.1 (a) à (f) montrent un exemple de réalisation de chacun des six modèles considérés. Le bruit ajouté à

ces images est suffisamment important de sorte qu'il n'est pas possible à la simple vue des histogrammes de retrouver le nombre de classes présentes dans les données originales (figure 6.2).
Les tableaux 6.1 (a) à (f) reportent le pourcentage du nombre d'expériences (parmi les 100) où un nombre K de classes a été selectionné, pour $K_{min} \leq K \leq K_{max}$ et pour les différents critères. On observe sur l'ensemble des six cas traités que le critère $\mathrm{ICL}^{\mathrm{EM}}$ sélectionne toujours un modèle à une classe. Le critère $\mathrm{BIC}^{\mathrm{EM}}$ conduit à de meilleurs résultats puisqu'il sélectionne majoritairement la valeur K_T sur quatre exemples et la valeur $K_T - 1$ dans les deux autres cas.
Parmi les critères étudiés qui prennent en compte l'information spatiale présente dans les données, trois critères se distinguent par leur bonnes performances en ce qui concerne la sélection du nombre de classes sur l'ensemble des tests effectués et en ce qui concerne la stabilité des résultats fournis. Il s'agit des critères PLIC, ICL_p et BIC_w. Les deux premiers critères conduisent au choix K_T sauf pour l'exemple d'un modèle à 2 classes de paramètre spatial $\beta = 0.2$. Le critère BIC_w permet toujours de retrouver K_T. On observe une tendance marquée du critère BIC_p à surestimer le nombre de classes présentes et inversement un tendance du critère ICL_w à le sous-estimer.
Nous constatons l'importance du choix de la méthode d'estimation des paramètres pour le calcul de l'approximation $\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_K)$ de $\mathrm{BIC}(M_K)$. En effet, nous rappelons que PLIC, comme BIC_p, est une approximation de type $\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_K)$. Or les performances des critères PLIC et BIC_p sont très différentes sur ces exemples simulés. Notons que cette famille d'approximations utilise non seulement une estimation $\hat{\Psi}_K$ du paramètre du modèle, mais aussi un champ voisin $\tilde{\mathbf{z}}$, ce qui peut expliquer les variations dans les résultats suivant que ces deux quantités sont obtenues par ICM ou l'algorithme en champ simulé. Notamment, le champ $\tilde{\mathbf{z}}$ est obtenu par application de la règle du MAP dans le cas de ICM, et par simulation dans le cas de l'algorithme en champ simulé.

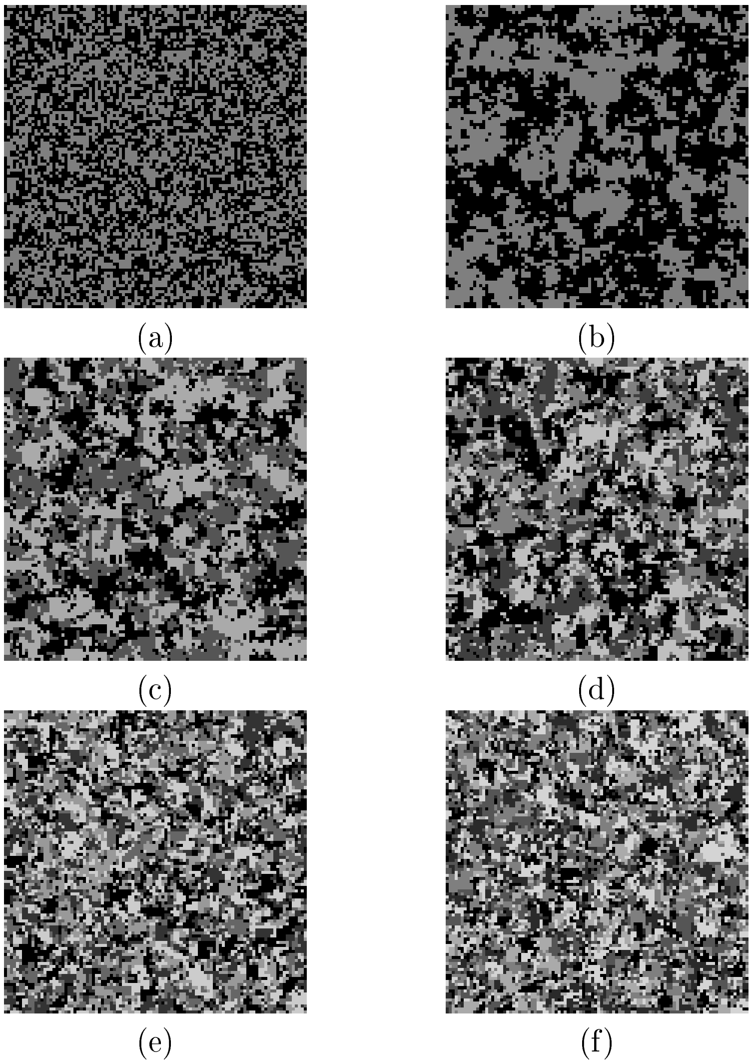


FIG. 6.1 – *simulation d'un modèle de Potts simple pour différentes valeurs de K_T et de β : (a) $K_T = 2$, $\beta = 0.2$, (b) $K_T = 2$, $\beta = 0.78$, (c) $K_T = 3$, $\beta = 0.9$, (d) $K_T = 4$, $\beta = 1$, (e) $K_T = 5$, $\beta = 1$, (f) $K_T = 6$, $\beta = 1.1$.*

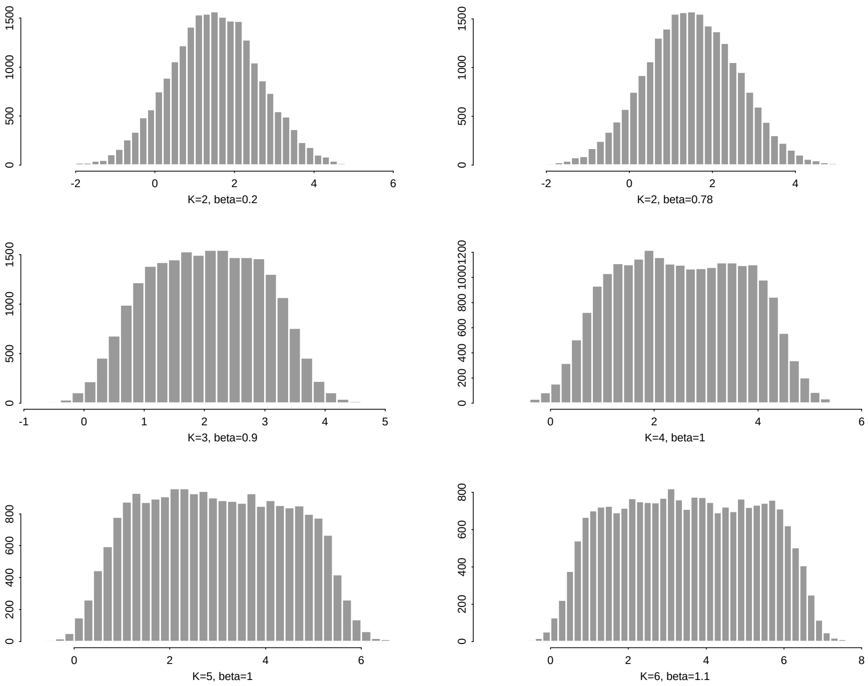


FIG. 6.2 – *exemples d'histogrammes d'observations obtenues par simulation d'un modèle de Potts simple dégradé par un bruit gaussien, pour différentes valeurs de K_T et de β.*

$K_T = 2,\ \beta = 0.2$				
$\hat{K}$	1	2	3	4
BIC^{EM}	0	100	0	0
ICL^{EM}	100	0	0	0
PLIC	69	31	0	0
BIC_p	0	43	22	35
ICL_p	100	0	0	0
BIC_w	0	67	15	18
ICL_w	100	0	0	0

$K_T = 2,\ \beta = 0.78$				
$\hat{K}$	1	2	3	4
BIC^{EM}	0	100	0	0
ICL^{EM}	100	0	0	0
PLIC	0	100	0	0
BIC_p	0	8	32	60
ICL_p	0	100	0	0
BIC_w	0	100	0	0
ICL_w	0	100	0	0

$K_T = 3,\ \beta = 0.9$						
$\hat{K}$	1	2	3	4	5	6
BIC^{EM}	0	0	100	0	0	0
ICL^{EM}	100	0	0	0	0	0
PLIC	0	0	100	0	0	0
BIC_p	0	0	2	8	22	68
ICL_p	0	0	100	0	0	0
BIC_w	0	0	100	0	0	0
ICL_w	0	24	76	0	0	0

$K_T = 4,\ \beta = 1$							
$\hat{K}$	1	2	3	4	5	6	7
BIC^{EM}	0	0	38	62	0	0	0
ICL^{EM}	100	0	0	0	0	0	0
PLIC	0	0	0	100	0	0	0
BIC_p	0	0	0	9	2	30	59
ICL_p	0	0	0	100	0	0	0
BIC_w	0	0	0	99	1	0	0
ICL_w	0	4	38	58	0	0	0

