Modèles markoviens et extensions pour la classification de données complexes

Juliette Blanchet

Université Joseph Fourier

Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique

Modèles Markoviens et extensions pour la classification de données complexes

THÈSE présentée par

Juliette BLANCHET

pour l'obtention du titre de

DOCTEUR EN SCIENCES
Spécialité Mathématiques Appliquées

Directeurs de thèse :	**Cordelia Schmid** Directeur de recherche, INRIA Rhône-Alpes
	Florence Forbes Chargée de recherche, INRIA Rhône-Alpes
Rapporteurs :	**Xavier Descombes** Chargé de recherche, INRIA Sophia-Antipolis
	Wojciech Pieczynski Professeur, INT, Evry
Examinateurs :	**Gersende Fort** Chargée de recherche au CNRS, ENST, Paris
	Bernard Ycart Professeur, Université Joseph Fourier, Grenoble

Modèles markoviens et extensions pour la classification de données complexes

Résumé : Nous abordons le problème de la classification d'individus à partir d'observations dites « complexes » en ce sens qu'elles ne vérifient pas certaines des hypothèses simplificatrices classiquement adoptées. Dans ce travail, les individus à classer sont supposés dépendants les uns des autres. L'approche adoptée est une approche probabiliste fondée sur une modélisation markovienne. Trois problèmes de classification sont abordés.
Le premier concerne la classification de données lorsque celles-ci sont de grande dimension. Pour un tel problème, nous adoptons un modèle markovien gaussien non diagonal tirant partie du fait que la plupart des observations de grande dimension vivent en réalité dans des sous-espaces propres à chacune des classes et dont les dimensions intrinsèques sont faibles. De ce fait, le nombre de paramètres libres du modèle reste raisonnable.
Le deuxième point abordé s'attache à relâcher l'hypothèse simplificatrice de bruit indépendant unimodal, et en particulier gaussien. Nous considérons pour cela le modèle récent de champ de Markov triplet et proposons une nouvelle famille de Markov triplet adaptée au cadre d'une classification supervisée. Nous illustrons la flexibilité et les performances de nos modèles sur une application à la reconnaissance d'images réelles de textures.
Enfin, nous nous intéressons au problème de la classification d'observations dites incomplètes, c'est-à-dire pour lesquelles certaines valeurs sont manquantes. Nous développons pour cela une méthode markovienne ne nécessitant pas le remplacement préalable des observations manquantes. Nous présentons une application de cette méthodologie à un problème réel de classification de gènes.

Mots clés : classification, champ de Markov caché, indépendance conditionnelle, champ de Markov triplet, données de grande dimension, observations manquantes, algorithme EM, approximation de type champ moyen.

Markovian models and extensions for complex data clustering

Abstract : We address the issue of clustering individuals from « complex » observations in the sense that they do not verify some of the classically adopted simplifying assumptions. In this work, the individuals to be clustered are assumed to be dependant upon one another. We adopt a probabilistic approach based on Markovian models. Three clustering problems are considered.
The first of these relates to high-dimensional data clustering. For such a problem, we adopt a non-diagonal Gaussian Markovian model which is based upon the fact that most high-dimensional data actually lives in class dependent subspaces of lower dimension. Such a model only requires the estimation of a reasonable number of parameters.
The second point attempts go beyond the simplifying assumption of unimodal, and in particular Gaussian, independent noise. We consider for this the recent triplet Markov field model and propose a new family of triplet Markov field models adapted to the framework of a supervised classification. We illustrate the flexibility and performances of our models, applied through real texture image recognition.
Finally, we tackle the problem of clustering with incomplete observations, i.e. for which some values are missing. For this we develop a Markovian method which does not require preliminary imputation of the missing data. We present an application of this methodology on a real gene clustering issue.

Key-words : clustering, hidden Markov field, conditional independence, triplet Markov field, high dimensional data, missing data, EM algorithm, mean field-like approximation.

Remerciements

Je tiens tout d'abord à remercier les membres de mon jury, Madame Gersende Fort et Messieurs Xavier Descombes, Wojciech Pieczynski et Bernard Ycart. Je leur suis reconnaissante d'avoir participé à l'évaluation de ma thèse. Leurs nombreux commentaires m'ont été d'un grand apport.

Je tiens également à remercier mes directrices de thèse Cordelia Schmid et Florence Forbes. Merci de m'avoir aidé à mener cette thèse jusqu'au bout. Je tiens en particulier à remercier Florence avec qui j'ai pris beaucoup de plaisir à travailler, à discuter, ... et à courir !

Je tiens encore à remercier tous ceux qui m'ont soutenue, mes parents, mon grand frère, ceux qui ont eu la malchance de partager mon bureau (et de supporter mes affaires de course à pieds dans l'armoire) et de manière générale les membres des équipes Mistis et Lear de l'INRIA. Un merci particulier :

- à Charles pour m'avoir rendu de nombreux services (SOS Linux en particulier).
- à Matthieu avec qui ce fut un plaisir de travailler, de courir ... et de boire occasionnellement du houblon.
- à Caroline pour avoir transformé notre bureau en cafétéria.
- à Vasil pour sa gentillesse, ses chocolats et ses fleurs !
- à Sophie à qui je souhaite bonne chance !
- à Alexandre pour son rire et sa bonne humeur.
- à Stéphane et Laurent pour leur présence et leurs conseils.

J'ai gardé le meilleur pour la fin : merci à Cyril pour tout ce que je ne peux énumérer ici.