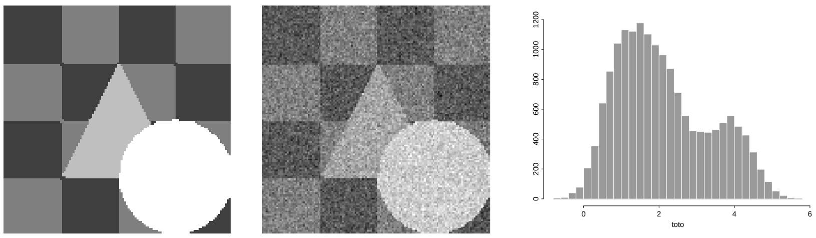
$K_T = 5,\ \beta = 1$								
$\hat{K}$	1	2	3	4	5	6	7	8
BIC^{EM}	0	0	0	79	21	0	0	0
ICL^{EM}	100	0	0	0	0	0	0	0
PLIC	0	0	0	0	100	0	0	0
BIC_p	0	0	0	0	0	2	23	75
ICL_p	0	0	5	21	74	0	0	0
BIC_w	0	0	0	0	92	8	0	0
ICL_w	0	0	68	27	5	0	0	0

$K_T = 6,\ \beta = 1.1$										
$\hat{K}$	1	2	3	4	5	6	7	8	9	10
BIC^{EM}	0	0	0	13	80	7	0	0	0	0
ICL^{EM}	100	0	0	0	0	0	0	0	0	0
PLIC	0	0	0	0	2	98	0	0	0	0
BIC_p	0	0	0	0	0	2	1	18	32	47
ICL_p	0	0	0	2	4	94	0	0	0	0
BIC_w	0	0	0	0	0	99	1	0	0	0
ICL_w	0	0	60	31	8	1	0	0	0	0

TAB. 6.1 – *sélection du nombre de classes dans une image simulée suivant un modèle de Potts puis dégradée par un bruit gaussien, pour différentes valeurs de K_T et de β. Les valeurs reportées dans les tableaux représentent le pourcentage du nombre d'expériences (parmi 100) où un nombre de classes donné a été sélectionné.*

6.3.2 Images avec bruit simulé

Le comportement des critères est maintenant étudié sur deux exemples d'images dégradées par un bruit gaussien (figures 6.3 (a) et (b) et 6.5 (a) et (b)) . À la différence des exemples présentés dans la partie précédente, les images originales ne sont pas des réalisations d'un modèle de champ de Markov. Il s'agit de l'image Damier et de l'image Logo, déjà présentées dans la partie 4.9, pour lesquelles le nombre de classes est connu et vaut respectivement $K_T = 4$ et $K_T = 2$. Dans le cas de l'image Damier, nous avons utilisé un bruit avec un écart-type, commun à chaque classe, égal à 0.5. Pour l'image Logo, les écarts-types des densités conditionnelles des observations dépendent de la classe et sont égaux respectivement à 130 et à 300. L'histogramme de la version bruitée de l'image Damier permet de distinguer de manière évidente au moins 2 classes (*cf.* figure 6.3 (c)). On devine également 2 classes à partir de l'histogramme de l'image Logo bruitée (*cf.* figure 6.5 (c)) : le pic et la queue de distribution.
L'étude de l'exemple de l'image Damier montre que l'ajout de l'information spatiale dans le modèle permet de sélectionner plus justement le nombre de classes. En effet, contrairement à BIC^{EM} et ICL^{EM}, tous les critères basés sur un modèle de champ de Markov caché retrouvent la valeur $K_T = 4$ (tableau 6.2). Sur cet exemple, l'attraction du modèle à 4 classes par rapport aux autres modèles est de même intensité quel que soit le critère utilisé pour un modèle spatial.
L'importance de la prise en compte de la composante spatiale apparaît aussi sur l'exemple de l'image Logo : ici BIC^{EM} sélectionne bien 2 classes mais ICL^{EM} n'en sélectionne qu'une. Par contre les deux critères ICL approximés, ICL_w et ICL_p, retrouvent bien 2 classes (tableau 6.2). Parmi les trois approximations de BIC étudiées pour un modèle de champ de Markov caché, BIC_p et BIC_w fournissent un bon résultat alors que PLIC surestime le nombre de classes pour l'image Logo. Les courbes de la figure 6.6 traduisent une attraction du modèle à $K_T = 2$ classes plus marquée avec les approximations de BIC et de ICL obtenues par approximation des fonctions de partition qu'avec celles basées sur une approximation de toute la distribution de Gibbs marginale.



(a) (b) (c)

FIG. 6.3 – *image Damier : (a) image originale, (b) image bruitée, (c) histogramme des données bruitées.*

Image Damier	
critère	$\hat{K}$
$\mathrm{BIC}^{\mathrm{EM}}$	3 (4)
$\mathrm{ICL}^{\mathrm{EM}}$	2
PLIC	4
BIC_p	4
ICL_p	4
BIC_w	4
ICL_w	4

Image Logo	
critère	$\hat{K}$
$\mathrm{BIC}^{\mathrm{EM}}$	2 (1,3)
$\mathrm{ICL}^{\mathrm{EM}}$	1
PLIC	3
BIC_p	2
ICL_p	2
BIC_w	2
ICL_w	2

TAB. 6.2 – *sélection du nombre de classes pour les images Damier et Logo.*

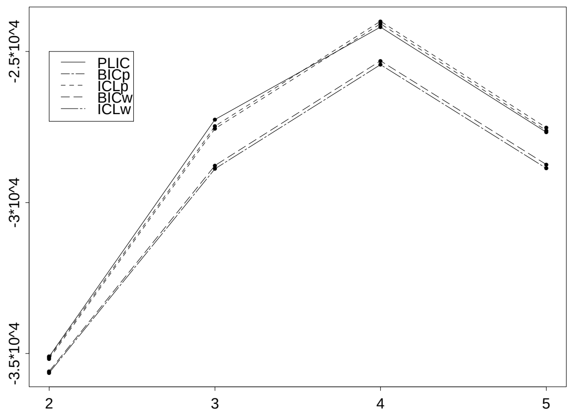
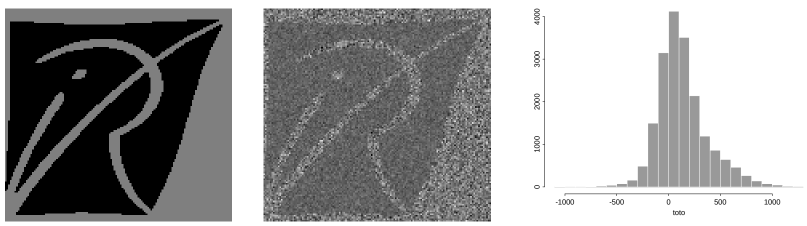


FIG. 6.4 – *courbes d'évolution, en fonction du nombre de classes, des critères BIC et ICL approximés par champ moyen et du critère PLIC dans le cas de l'image Damier.*



(a) (b) (c)

FIG. 6.5 – *image Logo : (a) image originale, (b) image bruitée, (c) histogramme des données bruitées.*

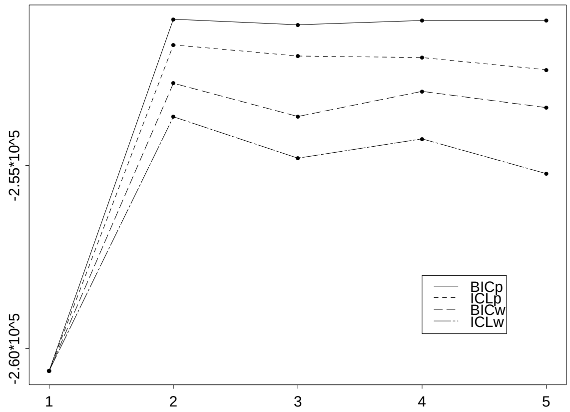
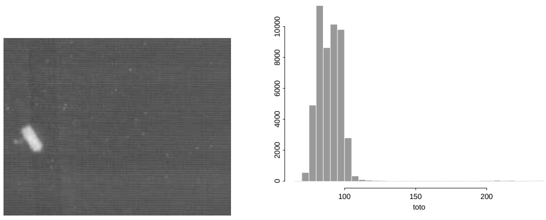


FIG. 6.6 – *courbes d'évolution, en fonction du nombre de classes, des critères BIC et ICL approximés par champ moyen dans le cas de l'image Logo.*