Table des matières

Introduction

La classification est le traitement qui consiste à regrouper des individus en groupes homogènes par rapport aux mesures effectuées sur ces individus. Le terme "individu" est ici à prendre au sens large : il s'agit de l'entité sur laquelle on fait des observations (les pixels d'une image, des gènes, des segments de texte etc...). L'objectif de la classification est alors de regrouper les individus qui se "ressemblent" dans une même classe. Les mesures effectuées sur les individus peuvent être de nature variable (réelles, entières, dans l'intervalle $[0, 1]$...), uni- ou multi-dimensionnelles. Cela peut par exemple être des niveaux d'expression de gènes à différents instants. Les individus sont alors les gènes, les mesures des niveaux d'expression multi-dimensionnels, chaque dimension correspondant à un instant ou à une condition expérimentale particulière. Cela peut encore être des intensités en imagerie hyperspectrale, les individus étant les pixels et les mesures des intensités à différentes longueurs d'onde. L'intérêt de la classification est alors de fournir une vue résumée de l'ensemble des données. Cela peut permettre, par exemple, de regrouper les gènes intervenant dans la réplication de l'ADN, ou encore de mettre en évidence les zones de $C0_2$ dans le sol de la planète Mars.

Une approche possible pour une telle problématique est l'approche probabiliste sous laquelle observations et classes sont supposées être des réalisations de variables aléatoires. Le problème de la classification peut être vu comme un problème à données manquantes, les classes à associer aux individus étant non observées, donc manquantes. L'approche probabiliste repose alors sur la donnée d'un modèle pour le couple des observations et des classes, généralement décomposé en un modèle régissant les classes et un modèle régissant la génération des observations lorsque les classes sont connues (ou encore conditionnellement aux classes). On parle plus communément de *modèle de bruit*. Dans la pratique, des hypothèses simplificatrices sont souvent adoptées :

- Au niveau de la modélisation, on suppose en général que les classes sont indépendantes les unes des autres et que le modèle de bruit se factorise sur les individus (on parle alors de *bruit indépendant*). Sous ces deux hypothèses, les individus sont alors implicitement supposés indépendants les uns des autres. Enfin, le bruit est supposé être de forme assez simple, gaussien en général, ou au moins unimodal.

- Au niveau des cas traités, les observations sont en général supposées de dimension raisonnable. Dans le cas contraire, les composantes de chaque observation sont supposées indépendantes les unes des autres. Enfin, aucune observation n'est manquante. Lorsque, pour différentes raisons, certaines observations viennent à manquer, soit cette observation n'est pas traitée (comme si aucune mesure n'avait été faite sur l'individu correspondant), soit les valeurs manquantes sont remplacées de manière

brutale (par des zéros, la moyenne...).

Or, en pratique, il existe beaucoup de cas où ces hypothèses sont mises en défaut et ne donnent pas de résultats satisfaisants. En particulier, les observations effectuées sont souvent dépendantes les unes des autres (les niveaux de gris des pixels d'une image par exemple). De plus les données modernes, du fait des progrès des appareils de mesure et des capacités de stockage, sont souvent en grande dimension. Notons encore que l'hypothèse de bruit indépendant gaussien est mal adapté à certains cas réels, notamment pour la modélisation de textures ou de manière général de classes non unimodales. Enfin, il est très fréquent que certaines observations soient manquantes (certains pixels d'une image, lorsque des réponses à certaines questions d'un sondage n'ont pas été remplies...). De manière générale, nous entendons par l'expression "données complexes" des données sujettes à différentes sources de complexité et donc ne suivant pas le cadre idéal des hypothèses précédemment décrites. L'objet de cette thèse est alors de proposer des modèles et méthodes permettant de classer de telles données "complexes" en nous limitant aux différentes sources de complexité mentionnées et donc de relâcher ces hypothèses trop simplificatrices.

Pour supprimer l'hypothèse d'indépendance des individus, nous nous plaçons dans le cadre d'une modélisation markovienne. De tels modèles permettent, de part leur structure, de prendre en compte explicitement les dépendances entre les différents individus. Celles-ci y sont définies à l'aide d'un système de voisinage, ou, de manière équivalente, d'un graphe (ou réseau) d'interactions. Les champs de Markov ou, de manière équivalente les champs de Gibbs, sont issus, à l'origine, de la physique statistique où ils ont été introduits notamment pour étudier les phénomènes de transition de phase [71]. Leur intérêt pour le traitement d'images n'est apparu que plus tard avec [73] et l'article fondateur de Geman et Geman [58]. L'utilisation de tels modèles ira croissant avec les années et les applications sont nombreuses et variées (analyse d'images, d'enregistrement sonores, de données génomiques, de texte, ...).

Les travaux réalisés dans cette thèse portent sur les trois points suivants, que nous détaillons ensuite :

- le développement de méthodes pour la classification de données de grande dimension sous hypothèse markovienne.
- le développement de méthodes pour la classification supervisée lorsque le modèle de bruit n'est ni indépendant, ni unimodal.
- le développement de méthodes pour la classification d'observations incomplètes (certaines mesures sont manquantes).

Classification de données de grande dimension Les données de grande dimension souffrent de deux problèmes connus sous le nom de *fléau de la dimension* et du *phénomène de l'espace vide*. Le fléau de la dimension est un terme générique introduit par Bellman en 1957 dans [84] pour parler de la difficulté d'optimiser une fonction par une recherche exhaustive de l'optimum dans un espace discrétisé (par exemple une grille régulière de pas 0.1 sur le cube unité dans un espace de dimension 10). Plus généralement, on parle de fléau de la dimension pour exprimer la difficulté de trouver des paramètre optimaux

en grande dimension. Le phénomène de l'espace vide, dont la paternité est usuellement attribuée à Scott et Thompson [115], met en évidence un effet surprenant de la grande dimension allant à l'encontre de la représentation habituelle : les données de grande dimension vivent en réalité dans des espaces de dimension beaucoup plus petite. C'est d'ailleurs la définition du terme "grande dimension" adoptée par certains auteurs, dont Verleysen [125] : le phénomène de l'espace vide définit la frontière entre les espaces de petite et de grande dimension.
Pour faire face au problème du fléau de la dimension, des modèles parcimonieux peuvent être utilisés, comme l'un des 14 modèles particuliers proposés dans [5] pour le cas gaussien, allant du modèle le plus parcimonieux (le modèle sphérique) au modèle le moins parcimonieux (le modèle gaussien classique à matrice de covariance pleine). Néanmoins de tels modèles n'ont pas été spécifiquement élaborés pour les données de grande dimension et, en particulier, ne tiennent pas compte du phénomène de l'espace vide.
Une deuxième solution serait d'utiliser des méthodes de réduction de dimension (ACP, sélection de variables...). Néanmoins, en classification, ces méthodes de réduction de dimension se font au prix d'une perte d'information car, certes, toutes les variables ne sont peut être pas informatives, mais l'ensemble des variables est souvent nécessaire pour discriminer les classes les unes par rapport aux autres.
Pour classer les données de grande dimension, nous proposons de nous baser sur le modèle développé par C. Bouveyron [23] dans le cadre d'une classification par mélange indépendant. Nous étendons sa méthode à une classification sous hypothèse markovienne (chapitre III). Une telle technique permet alors de tenir compte de la dépendance entre les individus sur lesquels sont effectuées les mesures, ainsi que de la dépendance entre les composantes d'une mesure, tout en ne nécessitant l'estimation que d'un nombre raisonnable de paramètres.

Classification supervisée par modèle de Markov triplet Dans de nombreux cas pratiques, et notamment en modélisation de textures et plus généralement de classes non unimodales, l'hypothèse largement utilisée de bruit indépendant est trop restrictive et la relâcher est indispensable. Pour cela, différents modèles markoviens ont été proposés dans la littérature, notamment l'utilisation de champs de Markov gaussiens [39], de champs de Markov cachés à bruit non corrélé ou de champs de Markov cachés généraux [6]. De manière générale, le succès des modèles de Markov caché à bruit indépendant est dû au fait qu'il est possible, sous une telle modélisation, d'utiliser les méthodes bayésiennes classiques pour estimer les paramètres et classer les individus. En effet, la distribution des classes conditionnellement aux observations (encore appelée *loi a posteriori*) reste alors markovienne, si bien que les techniques générales de simulations dites "méthodes de Monte-Carlo par Chaînes de Markov" (ou encore *méthodes MCMC*) [105] sont utilisables. Or, la double hypothèse de champ de Markov caché (sous laquelle les classes suivent un champ de Markov) et de bruit indépendant est suffisante mais non nécessaire pour que la distribution *a posteriori* soit markovienne. A partir de cette observation fondamentale, W. Pieczynski et A. Tebbache ont proposé dans [100] un modèle plus général, les *champs de Markov couples*. De tels modèles autorisent la prise en compte de la non-indépendance du bruit tout en permettant l'utilisation des mêmes traitements bayésiens que sous l'hypothèse de bruit indépendant. Néanmoins, ils ne peuvent modéliser efficacement le caractère multi-modal des classes, en conservant leur flexibilité au niveau de l'estimation

des paramètres. A cette fin, D. Benboudjema et W. Pieczynski ont récemment proposé dans [6] un modèle plus général que les champs de Markov couples, les champs de Markov triplets, permettant, par l'ajout d'un champ auxilliaire, de modéliser un bruit plus riche avec des coûts algorithmiques similaires. Plus précisément, l'intérêt d'un tel modèle est de pouvoir modéliser des cas plus complexes que ce que permet le modèle standard de champ de Markov caché à bruit indépendant sans nécessiter d'autre méthodes que celles classiquement utilisées sous ce modèle simple. En pratique cependant, les modèles de Markov triplets ne sont pas utilisés dans toute leur généralité et des hypothèses particulières sont faites dans les applications (voir [6] et [7]).
Dans cette thèse, nous nous sommes intéressés à des modèles de Markov triplets différents de ceux utilisées dans [6] et [7]. Nous avons initialement construits ces modèles avec pour objectif la classification supervisée d'individus issus de classes complexes ou soumis à des modèles de bruits non standards (non unimodal et non indépendant). Le terme "supervisé" signifie que nous disposons d'individus étiquetés (nous connaissons leurs classes). A partir des observations correspondantes (formant ce qu'on appelle la *base d'apprentissage*), nous désirons classer d'autres individus (la *base de test*) dans ces mêmes classes. Une telle classification supervisée peut être très utile en pratique. Pensons par exemple à la classification supervisée d'images IRM pour détecter automatiquement (donc rapidement et à grande échelle) la présence, ou non, d'une tumeur à partir de quelques images étiquetées par un médecin. Pour une telle problématique, nous proposons de nouveaux modèles basés sur les modèles de Markov triplets et adaptés à un cadre supervisé. Nous détaillons les étapes d'apprentissage et de test par application de l'algorithme de type champ moyen-EM proposé par G.Celeux, F. Forbes et N.Peyrard dans [29]. Il est important de remarquer que toute autre méthode classique d'estimation (la procédure ICE, la gradient stochastique...) aurait pu être utilisée du fait de la définition même des champs de Markov triplets. Enfin, nous illustrons notre modèle sur une application à la reconnaissance de textures (chapitre VI). Par ailleurs, l'étude de ces modèles nous a permis de mettre en évidence, dans le cas non supervisé, un potentiel problème de non-identifiabilité des paramètres des modèles de Markov triplets dès lors que l'on cherche à utiliser des modèles de bruits trop généraux.