6.3.3 Images réelles

Cette dernière série d'expérimentations concerne des images réelles, pour lesquelles il n'exite pas de valeur exacte de K_T. Cependant il est possible d'obtenir un avis d'expert sur un nombre de classes raisonnable en fonction du contexte. L'objectif est de savoir si les critères considérés, basés sur les caractéristiques de l'image observée, sélectionnent un nombre de classes pertinent.
La première image (figure 6.7 (a)) est une image aérienne d'une bouée flottant sur une eau sombre. L'image est donc a priori composée de deux zones principales, information que l'on retrouve sur l'histogramme des observations (figure 6.7 (b)). Ce dernier distingue deux groupes principaux correspondant aux niveaux de gris inférieurs à 100 (l'eau) et aux niveaux de gris supérieurs à 100 (la bouée). Sur cet exemple nous avons observé de mauvaises performances des critères BIC_p, BIC_w, ICL_p et ICL_w basés sur les estimations fournies par l'algorithme en champ moyen ou en champ simulé : ils sélectionnent 4 classes. Ainsi, les résultats reportés dans le tableau 6.3 et sur la figure 6.8 pour ces critères ont été obtenus avec les estimations $\hat{\Psi}_K$ et les configurations $\tilde{\mathbf{z}}$ et $\hat{\mathbf{z}}$ fournies par l'algorithme ICM (contrairement à tous les autres exemples pour lesquels nous avons utilisé l'algorithme en champ simulé). De ce fait, $\mathrm{PLIC} = \mathrm{BIC}_p$. Les mêmes résultats restent valables si l'on utilise l'algorithme en champ modal.
On constate que seuls les critères $\mathrm{ICL}^{\mathrm{EM}}$, BIC_w et ICL_w sélectionnent 2 classes et parmi les critères spatiaux, ICL_w marque plus l'attraction du modèle à 2 classes (figure 6.8). Le bon comportement de BIC_w montre que la prise en compte de l'information spatiale dans le modèle pour le calcul de BIC vient corriger le résultat de $\mathrm{BIC}^{\mathrm{EM}}$. Le critère $\mathrm{ICL}^{\mathrm{EM}}$ a le même effet correcteur, dû cette fois à la prise en compte dans le critère de la pertinence de la segmentation correspondant au choix $\hat{K}$. Le critère ICL_w combine les deux informations. Le comportement parcimonieux de $\mathrm{ICL}^{\mathrm{EM}}$ se retrouvait déjà dans les exemples simulés de la partie 6.3.1. C'est un avantage ici car l'objectif est de choisir K de sorte à obtenir une segmentation en composantes homogènes facilement exploitables, contrairement au cas des images simulées où il s'agissait d'estimer un modèle.
La figure 6.9 (a) montre une image PET de poumons de chien. L'image est en fait une image circulaire, et la figure 6.9 (a) a été obtenue en remplissant le reste d'une grille carrée avec un niveau de gris constant, correspondant au pic de l'histogramme (figure 6.9 (b)). Une segmentation satisfaisante doit permettre de distinguer les poumons du reste de l'image dans le but de mesurer l'hétérogénéité du tissu dans la région d'intérêt. Les critères $\mathrm{BIC}^{\mathrm{EM}}$, BIC_w et ICL_w sélectionnent 3 classes avec la même évidence (tableau 6.3 et figure 6.11). Il est intéressant de noter que cela ne correspond pas aux mêmes segmentations. En effet, la segmentation obtenue pour un modèle de mélange fini indépendant estimé par EM retrouve le fond de l'image, la zone circulaire et les poumons. Lorsque K augmente, le disque disparaît de la segmentation. La segmentation obtenue pour un modèle de champ de Markov caché estimé par l'algorithme en champ simulé ne distingue pas le disque dans le fond de l'image mais retrouve deux zones dans les poumons (figure 6.10). Ce résultat n'est pas surprenant puisque les niveaux de gris de l'image reflètent l'intensité de l'émission radioactive du gaz présent dans les poumons. Les deux classes correspondent à une zone de haute densité en gaz,

à l'intérieur des poumons, et une zone plus périphérique de densité plus basse. Les solutions en trois classes fournies par EM indépendant et l'algorithme en champ simulé sont toutes les deux satisfaisantes, puisqu'elles permettent de distinger les contours des poumons. Lorsque K augmente, dans les segmentations obtenues par l'algorithme en champ simulé, il apparaît des rayons autour des poumons, résidus du disque original. D'un point de vue visuel, les segmentations correspondantes ne sont pas satisfaisantes. Les critères PLIC, BIC_p et ICL_p sélectionnent cependant 5 et 6 classes.



(a) (b)

FIG. 6.7 – *image Bouée: (a) image observée, (b) histogramme des observations.*

Image Bouée	
critère	$\hat{K}$
$\mathrm{BIC}^{\mathrm{EM}}$	4 (5,6)
$\mathrm{ICL}^{\mathrm{EM}}$	2
PLIC	4
ICL_p	4
BIC_w	2
ICL_w	2

Image Poumons	
criterion	$\hat{K}$
$\mathrm{BIC}^{\mathrm{EM}}$	3
$\mathrm{ICL}^{\mathrm{EM}}$	2
PLIC	6
BIC_p	5 (6)
ICL_p	5 (6)
BIC_w	3
ICL_w	3

TAB. 6.3 – *sélection du nombre de classes pour les images Bouée et Poumons.*

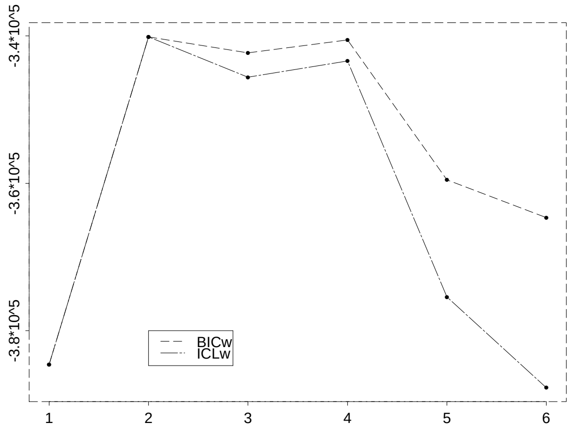


FIG. 6.8 – *courbes d'évolution, en fonction du nombre de classes, des critères BIC et ICL approximés par champ moyen dans le cas de l'image Bouée.*

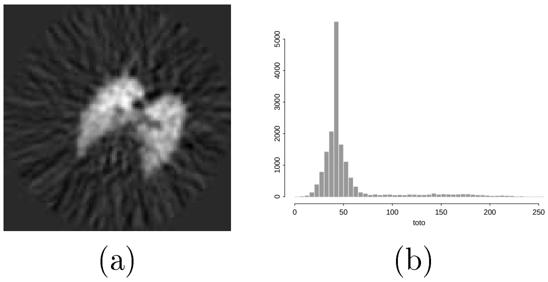


FIG. 6.9 – *image Poumons : (a) image observée, (b) histogramme des observations.*

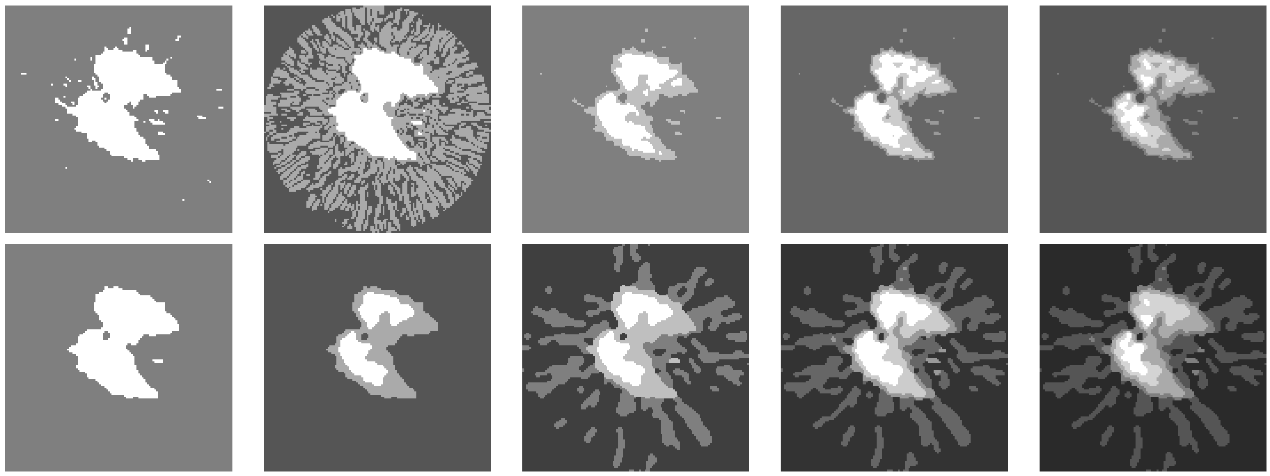


FIG. 6.10 – *segmentations en 2, 3, 4, 5 et 6 classes de l'image Poumons, par l'algorithme EM pour mélange fini indépendant (en haut) et l'algorithme en champ simulé (en bas).*