Classification d'individus avec observations manquantes Il est très courant en pratique de ne pas disposer de toutes les mesures pour tous les individus. Par exemple, l'appareil de mesure peut être défaillant et les pixels les plus brillants de l'image ne pas être mesurés. Ou bien une expérience biologique peut avoir échoué sur certains gènes (parce qu'ils n'ont pas ou mal réagit). La méthode la plus simple pour faire face à un tel problème est d'éliminer brutalement les individus pour lesquels certaines observations sont manquantes. On comprend aisément qu'une telle technique soit à éviter. Une deuxième technique très populaire est de remplacer les données manquantes par des valeurs (des zéros, la moyenne etc...) en pré-traitement (on parle encore d'*imputation*), puis d'effectuer la classification à partir des observations ainsi complétées. Cette technique, si elle est très couramment utilisée du fait de sa simplicité, tend à introduire un biais dans l'échantillon. De plus, concernant le choix des valeurs imputées, aucune solution universelle n'existe et des choix différents peuvent conduire à des résultats très différents. Des méthodes plus perfectionnées ont été proposées dans le cadre du modèle de mélange indépendant, notamment d'appliquer l'algorithme EM de [44] pour estimer un tel modèle (voir [108] par exemple). Néanmoins, à notre connaissance, aucune méthode n'a été développée dans un cadre de dépendance markovienne des individus. Nous proposons des méthodes pour clas-

ser de telles observations non complètes sous modélisation markovienne. Nous détaillons en particulier le fonctionnement de l'algorithme champ moyen-EM proposé dans [29] et l'adaptons à ce problème sous lequel deux champs sont manquants : les classes et les variables non observées. Nous appliquons une telle démarche à la classification de données d'expression de gènes issues de puces ADN.

Organisation du mémoire

La première partie est consacrée à la description des modèles de champ de Markov cachés pour la classification. Notre contribution principale y est de synthétiser les connaissances sur ces modèles et d'étendre certaines formules classiques sur les modèles d'Ising ou de Potts au modèle plus général de champ de Markov avec intéraction d'ordre 2 (modèle que nous appellerons *modèle de Potts étendu*). La seconde partie traite de la classification de données sous l'hypothèse d'un bruit non standard : de grande dimension, non unimodal et/ou non indépendant. Nous présentons entre autre le modèle de champ de Markov triplet de [6] généralisant celui de champ de Markov couple de [100]. Nous proposons alors une nouvelle famille de modèles de Markov triplets pour la classification supervisée et précisons les étapes d'apprentissage et de test. Nous relevons également les possibles problèmes de non-identifiabilité que posent ces modèles dans le cas général. La troisième partie concerne la classification d'individus avec observations manquantes sous modélisation markovienne. Nous adaptons l'algorithme champ moyen-EM proposé dans [29] à ce problème. Enfin, une annexe contient différentes précisions et notamment une description du logiciel SpaCEM^3 mis au point durant la thèse.

Partie A

Classification de données structurées

Chapitre I

Les champs de Markov

Les modèles markoviens sont des modèles largement utilisés dans de nombreux domaines d'application, notamment en traitement d'images [34] (restauration [30], segmentation [45], analyse de mouvement [1, 22], reconstruction 3D [70], tomographie [110], IRM [113]...), analyse de textes [19, 92], reconnaissance vocale [66], classification de gènes [52], modélisation de musique [98]. Leur caractéristique est de fournir une modélisation explicite des dépendances entres individus via l'utilisation d'une structure de voisinage ou, de manière équivalente, d'un graphe (ou réseau) d'intéractions. Le terme "individu" est ici à prendre au sens large : il s'agit de l'entité sur laquelle on fait des observations, les pixels d'une image, des gènes, des segments de texte, un signal sonore etc... Les individus étant définis sur un graphe d'intéraction, par abus de langage, on les dénomme encore *sites* bien que leurs dépendances ne soient pas nécessairement géographiques.

Ce chapitre a pour but de donner une vue globale des champs de Markov. Après une présentation générale, nous développons certaines méthodes d'approximation d'une distribution markovienne, notamment l'approximation en *champ moyen*. Enfin, nous décrivons des procédures d'estimation des paramètres d'un champ de Markov.

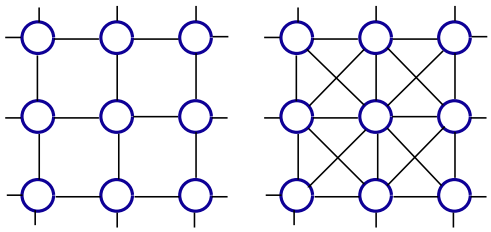
1 Généralités

1.1 Système de voisinage

La définition d'un champ de Markov repose sur celle d'un système de voisinage, c'est-à-dire la donnée, pour chaque individu $i \in \mathcal{I}$ de ses voisins $N_i = \{j \in \mathcal{I} \text{ t.q. } j \text{ est voisin de } i\}$. Ce système de voisinage peut encore être vu comme un graphe $\mathcal{G}$ reliant les individus $i \in \mathcal{I}$ (encore appelés *sommets*), par des branches ou *arêtes* : si j est voisin de i (c'est-à-dire si $j \in N_i$), alors une arête relie i à j. En général, le système de voisinage est supposé symétrique :

$$j \in N_i \Leftrightarrow i \in N_j,$$

ce qui revient à dire que les arêtes du graphe $\mathcal{G}$ sont non-orientées. Notons qu'il est possible de définir un champ de Markov à partir d'un système de voisinage non symétrique ($\mathcal{G}$ est alors orienté) mais cette asymétrie entraîne des complications supplémentaires sur le modèle, même pour des graphes très simples. Mentionnons l'approche proposée dans [83] permettant de considérer des voisinages non stationnaires et éventuellement



(a) (b)

FIG. I.1 – (a) Système des 4-plus proches voisins (voisinage de premier ordre), (b) Système des 8-plus proches voisins (voisinage de second ordre)

non symétriques. Pour cela, les auteurs introduisent un champ auxilliaire supplémentaire de manière à se ramener au cas classique de champ de Markov avec voisinage symétrique.

Sauf mention contraire, nous supposons dans la suite que le système de voisinage est symétrique.

La première question à se poser est alors, à partir d'individus $\mathcal{I}$, comment définir le sytème de voisinage (ou de manière équivalente le graphe) à leur associer. Nous définissons dans cette partie deux grandes familles de graphes, les *graphes de proximité* construits à partir d'une métrique (section 1.1.1) et les graphes aléatoires, construits à partir d'une loi de probabilité (section 1.1.2).

1.1.1 Graphe de proximité

Le terme "graphe de proximité" désigne la famille de graphes construits à partir d'une métrique : deux sites sont reliés l'un à l'autre s'ils sont "suffisamment proches". Cette proximité peut être fonction de la disposition spatiale des sites, ou d'une certaine *similarité* entre ces individus. Il existe de nombreuses manières de construire un graphe de proximité à partir d'une métrique. Nous présentons ci-après les graphes les plus utilisés.

Graphe des k-voisins réciproques Pour chaque individu, on cherche ses k-plus proches voisins (au sens d'une distance à définir) et on ne retient que les voisins réciproques, de manière à avoir un graphe symétrique (non orienté) (Figure I.2 (b)).
C'est en particulier le type de graphe utilisé dans le cas d'une grille régulière, pour les pixels d'une image par exemple. Plus précisément, on utilise alors en général (voir Figure I.1) :

- les 4-plus proches voisins, aussi appelé *voisinage de premier ordre* : chaque individu est relié à ses deux voisins horizontaux et à ses deux voisins verticaux.
- les 8-plus proches voisins, aussi appelé *voisinage de second ordre* : chaque individu est relié à ses quatre voisins horizontaux et verticaux et à ses quatre voisins diagonaux.

Graphe des ε-voisins Chaque point est connecté aux points situés à une distance inférieure à ε (voir un exemple en Figure I.2 (c) pour la distance euclidienne). Notons que le système des 4-plus proches voisins sur une grille régulière est également un ε-graphe pour la distance euclidienne, où ε est égal à l'écartement entre deux sites consécutifs sur une même ligne (ou colonne). De même, le système des 8-plus proches voisins sur un

grille régulière est également un ε-graphe pour la distance euclidienne, où ε est égal à l'écartement entre deux sites consécutifs sur une diagonale.

Triangulation de Delaunay Une telle tesselation est obtenue comme dual du diagramme de Voronoï. Une cellule de Voronoï associée à un site i de $\mathcal{I}$ est composée de l'ensemble des points qui sont plus proches de i que de tout autre point de $\mathcal{I}$ (Figure I.2 (d)). Deux points i et j créent alors une arête dans le graphe de Delaunay si et seulement si les régions de Voronoï associées à i et j sont adjacentes (Figure I.2 (e)).

Graphe de voisinage relatif (ou *Relative Neighborhood Graph*) La construction d'un tel graphe est fondée sur la notion de voisins "relativement proches" définie par Lankford [80]. Deux sites i et j de $\mathcal{I}$ sont dits "relativement proches" s'il n'existe pas de point plus proche à la fois de i et de j, c'est-à-dire si :

$$d(i,j) \leq \min_{s \in \mathcal{I} \setminus \{i,j\}} \max\{d(i,s), d(j,s)\}$$

où d est une distance (la distance Euclidienne en général). Le graphe de voisinage relatif [122] est alors obtenu en reliant les points "relativement proches"(Figure I.2 (f)).

Graphe de Gabriel Deux sites i et j sont voisins dans le graphe de Gabriel [56] s'il n'existe pas d'autre point dans la boule passant par i et j et de rayon $\frac{d(i,j)}{2}$ (Figure I.2 (g)).