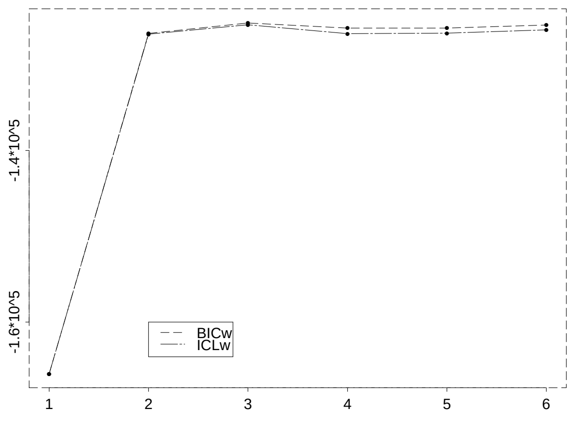


FIG. 6.11 – *courbes d'évolution, en fonction du nombre de classes, des critères BIC et ICL approximés par champ moyen dans le cas de l'image Poumons.*

6.3.4 Conclusion

Les résultats de ces expérimentations permettent de tirer plusieurs conclusions. Nous avons proposé en début de ce chapitre deux méthodes d'approximation des critères BIC et ICL pour champ de Markov caché basées sur le principe du champ moyen. Ces deux méthodes présentent l'avantage de conduire à des critères simples à évaluer. La première consiste à approximer toute la distribution de Gibbs marginale alors que la seconde approxime uniquement les fonctions de partition mises en jeu, ce qui permet d'introduire moins d'approximation dans les critères.
Nous avons observé de bonnes performances du critère BIC_w, obtenu par approximation de BIC par la seconde méthode, et ce notamment sur les exemples d'images réelles. En ce qui concerne ICL, si l'on considère que l'intérêt d'un critère de sélection se trouve dans le contexte d'images réelles, alors ICL_w, également obtenu par approximation uniquement des fonctions de partition, semble un critère prometteur.
Si l'on compare maintenant les performances des critères BIC_w et ICL_w avec celles des critères basés sur un modèle de mélange indépendant, $\mathrm{BIC}^{\mathrm{EM}}$ et $\mathrm{ICL}^{\mathrm{EM}}$, dans le cas des images non simulées, les résultats mettent en évidence l'intérêt de la prise en compte de la structure spatiale des données dans la sélection du nombre de classes de l'image. Si cela peut paraître évident dans le cas de critères exacts, il n'était pas certain que l'information spatiale soit préservée par les approximations faites. On remarque également que si dans les exemples présentés et pour un modèle de mélange fini indépendant, $\mathrm{BIC}^{\mathrm{EM}}$ sélectionne en général plus de composantes que $\mathrm{ICL}^{\mathrm{EM}}$, ce phénomène n'apparaît pas pour un modèle de champ de Markov caché entre les approximations BIC_w et ICL_w. Ainsi, on peut penser que l'ajout de l'hypothèse de dépendance markovienne entre les étiquettes de l'image, en favorisant des segmentations plus homogènes et plus lisses, peut avoir le même effet que la pénalisation par un terme d'entropie.
Ce bilan permet de mettre en avant l'intérêt du principe du champ moyen pour le calcul des critères BIC et ICL et notamment nous distinguons le critère BIC_w, basé sur l'approximation en champ moyen d'une fonction de partition d'une distribution de Gibbs. Néanmoins, il serait intéressant d'étudier comment les performances des critères proposés, et donc de BIC_w en particulier, peuvent dépendre de la méthode d'estimation des paramètres choisie.

Conclusion

Bilan

Les travaux présentés dans ce mémoire s'inscrivent dans la recherche actuelle en segmentation markovienne d'images. L'image observée et la segmentation cherchée sont considérées comme des variables aléatoires, réalisations d'un modèle de champ de Markov caché. Dans ce contexte, l'objectif global de cette thèse est de proposer des méthodes simples et peu coûteuses pour la segmentation non supervisée, dans un cas où la complexité du modèle oblige à passer par des approximations. Pour l'estimation du modèle de champ de Markov caché, nous avons adopté l'approche *mélange* qui consiste à voir le problème comme un problème de données manquantes, puisque la segmentation est inconnue. Dans ce contexte, nos contributions interviennent à plusieurs niveaux.
Tout d'abord, nous proposons une famille d'algorithmes, nommés algorithmes EM de type champ moyen, pour l'estimation des paramètres d'un champ de Markov caché. Ces algorithmes sont dérivés du principe d'approximation en champ moyen d'un champ de Markov. Ils présentent l'avantage d'être moins lourds que les versions de EM faisant intervenir des simulations de Monte-Carlo et évitent le problème de biais induit par les méthodes d'estimations de l'approche *image*. De plus, une étude sur des images simulées et réelles a montré l'intérêt de ces nouveaux algorithmes pour traiter la complexité de EM dans des problèmes où l'information spatiale est importante. En particulier, nous présentons un nouvel algorithme dans la famille proposée, l'algorithme en champ simulé, qui apparait comme très prometteur par ses bonnes performances et sa simplicité.
Nous présentons ensuite de nouveaux résultats sur la convergence de l'algorithme MCEM pour l'estimation des paramètres d'un champ de Markov caché, dans un cadre où peu de résultats sont disponibles du fait de la complexité du modèle.
Enfin, ces travaux sont complétés par un développement sur l'approximation de critères classiques pour le choix du nombre de classes dans un modèle de champ de Markov caché. Nous proposons des critères simples basés sur l'approximation en champ moyen, qui conduisent à des résultats satisfaisants en pratique dans le cas d'images simulées et réelles.
L'ensemble de ces travaux a permis d'une part de mieux reconnaître l'intérêt de l'approximation en champ moyen pour la segmentation markovienne non supervisée. D'autre part, nous avons développé à partir de ce principe une méthode cohérente et simple pour traiter les trois étapes de sélection, d'estimation et de segmentation pour l'analyse markovienne

d'images.

Perspectives

Dans la continuité de ce travail, nous voyons deux prolongements essentiels correspondant à une volonté de compléter la compréhension des algorithmes EM de type champ moyen et d'étendre ces procédures à des contextes plus complexes.
Le premier axe de recherche concerne l'étude du rôle effectif du paramètre spatial dans le modèle de Potts. Pour les algorithmes que nous proposons, il est estimé. Or, nous avons constaté au cours de nos expérimentations qu'une certaine variation dans l'estimation de ce paramètre ne conduit pas forcément à une segmentation différente. D'autre part, si l'estimation des paramètres du modèle de bruit par l'algorithme EM nécessite uniquement d'approcher la distribution conditionnelle du champ de Markov caché, l'estimation du paramètre spatial apporte une difficulté supplémentaire, puisque la distribution marginale doit également être approximée. Cette dernière étape vient alors augmenter la complexité des algorithmes EM de type champ moyen. Parmi les expériences publiées sur différentes méthodes de segmentation markovienne non supervisée, il est suggéré, au lieu d'estimer le paramètre spatial, soit de le fixer dans un intervalle pertinent (Zhang 1992, Forbes et Raftery 1998), soit de l'augmenter au cours de l'algorithme suivant un certain schéma d'évolution (Besag 1986). Avec cette dernière solution le paramètre spatial peut être interprété comme l'inverse d'une température, à la manière du recuit simulé. Il serait intéressant d'étudier comment les variations de ce paramètre influent sur l'allure de l'image segmentée, afin d'envisager d'utiliser de manière efficace les stratégies proposées par Besag (1986) et Zhang (1992). Il s'agirait alors de déterminer si ces solutions ne permettent pas d'obtenir des résultats aussi satisfaisants que dans le cas où le paramètre spatial est estimé, et ce à moindre coût.
Un élargissement possible de nos travaux consiste à utiliser des modèles de champs de Markov plus complexes. Nous avons considéré dans nos expérimentations un modèle de Potts simple qui, bien que montrant de bonnes performances dans beaucoup de situations, peut ne pas toujours saisir complètement les caractéristiques de l'image. Les méthodes que nous proposons pour l'estimation des paramètres s'étendent à des modèles plus complexes sans difficulté théorique supplémentaire et les critères proposés pour le choix du nombre de classes peuvent aussi être utilisés de manière plus générale pour la sélection de modèles markoviens. Nous pensons en particulier au modèle de Potts avec champ externe, pour lequel une étude des algorithmes EM de type champ moyen est en cours. En pratique cela entraîne des difficultés de calcul supplémentaires pour l'estimation des potentiels sur les singletons, mais cela devrait conduire à plus de flexibilité et à de meilleurs résultats dans le cas où les proportions dans l'image sont très contrastées. Nous pensons également à des modèles avec une structure de dépendances spatiales plus fine, comme ceux proposés et étudiés par Descombes, Mangin, Pechersky, et Sigelle (1995) et Tjelmeland et Besag (1998). À une échelle de complexité au dessus se situent les modèles markoviens multi-échelle, notamment illustrés dans Luettgen, Karl, Willsky, et Tenney (1993), Bouman et Shapiro (1994), Kato, Zerubia, et Berthod (1999) ou encore Chardin et Pérez (1999). Dans ces derniers cas, l'utilisation des approximations

de type champ moyen serait d'autant plus intéressante !