1.1.2 Graphe aléatoire

Les graphes aléatoires (*random graphs*) furent initialement introduits en 1959 par Erdös and Rényi [47]. Le modèle de graphe aléatoire le plus étudié est celui noté $\mathcal{G}_{np}$ composé de n sommets, chaque paire de sommets étant reliés par une arête avec la probabilité p (voir Figure I.3). Lorsque n est grand, la probabilité p_r qu'un sommet quelconque soit de degré r (c'est-à-dire soit relié à r sommets) est alors :

$$p_r = C_n^r \, p^r \, (1-p)^{n-r} \approx \frac{\mu^r e^{-\mu}}{r!}$$

en notant $\mu = pn$ le degré moyen d'un sommet. La distribution des degrés peut donc être approximée par une loi de Poisson, d'où le nom de *graphe aléatoire de Poisson*. La structure d'un tel graphe aléatoire dépend alors grandement de la valeur de p (voir Figure I.3). Les graphes aléatoires ont été généralisés à des degrés non-Poisson, entre autre aux lois puissance, ainsi qu'aux graphes aléatoires exponentiels [121].
L'intérêt des graphes aléatoires est de pouvoir modéliser l'effet "petit monde" (*small-world*) [42] régissant les grands réseaux d'interaction et stipulant qu'il existe des chemins très courts (c'est-à-dire passant par peu de nœuds intermédiaires) entre tous les nœuds du réseau. Les observations expérimentales ont en effet montré que les grands réseaux d'interactions (sociales, informatiques, biologiques) présentaient l'effet petit monde (voir Figure I.4). Pour plus de détails sur les graphes aléatoires en général, on pourra se reporter à [94].

1.1.3 Clique

Un système de voisinage donné définit un ensemble $\mathcal{C}$ de parties de $\mathcal{I}$ appelées *cliques*. Une clique est définie comme étant soit un singleton, soit un ensemble de sites deux à

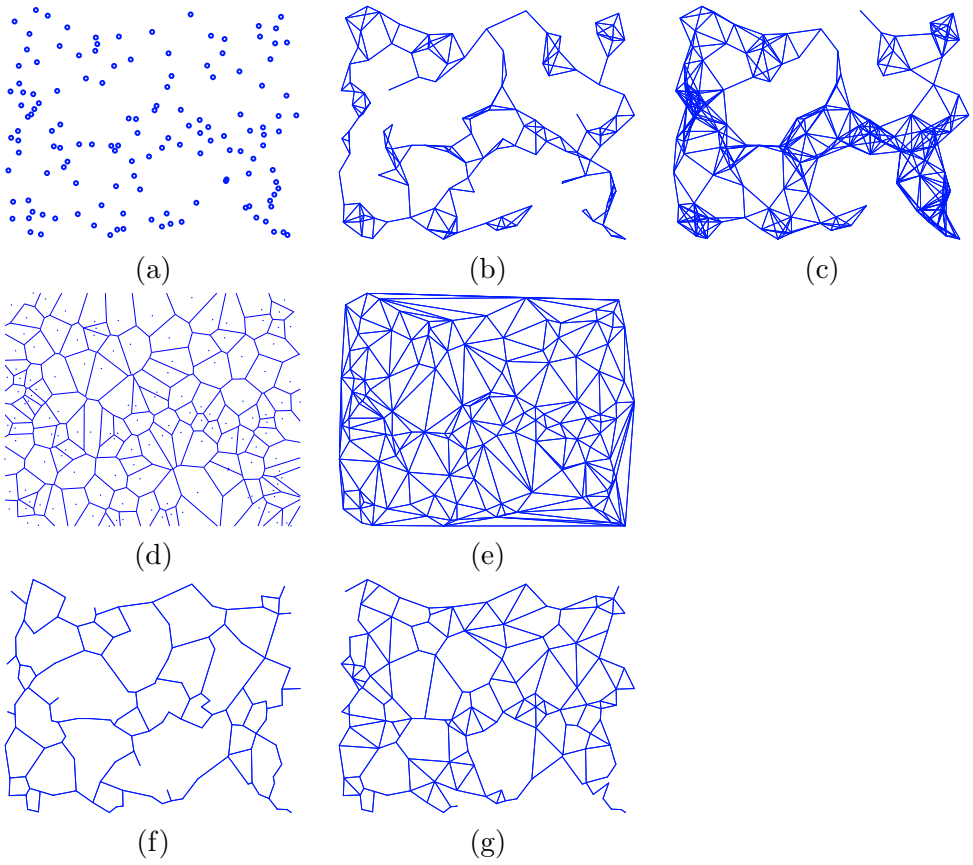


FIG. I.2 – (a) Point irrégulièrement espacés, (b) graphe des k-voisins réciproques ($k = 5$), (c) ε-graphe, (d) cellules de Voronoï, (e) triangulation de Delaunay, (f) graphe de voisinage relatif, (g) graphe de Gabriel

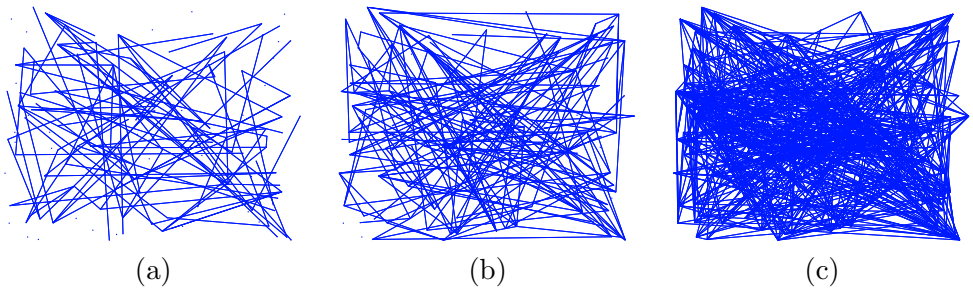
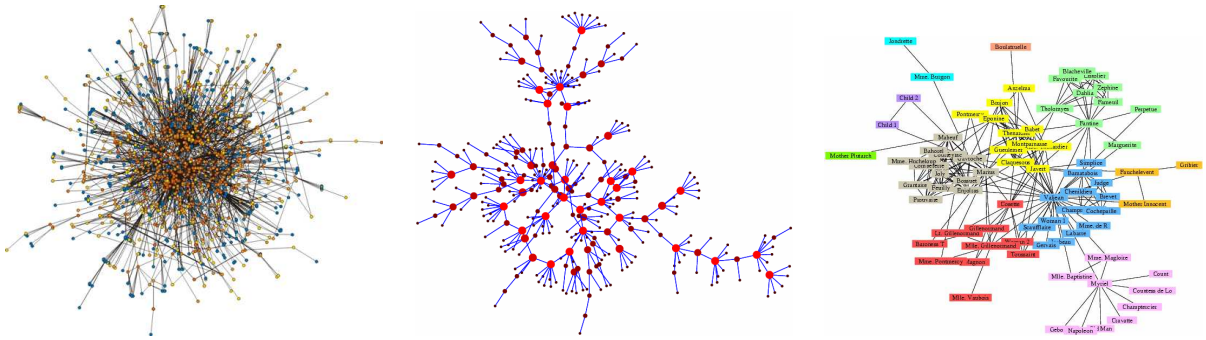
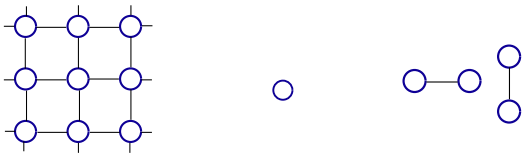


FIG. I.3 – (a) (b) (c) Simulation d'un graphe aléatoire de Poisson ($p = 0.01$, 0.02 et 0.06),



(a) (b) (c)

FIG. I.4 – Effet petit monde : (a) Graphe d'interaction de 1600 gènes orthologues [116] (deux gènes sont reliés s'il existe une association fonctionnelle entre leurs protéines), (b) Réseau de contact sexuels entre individus [94], (c) Réseau d'interaction entre les principaux personnages des *Misérables* de V. Hugo [95]



(a) (b) (c)

FIG. I.5 – (a) Système des 4 plus proches voisins, (b) Clique d'ordre 1, (c) Cliques d'ordre 2

deux voisins. On appelle *ordre* d'une clique le nombre de ses éléments. Les Figures I.5 et I.6 donnent les cliques induites respectivement par un système de voisinage des 4 et 8 plus proches voisins.

1.2 Définition d'un champ de Markov

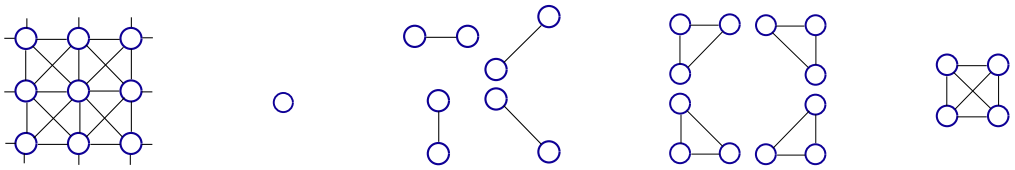
Soit $\mathcal{I}$ un ensemble de sites indicés par $\{1, \cdots, n\}$ sur lequel un système de voisinage $\mathcal{G}$ est défini. Notons $i \sim j$ la relation de voisinage entre deux sites i et j de $\mathcal{I}$ et N_i l'ensemble des sites voisins du site i. Considérons un système de variables aléatoires $\mathbf{Z} = \{Z_1, \cdots, Z_n\}$ définies sur $\mathcal{I}$. De manière générale, notons $\mathbf{Z}_A$ l'ensemble des variables $\{Z_i, i \in A\}$ pour un ensemble quelconque $A \subset \mathcal{I}$, et $\mathcal{I}\backslash\{i\}$ l'ensemble des sites $\mathcal{I}$ privé du site i.

Définition 1 (Champ de Markov)**.** *Soit* $\mathbf{Z} = \{Z_1, \cdots, Z_n\}$ *un système de variables aléatoires localisées en n sites, et* $\mathcal{G}$ *un graphe de voisinage défini sur ces sites. On dit que* $\mathbf{Z}$ *est un champ de Markov associé à* $\mathcal{G}$ *si les deux conditions suivantes sont satisfaites :*

$$\forall \mathbf{z}, \forall i, \quad P(Z_i = z_i | \mathbf{Z}_{\mathcal{I}\backslash\{i\}} = \mathbf{z}_{\mathcal{I}\backslash\{i\}}) = P(Z_i = z_i | \mathbf{Z}_{N_i} = \mathbf{z}_{N_i}) \tag{I.1}$$

$$\forall \mathbf{z}, \quad P(\mathbf{Z} = \mathbf{z}) > 0. \tag{I.2}$$

La condition (I.1) signifie que la distribution de la variable aléatoire Z_i au site i ne dépend des autres variables $\mathbf{Z}_{\mathcal{I}\backslash\{i\}}$ qu'à travers la valeur des variables aléatoires $\mathbf{Z}_{N_i}$ aux