D'autre part, plusieurs moyens d'améliorer l'approximation "naïve" en champ moyen d'un champ de Markov ont été proposés en statistique physique, qui consistent à introduire une prise en compte des interactions d'ordre 2. Notamment, la méthode de Bethe et plus généralement l'approximation de Kikuchi ont été étudiees dans le cadre des modèles graphiques (Yedidia, Freeman, et Weiss 2001). Il serait intéressant d'étudier l'intérêt de telles approximations au niveau des algorithmes EM de type champ moyen et plus généralement pour la segmentation markovienne d'images.

Nous souhaitons enfin signaler que les résultats expérimentaux présentés s'appuient sur un travail de programmation. Un outil informatique a été développé en langage C, qui regroupe tous les algorithmes d'estimation et de segmentation que nous avons testés (ICM, MCEM, algorithmes EM de type champ moyen, EM pour modèle de mélange indépendant) ainsi que le calcul des critères de sélection BIC et ICL approchés par la méthode du champ moyen. Nous envisageons d'incorporer cet outil dans le logiciel MIXMOD, logiciel pour l'analyse statistique de mélanges de distributions de probabilité, actuellement en cours de développement au sein du projet IS2 de l'Inria Rhône-Alpes.

Annexe : Équation de cohérence du champ moyen pour le modèle de Potts

Dans le cas du modèle de Potts avec champ externe, le problème de point fixe associé à l'approximation en champ moyen, $\mathbf{z}^{mf} = h(\mathbf{z}^{mf})$, est défini par les équations suivantes

$$\forall i \in S \ , \quad z_i^{mf} \quad = \quad \frac{\sum\limits_{z_i} z_i \exp\{V_i(z_i) + \beta z_i^t \sum\limits_{j \in N(i)} z_j^{mf}\}}{\sum\limits_{z_i} \exp\{V_i(z_i) + \beta z_i^t \sum\limits_{j \in N(i)} z_j^{mf}\}} \ . \tag{6.13}$$

Si les potentiels sur les singletons sont nuls ou égaux, quelle que soit la valeur du paramètre spatial β la configuration $\mathbf{u}^K$ uniforme en chaque site, définie par $u_i^K = (K^{-1}, \ldots, K^{-1})^t$ pour tout i dans S, est solution de cette équation. Elle disparaît dès que les potentiels sur les singletons sont non nuls et dépendent de la classe. Plusieurs questions théoriques importantes se posent autour de la résolution du problème de point fixe. La solution $\mathbf{u}^K$ est-elle stable? Existe-t-il toujours une solution pour le modèle de Potts avec champ externe? Si oui, y-a-t-il unicité de la solution? Enfin, est-ce que la procédure itérative de recherche du point fixe $\mathbf{z}^{q+1} = h(\mathbf{z}^{q+1})$ converge? Les propriétés établies dans cette annexe fournissent des éléments de réponse à ces questions.

Stabilité de la solution uniforme

Proposition 2 *Soit N_v le nombre de voisins d'un pixel, supposé identique pour tous les pixels. Dans le cas du modèle de Potts simple et pour un système de voisinage d'ordre 1 ou 2, quel que soit le nombre K de couleurs, la solution uniforme $\mathbf{u}^K$ est stable si et seulement si $\beta \leq \frac{K}{Nv}$.*

Démonstration : l'équation de cohérence associée au principe du champ moyen est issue du problème de minimisation de la distance de Kullback-Leibler entre une distribution factorisable $\mathbf{c}$ définie par $\mathbf{c} = \{c_{ik},\ i \in S,\ k \in \{1, \ldots, K\}\}$ et la distribution de Gibbs $P_G(\mathbf{z})$ (voir section 3.2). Plus précisement, nous avons l'équivalence suivante

$$\frac{\partial KL(\mathbf{c}, P_G)}{\partial \mathbf{c}} = 0 \quad \Leftrightarrow \quad \mathbf{c} = h(\mathbf{c}).$$

Cela signifie que l'ensemble des points stationnaires de la fonction $U(\mathbf{c}) = KL(\mathbf{c},P_G)$ est l'ensemble $\mathcal{S} = \{\mathbf{c} \ / \ \mathbf{c} = h(\mathbf{c})\}$. Pour montrer la stabilité du point stationnaire $\mathbf{u}^K$ (c'est-à-dire qu'il s'agit bien d'un minimum local) il suffit de montrer que la matrice hessienne de U au point $\mathbf{u}^K$ est définie positive. Pour cela nous appliquons les résultats de Elfadel et Yuille (1992), qui étudient les solutions du problème

$$\mathbf{c} \quad = \quad \arg\min_{\mathbf{c}} U(\mathbf{c}), \tag{6.14}$$

dans le cas d'une distribution de Gibbs définie sur une grille $R \times R$ par une énergie de la forme $H(\mathbf{z}) = \beta \sum_{i,j=1}^{R} \sum_{k,l=1}^{K} J_{ij} M_{kl} z_{ik} z_{jl}$. La matrice $J = [J_{ij}]$ est une matrice de taille $R^2 \times R^2$ dont les coefficients vérifient $J_{ij} = 1$ si $i \sim j$ et $J_{ij} = 0$ sinon. La matrice $M = [M_{kl}]$ est la matrice $K \times K$ dont les coefficients donnent l'intensité de l'interaction entre les pixels. Elfadel et Yuille (1992) remarquent que pour un modèle de champ de Markov à K couleurs défini par une telle distribution de Gibbs, $\mathbf{u}^K$ est toujours une solution triviale du problème (6.14). Le modèle de Potts simple est un cas particulier pour lequel les coefficients de M vérifient $M_{kl} = -1/2$ si $k = l$ et $M_{kl} = 0$ sinon. La matrice J dépend du système de voisinage choisi. Elfadel et Yuille (1992) donnent alors la forme des valeurs propres de la matrice hessienne de U au point $\mathbf{u}^K$, qui fait intervenir la matrice J.
Nous introduisons d'abord une notation pour les matrices particulières dont chaque ligne est une permutation de la première. Pour tout R-uplet $(a_1,a_2,\dots,a_R)$, la notation $Circulant(a_1,a_2,\dots,a_R)$ représente la matrice $R \times R$

$$A \quad = \quad \begin{bmatrix} a_1 & a_2 & \dots & \dots & a_R \\ a_R & a_1 & a_2 & \dots & a_{R-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_3 & a_4 & \dots & \dots & a_2 \\ a_2 & a_3 & \dots & \dots & a_1 \end{bmatrix}.$$

La notation s'étend au cas où $(a_1,a_2,\dots,a_R)$ est un R-uplet de matrices. Pour un voisinage d'ordre 1, Elfadel et Yuille (1992) donnent

$$J = Circulant(J_1,J_2,\dots,J_R),$$

avec

$$\begin{aligned} J_i &= 0, \forall i \neq 1,\ 2,\ R, \\ J_1 &= Circulant(0,1,0,\dots,0,1), \\ J_2 &= Circulant(1,0,\dots,0), \\ J_R &= Circulant(1,0,\dots,0). \end{aligned}$$

Si l'on note J_{ij} la $j^{\text{ième}}$ composante de la première ligne de la matrice J_i, d'après le théorème 2 de Elfadel et Yuille (1992), les R^2 valeurs propres de la matrice hessienne de U au point

$\mathbf{u}^K$ sont données par

$$\lambda_{pq} \;=\; -\hat{c}_{pq} + \frac{K}{\beta}, \quad 1 \leq p,q \leq R, \tag{6.15}$$

avec $\{\hat{c}_{pq} = \sum_{i,j=1}^{R} J_{ij}\, \xi_R^{p(i-1)} \xi_R^{q(j-1)}, 1 \leq p,q \leq R\}$, l'ensemble des coefficients de la transformée de Fourier discrète 2D de $\{J_{ij}, 1 \leq i,j \leq R\}$ et $\xi_R = e^{i\frac{2\pi}{R}}$. Les coefficients $\hat{c}_{pq}$ se simplifient,

$$\begin{aligned}\forall\, 1 \leq p,q \leq R, \quad \hat{c}_{pq} &= \xi_R^q + \xi_R^{q(R-1)} + \xi_R^p + \xi_R^{p(R-1)} \\ &= 2\cos(\frac{2\pi q}{R}) + 2\cos(\frac{2\pi p}{R}).\end{aligned}$$

Ainsi, les valeurs propres sont toutes positives si et seulement si $\beta \leq \frac{K}{4}$, et dans ce cas le point fixe $\mathbf{u}^K$ est stable.
Pour un voisinage d'ordre 2, la matrice J devient

$$\begin{aligned} J_i &= 0, \forall i \neq 1,\, 2,\, R, \\ J_1 &= Circulant(0,1,0,\ldots,0,1), \\ J_2 &= Circulant(1,1,0,\ldots,0,1), \\ J_R &= Circulant(1,1,0,\ldots,0,1). \end{aligned}$$

Les valeurs propres sont données par (6.15) avec cette fois

$$\begin{aligned}\hat{c}_{pq} &= \xi_R^q + \xi_R^{q(R-1)} + \xi_R^p + \xi_R^{p(R-1)} + (\xi_R^q + \xi_R^{q(R-1)})(\xi_R^p + \xi_R^{p(R-1)}) \\ &= 2\cos(\frac{2\pi q}{R}) + 2\cos(\frac{2\pi p}{R}) + 4\cos(\frac{2\pi q}{R})\cos(\frac{2\pi p}{R}).\end{aligned}$$

Dans ce cas, la valeur seuil est $\beta = \frac{K}{8}$.

□

Unicité du point fixe

Proposition 3 *Soit N_v le nombre de voisins d'un pixel, supposé identique pour tous les pixels. Pour un modèle de Potts avec champ externe à K couleurs, si $\beta < \frac{4}{N_v K}$ alors la fonction h est une contraction et admet un unique point fixe $\mathbf{z}^{mf}$. De plus la suite $\mathbf{z}^{q+1} = h(\mathbf{z}^{q+1})$ converge vers $\mathbf{z}^{mf}$.*

Démonstration : la fonction h est une fonction de $\mathcal{V}^n$ dans $\mathcal{V}^n$ dont les composantes h_{ik} sont données pour tout $\mathbf{z}$ par

$$\forall i \in S, \forall k \in \{1,\ldots,K\}, \quad h_{ik}(\mathbf{z}) = \frac{\exp\{V_i(e_k) + \beta e_k^t \sum\limits_{j\in N(i)} z_j\}}{\sum_{l=1}^K V_i(e_l) + \beta e_l^t \sum\limits_{j\in N(i)} z_j\}} = P_G(z_i = e_k \mid \mathbf{z}_{N(i)},\beta).$$

Les dérivées partielles $\frac{\partial h_{ik}}{\partial z_{jl}}(\mathbf{z})$ existent et sont continues pour tout $i,j \in S$, et $k \in \{1,\ldots,K\}$. Elles sont données par

$$\frac{\partial h_{ik}}{\partial z_{jl}}(\mathbf{z}) = \begin{cases} 0,\ si\ j \notin N(i) \\ \beta h_{ik}(\mathbf{z})(1 - h_{ik}(\mathbf{z})),\ si\ j \in N(i)\ et\ k = l \\ -\beta h_{ik}(\mathbf{z}) h_{il}(\mathbf{z}),\ si\ j \in N(i)\ et\ k \neq l \end{cases}. \tag{6.16}$$

La fonction h est contractante si la norme de la matrice jacobienne $h'(\mathbf{z}) = [\frac{\partial h_{ik}}{\partial z_{jl}}(\mathbf{z})]$ est strictement inférieure à 1. Dans ce cas, d'après le théorème du point fixe, la proposition est démontrée. Considérons la norme infinie de la matrice $h'(\mathbf{z})$,

$$\|h'(\mathbf{z})\|_\infty = \max_{i,k} \sum_{j\in S} \sum_{l=1}^K |\frac{\partial h_{ik}}{\partial z_{jl}}(\mathbf{z})|\,.$$

On suppose que le paramètre spatial est positif. Pour tout $i \in S$ et $k \in \{1,\ldots,K\}$, nous avons

$$\sum_{j\in S} \sum_{l=1}^K |\frac{\partial h_{ik}}{\partial z_{jl}}(\mathbf{z})| = \sum_{j\in N(i)} \beta h_{ik}(\mathbf{z})(1 - h_{ik}(\mathbf{z})) + \sum_{j\in N(i)} \sum_{l\neq k} \beta h_{ik}(\mathbf{z}) h_{il}(\mathbf{z}).$$

L'étude de la fonction $x(1-x)$ permet de montrer l'inégalité suivante

$$\forall i \in S, \forall k \in \{1,\ldots,K\}, \quad h_{ik}(\mathbf{z})(1 - h_{ik}(\mathbf{z})) \leq \frac{1}{4}.$$

Puisque $P_G(Z_i = e_l \mid \mathbf{z}_{N(i)},\beta) < 1 - P_G(Z_i = e_k \mid \mathbf{z}_{N(i)},\beta)$, nous avons aussi l'inégalité

$$\forall i \in S, \forall k \in \{1,\ldots,K\}, \quad h_{ik}(\mathbf{z}) h_{il}(\mathbf{z}) < \frac{1}{4}.$$

Finalement,

$$\sum_{j\in S} \sum_{l=1}^K |\frac{\partial h_{ik}}{\partial z_{jl}}(\mathbf{z})| < \beta \frac{N_v K}{4}.$$

On en déduit que si $\beta < \frac{4}{N_v K}$, la fonction h est contractante.

$\square$

Ce résultat est valable pour des potentiels sur les singletons quelconques. Ainsi, on en déduit que pour une valeur de β inférieure à $\frac{4}{N_v K}$, l'unique solution de l'équation de cohérence associée au modèle de Potts simple est la solution uniforme $\mathbf{u}^K$ et la procédure itérative de recherche du point fixe converge bien vers cette solution. Pour un modèle à 2 classes, les deux seuils $\frac{4}{N_v K}$ et $\frac{K}{Nv}$ sont égaux. Pour un nombre de classes $K \geq 3$, lorsque $\frac{4}{N_v K} < \beta \leq \frac{K}{Nv}$, les résultats établis permettent seulement d'assurer la stabilité de cette solution.

Publications

Groupes de travail

Celeux, G., F. Forbes et N. Peyrard (1999). Mean field approximation principle and Markov random fields model-based image segmentation. *Model-Based Clustering, Spatial Data Analysis, and Bayesian Model Selection Workshop, Seattle - USA.*

Celeux, G., F. Forbes et N. Peyrard (2000). Mean field approximation methods for parameter estimation and model selection in hidden Markov models for images. *Model-Based Clustering and Applications Workshop, Seattle - USA.*

Forbes, F. et N. Peyrard (2001). Mean field like approximations for choosing the number of components in hidden Markov models. *Model-Based Clustering and Applications Workshop, Seattle -USA.*

Conférences nationales

Celeux, G., F. Forbes et N. Peyrard (1999). Approximation du champ moyen et segmentation d'images. *XXXIème Journées de Statistique, Grenoble - France.*

Celeux, G., F. Forbes et N. Peyrard (2000). Critères pour la sélection d'un modèle de champ de Markov caché. *XXXIIème Journées de Statistique, Fès - Maroc.*

Conférences internationales

Celeux, G., F. Forbes et N. Peyrard (1999). Mean field approximation principle and Markov random fields model-based image segmentation. *Second European Conference on Highly Structured Stochastic Systems, Pavia - Italy.*

Celeux, G., F. Forbes et N. Peyrard (2000). Mean field approximation principle for parameter estimation in hidden Markov models. *First European Conference on Spatial and Computational Statistics, Ambleside - England.*

Rapports de recherche

Celeux, G., F. Forbes et N. Peyrard (2001). EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation, Rapport de Recherche Inria, no. 4105 (soumis à *Pattern Recognition*).

Peyrard, N. (2001). Convergence of MCEM and Related Algorithms for Hidden Markov Random Field, Rapport de Recherche Inria, no. 4146.

Forbes F., C. Fraley, D. Georgian-Smith, D. Goldhaber, N. Peyrard, A. Raftery (2001) Region-Of-Interest Selection and Statistical Analysis of Dynamic Breast Magnetic Resonance Imaging Data, Rapport de Recherche Inria, no. 4249 (soumis à *IEEE Transactions on Médical Imaging*).

Références

Abend, K., T. Harley, et L. Kanal (1965). Classification of binary patterns. *IEEE Transactions on Information Theory 11*, 538–544.

Aitkin, M. et D. B. Rubin (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society B 47*, 67–75.

Akaike, M. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox et F. Caski (Eds.), *Second International Symposium on Information Theory*, pp. 267.

Ambroise, C. (1996). *Approche probabiliste en classification automatique et contraintes de voisinage.* Thèse de Doctorat, Université de Technologie de Compiègne, France.

Ambroise, C., V. M. Dang, et G. Govaert (1997). Clustering of spatial data by the EM algorithm. In A. Soares, J. Gómez-Hernandez, et R. Froidevaux (Eds.), *geoENV I - Geostatistics for Environmental Applications, Quantitative Geology and Geostatistics*, Volume 9, pp. 493–504. Kluwer Academic Publisher.

Ambroise, C. et G. Govaert (1998). Convergence proof of an EM-type algorithm for spatial clustering. *Pattern Recognition Letters 19*, 919–927.

Archer, G. E. B. et D. M. Titterington (2000). Parameter estimation for hidden Markov chains. *To appear in Journal of Statistical Planning Inference*.

Barker, S. A. et P. J. W. Rayner (1997). Unsupervised image segmentation using Markov random field models. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 165–178.

Baum, T. P., G. Soules, et N. Weiss (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics 41*(1), 164–171.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B 35*, 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24*, 179–195.

Besag, J. (1983). Discussion on a paper by P. Switzer. *Bulletin of International Statistical Institute 47*, 77–92.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B 48*(3), 259–302.

Besag, J., J. York, et A. Mollié (1991). Bayesian image restoration with two applications in spatial statistics, with discussion. *Annals of the Institute of Statistical Mathematics 43*(1), 1–59.

Bethe, H. A. (1935). *Proceedings of the Royal Society of London*, 150–552.

Biernacki, C., G. Celeux, et G. Govaert (2000). Assessing a mixture model for clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 719–725.

Biscarat, J. C. (1994). Almost sure convergence of a class of stochastic approximation algorithms. *Stochastic Processes and their Applications 50(1)*, 83–100.

Bouman, C. et M. Shapiro (1994). A multiscale random field model for Bayesian image segmentation. *IEEE Transactions on Image Processing 3*(2), 162–177.

Broniatowski, M., G. Celeux, et J. Diebolt (1983). Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data Analysis and Informatics 3*, 359–374.

Celeux, G. (1990). *Analyse discriminante sur variables continues* (Collection didactique, Inria ed.), Volume 7.

Celeux, G. et J. Diebolt (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly 2*(1), 73–82.

Celeux, G. et J. Diebolt (1990). Une version de type recuit simulé de l'algorithme EM. *Comptes Rendus de l'Académie des Sciences - Série I - Mathématique 310*, 119–124.

Celeux, G. et J. Diebolt (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports 41*, 119–134.

Celeux, G. et J. Diebolt (1993). Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Stochastic Models 9*(4), 599–613.

Celeux, G., F. Forbes, et N. Peyrard (2001). EM procedures using mean field-like approximations for Markov model-based image segmentation. Rapport de recherche 4105, Inria.

Celeux, G. et G. Govaert (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis 14*, 315–332.

Chalmond, B. (1988). Image restoration using an estimated Markov model. *Signal Processing 15*, 115–129.

Chalmond, B. (1989). An iterative gibbsian technique for reconstruction of m-ary images. *Pattern Recognition 22*(6), 747–761.

Chalmond, B. (2000). *Éléments de modélisation pour l'analyse d'images.* Mathématiques et Applications 33 - Springer-Verlag Berlin.

Chan, K. S. et J. Ledolter (1995). Monte Carlo estimation for time series models involving counts. *Journal of the American Statistical Association - Theory and Methods 90*(429), 242–252.

Chandler, D. (1987). *Introduction to Modern Statistical Mechanics.* Oxford University Press.

Chardin, A. et P. Pérez (1999). Unsupervised image classification with a hierarchical em algorithm. In *IEEE International Conference on Computer Vision*, Volume 2, pp. 969–974.

Chellapa, R., S. Chatterjee, et R. Bagdazian (1985). Texture synthesis and compression using Gaussian-Markov random fields models. *IEEE Transactions on Systems Man and Cybernetics, SMC 15*, 298–303.

Cocquerez, J.-P. et S. Philipp (1995). *Analyse d'images : filtrage et segmentation*. Masson.

Comer, M. L. et E. J. Delp (2000). The EM/MPM algorithm for segmentation of textures images: Analysis and further experimental results. *IEEE Transactions on Image Processing 9*(10), 1731–1744.

Dang, M. V. (1998). *Classification de données spatiales : modèles probabilistes et critères de partitionnement*. Thèse de Doctorat, Université de Technologie de Compiègne.

Dang, M. V. et G. Govaert (1998). Spatial fuzzy clustering using EM and Markov random fields. *International Journal of Systems Research and Information Science 8*(2-3), 183–202.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Descombes, X., J.-F. Mangin, E. Pechersky, et M. Sigelle (1995). Fine structures preserving Markov model for image processing. In *9th Scandinavian Conference on Image Analysis*.

Descombes, X., R. Morris, et J. Zerubia (1996). Quelques améliorations à la segmentation d'images bayésienne. Rapport de recherche 2916, Inria.

Descombes, X., R. Morris, J. Zerubia, et M. Berthod (1999). Estimation of Markov random field prior parameters using Markov Chain Monte Carlo maximum likelihood. *IEEE Transactions on Image Processing 8*(7), 954–963.

Dobrushin, R. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability and its Applications 13*, 197–224.

Dunmur, A. et D. Titterington (1997). On a modification to the mean field EM algorithm in factorial learning. In M. Mozer, M. Jordan, et P. T. (Eds.), *Advances in Neural Information Processing Systems*, Volume 9, pp. 431–437. MIT Press.

Dunmur, A. et D. Titterington (1998). Mean fields and two-dimensional Markov random fields in image analysis. *Pattern Analysis and Applications 1*, 248–260.

Elfadel, I. et A. Yuille (1992). Mean-field phase transitions for Gibbs random fields. In S. shing Chen (Ed.), *SPIE- Neural and Stochastic Methods in Image and Signal Processing*, Volume 1766, pp. 257–268.

Forbes, F., C. Fraley, D. Georgian-Smith, D. Goldhaber, N. Peyrard, et A. Raftery (2001). Region-of-interest selection and statistical analysis of dynamic breast magnetic resonance imaging data, Rapport de recherche 4249, Inria.

Forbes, F. et N. Peyrard (2001). Mean field like approximations for choosing the number of components in hidden Markov models. In *Model-Based Clustering and Applications Workshop*, Seattle - USA.

Forbes, F. et A. Raftery (1998). Bayesian morphology: fast unsupervised bayesian image analysis. Rapport de recherche 3374, Inria.

Forsythe, G., M. Malcolm, et C. Moler (1980). *Computer methods for mathematical computations*, pp. 180 de l'édition russe.

Fort, G. (2001). *Contrôle d'ergodicité de chaîne de Markov : application à l'analyse de convergence de l'algorithme Monte-Carlo EM.* Thèse de Doctorat, Université PARIS 6 Pierre et Marie Curie.

Fraley, C. et A. Raftery (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal 41*, 578–588.

Geiger, D. et F. Girosi (1991). Parallel and deterministic algorithms from MRFs: Surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*(5), 401–412.

Geman, S. et D. Geman (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

Geman, S. et C. Graffigne (1987). Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, Berkley.

Graffigne, C., F. Heitz, P. Perez, F. Preteux, M. Sigelle, et J. Zerubia (1995). Hierarchical Markov random field models applied to image analysis: a review. In *SPIE Conference on neural, morphological, stochastics methods in image and signal processing.*

Green, P. (1995). Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–32.

Guyon, X. (1993). *Champs aléatoires sur un réseau : modélisation, statistique et application.* Masson.

Hathaway, R. (1986). Another interpretation of th EM algorithm for mixture distributions. *Journal of Statistics and Probability Letters 4*, 53–56.

Higdon, D. (1993). Discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B 55*, 78.

Humphreys, K. et D. M. Titterington (2000). Improving the mean field approximation in belief networks using Bahadur's reparameterisation of the multivariate binary distribution. *Neural Processing Letters 12*, 183–197.

Jaakkola, T. S. et M. Jordan (1997). Improving the mean field approximation via the use of mixture distributions. In *NATO ASI on Learning in graphical models*, pp. 163–173. Dordrencht, Kluwer Academic Publishers.

Jain, A. (1981). Advances in mathematical models for image processing. In *Proceedings of the IEEE*, Volume 69, pp. 502–528.

Ji, C. et L. Seymour (1996). A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *Annals of Applied Probability 6*, 423–443.

Kappen, H. et F. Rodriguez (1998). Efficient learning in Boltzmann machines using linear response theory. *Neural Computation 10*(5), 1137–1156.

Kass, R. et A. Raftery (1995). Bayes factors. *Journal of the American Statisticial Association 90*, 733–795.

Kato, Z. (1999). Bayesian color image segmentation using Reversible Jump Markov Chain Monte Carlo, Rapport de Recherche ERCIM R055.

Kato, Z., J. Zerubia, et M. Berthod (1999). Unsupervised parallel image classification using Markovian models. *Pattern Recognition 32*, 591–604.

Lakshmanan, S. et H. Derin (1989). Simultaneous parameter estimation and segmentation of Gibbs random fields using Simulated Annealing. *Transactions on Pattern Analysis and Machine Intelligence 11*(8), 799–813.

Leahy, R. M., J. Qi, et Z. Zhou (1997). Approximate maximum likelihood hyperparameter estimation for Gibbs priors. *IEEE Transactions on Image Processing 6*(6), 844–861.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics 20*, 1350–1360.

Lévy, P. (1948). Chaînes doubles de Markov et fonctions aléatoires de deux variables. *Comptes Rendus de l'Académie des Sciences - Série I 226*, 53–55.

Li, S. Z. (1995). *Markov random field modeling in computer vision.* Springer.

Luettgen, M. R., W. Karl, A. S. Willsky, et A. R. Tenney (1993). Multiscale representations of Markov random fields. *IEEE Transactions on Signal Processing 41*(12), 3377–3396.

Marroquin, J., S. Mitter, et T. Poggio (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association 82*, 76–89.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley series in probability and mathematical statistics - New York.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, et E. Teller (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics 21*, 1087–191.

Morris, R. D. (1999). Auxilliary variables for markov random fields with higher order interactions. In *Energy minimization Methods in computer vision and Pattern Recognition*, Number 1654 in Lecture Notes in Computer Science, pp. 131–141. Springer.

Neal, R. M. et G. E. Hinton (1998). A view of the EM algorithm that justifies incremental, sparse and other variants. In *Learning in graphical models*, pp. 355–368. Dordrencht, Kluwer Academic Publishers.

Nielsen, S. F. (2000). The asymptotic EM algorithm: estimation and asymptotic results. *Bernoulli 6*(3), 457–489.

Opper, M. et O. Winther (2000). From naive mean field theory to the TAP equations. In *Advances in mean field methods: Theory and Practice.* MIT Press.

Parisi, G. (1988). *Statistical Field Theory.* Addison-Wesley.

Peterson, C. et J. Anderson (1987). A mean field theory learning algorithm for neural networks. *Complex Systems 1*, 995–1019.

Peyrard, N. (2001). Convergence of MCEM and related algorithms for hidden Markov random field. Rapport de recherche 4146, Inria.

Qian, W. et D. M. Titterington (1991). Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London, Series A 337*, 407–428.

Qian, W. et D. M. Titterington (1992). Stochastic relaxations and EM algorithms for Markov random fields. *Journal of Statistical Computation and Simulation 40*, 55–69.

Redner, R. et H. Walker (1984). Mixtures densities, maximum likelihood and the EM algorithm. *Society for Industrial and Applied Mathematics Review 26*(2), 195–239.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica 14*, 465–471.

Robert, C. P. et G. Casella (1999). *Monte Carlo statistical methods*. Springer.

Roeder, K. et L. A. Wasserman (1997). Practical Bayesian density estimation using mixtures of Normals. *Journal of the American Statistical Association 92*, 894–902.

Rydén, R. et D. M. Titterington (1998). Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics 7*(2), 194–211.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Serra, J. (1982). *Image Analysis and Mathematical Morphology*. Academic Press.

Seymour, L. et C. Ji (1996). Approximate Bayes model selection procedures for Gibbs-Markov random fields. *Journal of Statistical Planning and Inference 51*, 75–97.

Stanford, D. et A. Raftery (2001). Determining the number of colors or gray levels in an image using approximate Bayes factors: The pseudolikelihood information criterion (PLIC). Technical Report 390, University of Washington.

Swendsen, R. et J.-S. Wang (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters 58*, 86–88.

Switer, P. (1969). Mapping a geographically correlated environment (with discussion). In *International Symposium on Statistical Ecology: Spatial Patterns and Statistical Distribution*, Volume 1, pp. 235–269.

Thouless, D., P. Anderson, et R. Palmer (1977). A solution to a "solvable" model of spin glass. *Philosophical Magazine 35*, 593.

Titterington, D. M. (1984). Comments on a paper by S. L. Sclove. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6), 656–658.

Titterington, D. M., A. F. Smith, et U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley series in probability and mathematical statistics - Chichester.

Tjelmeland, H. et J. Besag (1998). Markov random field with higher-order interactions. *Scandinavian Journal of Statistics 25*, 415–433.

Wei, G. et M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's Data Augmentation algorithms. *Journal of the American Statistical Association 85*(411), 699–704.

Weiss, Y. (2000). Comparing the mean field method and belief propagation for approximate inference in MRFs. In *Advances in mean field methods: Theory and Practice.* MIT Press.

Winkler, G. (1995). *Image analysis, random fields and dynamics Monte Carlo methods : a mathematical introduction.* Springer.

Wu, C. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics 11*, 95–103.

Wu, C.-H. et P. C. Doerschuk (1995). Cluster expansions for the deterministic computation of Bayesian estimators based on Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17*(3), 275–293.

Yedidia, J. S., W. T. Freeman, et Y. Weiss (2001). Bethe free energy, kikuchi approximations and belief propagation algorithms. Technical report, MERL.

Younes, L. (1988). Estimation and annealing for Gibbsian fields. *Annales de l'Institut Henri Poincaré 2*, 269–294.

Younes, L. (1989). Parameter estimation for unperfectly observed Gibsian fields. *Probability Theory and Related Fields 82*, 625–645.

Yuille, A. L. (2001). A Double-Loop algorithm to minimize the Bethe free energy. In M. Figueiredo, J. Zerubia, et A. K. Jain (Eds.), *Energy minimization Methods in computer vision and Pattern Recognition*, Lecture notes in Computer Science, pp. 3–18. Springer.

Zerubia, J. et R. Chellappa (1993). Mean field approximation using compound Gauss-Markov random field for edge detection and image estimation. *IEEE Transactions on Neural Networks 4*(4), 703–709.

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. *IEEE Transactions on Signal Processing 40*(10), 2570–2583.

Zhang, J. (1993). The mean field theory in EM procedures for blind Markov random field image restoration. *IEEE Transactions on Signal Processing 2*(1), 27–40.

TITLE:

Mean field-like approximations of Markov random fields for spatial data segmentation

ABSTRACT:

This work is part of the current research on Markov model-based segmentation of spatial data. It concerns problems such as model selection, parameter estimation and segmentation. The methods generally used to deal with this three steps cannot be applied directly in the context of hidden Markov fields due to the dependence structure in the model. Approximations are necessary. A solution, quite computationally expensive, is to use Monte Carlo simulations. We propose simple and less expensive methods, based on the *mean field approximation* of a Markov field. This principle allows to work with systems of independent variables, simpler to deal with.

In the context of parameter estimation, we studied, and generalized the mean field principle to make the EM algorithm tractable. A class of algorithms is derived, the *mean field-like* EM algorithms, which have the advantage of taking into account the Markovian structure while preserving the simplicity of EM in the independent case. Among these procedures, we point out the *simulated field algorithm*, related to the Monte Carlo EM (MCEM), for its performance in terms of estimation and segmentation on synthetic and real images.

Then, we present sufficient conditions for the almost sure convergence of MCEM, in a context where few results are available due to the complexity of the Markov model.

These results on estimation and segmentation in the case where the number of classes in the segmentation is known, are completed with a work on the problem of Markov model selection. For this problem as well, the use of the mean field principle is interesting since it leads to simple approximations of classical criteria (Bayesian Information Criterium, Integrated Completed Likelihood) which appear quite promising in practice.

RESUME :
Ce travail s'inscrit dans la recherche actuelle en segmentation markovienne d'images. Cela concerne les problèmes de sélection de modèle, d'estimation des paramètres et de segmentation. Les méthodes classiques pour traiter ces trois étapes ne peuvent être appliquées directement dans le cas des champs de Markov cachés du fait de la structure de dépendance spatiale du modèle. Des approximations sont nécessaires. Une solution, assez coûteuse, consiste à utiliser des simulations de Monte-Carlo. Nous proposons des méthodes simples et moins lourdes, basées sur l'approximation en champ moyen d'un champ de Markov. L'intérêt de cette méthode est de se ramener à un système de variables indépendantes, plus facile à manipuler.
Dans un premier temps nous avons étudié, puis généralisé le principe du champ moyen pour la mise en œuvre de l'algorithme EM pour l'estimation des paramètres du modèle. Il en résulte une classe d'algorithmes, *les algorithmes* EM *de type champ moyen*, qui présentent l'avantage de prendre en compte la structure markovienne tout en ayant une simplicité de mise en œuvre identique au cas indépendant. Parmi ces procédures, nous distinguons *l'algorithme en champ simulé*, proche de l'algorithme Monte-Carlo EM (MCEM), en raison de ses performances en termes d'estimation et de segmentation dans le cas d'images simulées et réelles.
Nous présentons ensuite des conditions suffisantes pour la convergence presque sûre de l'algorithme MCEM, dans un contexte où peu de résultats sont disponibles à cause de la complexité du modèle markovien.
Ces résultats sur l'estimation et la segmentation dans un cadre où le nombre de classes du modèle est connu sont complétés par un travail sur l'étape de sélection d'un modèle markovien. Là encore, l'outil champ moyen est intéressant puisqu'il permet de définir des approximations simples de critères classiques (Bayesian Information Criterion, Integrated Completed Likelihood) et qui s'avèrent prometteuses en pratique.

MOTS-CLES :
Segmentation, Champ de Markov caché, Approximation de type champ moyen, Algorithme EM, Algorithmes EM de type champ moyen, Monte-Carlo EM, Échantillonneur de Gibbs, Sélection de modèle.

DISCIPLINE : Mathématiques Appliquées

Thèse réalisée au sein du projet IS2 de l'Inria Rhône-Alpes.