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

FIG. I.6 – (a) Système des 8 plus proches voisins, (b) Clique d'ordre 1 , (c) Cliques d'ordre 2 , (c) Cliques d'ordre 3 , (d) Cliques d'ordre 4

sites voisins de i, et traduit donc la nature locale de la dépendance. Il s'agit d'une extension de la notion de chaîne de Markov. Dès lors que la condition de positivité (I.2) est vérifiée, la distribution jointe $P(\mathbf{z})$ est définie de manière unique à partir des caractéristiques locales. Cette condition de positivité (I.2) signifie que toute configuration doit avoir une probabilité non nulle d'occurrence. Elle est suffisante pour assurer la consistance de $P(\mathbf{z})$, mais non nécessaire et peut donc être relâchée [81].
Bien qu'intuitive, cette définition est en pratique peu exploitable ([58], p. 725) et on préfère caractériser un champ de Markov par une distribution jointe, ce qui est rendu possible grâce au théorème d'Hammersley-Clifford [9] :

Définition 2 (Distribution de Gibbs)**.** *Soit $\mathcal{G}$ un graphe de voisinage et $\mathcal{C}$ un ensemble de cliques associées à $\mathcal{G}$. Un champ aléatoire $\mathbf{Z}$ est régi par une distribution de Gibbs relativement à $\mathcal{G}$ si sa loi jointe est de la forme :*

$$P_G(\mathbf{z}) = W^{-1} \exp(-H(\mathbf{z})) \tag{I.3}$$

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

$$H(\mathbf{z}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{z}_c) \tag{I.4}$$

et W est la constante de normalisation, encore appelée fonction de partition,

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

Théorème 3 (Hammersley-Clifford)**.** *Le champ aléatoire $\mathbf{Z}$ est un champ de Markov si et seulement si sa loi jointe $P_G(\mathbf{z})$ est une distribution de Gibbs.*

Les termes "énergie", "potentiel" sont issus de la physique statistique. Les distributions de Gibbs ont en effet initialement été utilisées pour modéliser le comportement de systèmes moléculaires en interaction, de particules sous l'influence d'un champ magnétique

par exemple.

Notons qu'un champ de Markov peut être discret ou continu. Dans la suite, sauf mention contraire, nous nous plaçons dans le cas où le champ $\mathbf{Z}$ est discret. Pour obtenir les formules dans le cas continu, il suffit souvent de remplacer les sommes discrètes Σ par des intégrales $\int$.

Il est important de remarquer que, si la formulation gibbsienne (I.3) d'un champ de Markov fournit une expression analytique de la loi jointe $P_G(\mathbf{z})$, celle-ci est incalculable en pratique. En effet, son calcul requiert celui de la fonction de partition W. Or, si les variables aléatoires Z_i peuvent prendre K valeurs possibles et en notant n le nombre de sites, W est une somme de K^n termes, qui explose lorsque n est grand. A titre d'exemple, pour $n = 1000$ sites et $K = 2$, W est la somme de plus de 10^{301} termes ! De même, les distributions marginales $P(z_i)$ ne peuvent être calculées. Néanmoins, le calcul des probabilités conditionnelles ne posent pas de problème. En effet, dans le cas où $\mathbf{Z}$ est discret, on a :

$$\begin{aligned} P_G(z_i|\mathbf{z}_{\mathcal{I}\setminus\{i\}}) &= \frac{P_G(z_i,\mathbf{z}_{\mathcal{I}\setminus\{i\}})}{\sum_{z'_i} P_G(z'_i,\mathbf{z}_{\mathcal{I}\setminus\{i\}})} = \frac{W^{-1}}{W^{-1}} \frac{\exp(-\sum_{c\ni i} V_c(\mathbf{z}_c))\,\exp(-\sum_{c\not\ni i} V_c(\mathbf{z}_c))}{\sum_{z'_i}\exp(-\sum_{c\ni i} V_c(\underline{\mathbf{z}_c}))\,\exp(-\sum_{c\not\ni i} V_c(\mathbf{z}_c)}) \\ &= \frac{\exp(-\sum_{c\ni i} V_c(\mathbf{z}_c))}{\sum_{z'_i}\exp(-\sum_{c\ni i} V_c(\underline{\mathbf{z}_c}))} \end{aligned} \tag{I.5}$$

en notant $\underline{\mathbf{z}_c} = \{z'_i\} \cup \{z_j, j \in c, j \neq i\}$. La somme au numérateur de (I.5) ne porte que sur K termes et se calcule donc facilement.

1.3 Exemples

Nous donnons dans cette section des exemples de champs de Markov les plus utilisés.

1.3.1 Modèle d'Ising

Le modèle markovien le plus simple est le modèle d'Ising [71], issu de la mécanique statistique. Il correspond au cas où les variables Z_i ne peuvent prendre que deux valeurs, $+1$ ou -1. L'énergie du champ $\mathbf{Z}$ est donnée par :

$$H(\mathbf{z}) = h\sum_i z_i - J\sum_{i\sim j} z_i z_j \tag{I.6}$$

En mécanique statistique, cette énergie modélise l'interaction ferromagnétique entre *spins* voisins, les spins pouvant être orientés vers le haut ($z_i = +1$) ou vers le bas ($z_i = -1$). Le paramètre J témoigne du caractère ferromagnétique ($J > 0$) ou anti-ferromagnétique ($J < 0$) du modèle. Ce modèle est également utilisé en sciences économiques pour modéliser des systèmes de coopération ou non entre individus [75].

1.3.2 Modèle de Potts et extensions

Dans certaines applications, la variable Z_i représente une étiquette attribuée au site i, parmi K possibles. On parle alors de champ de Markov discret, et les K valeurs possibles

des Z_i sont appelées les *classes* (ou *labels*). Décomposons la fonction d'énergie H (I.4) en somme sur les cliques de différentes tailles :

$$H(\mathbf{z}) = \sum_i V_i(z_i) + \sum_{\substack{i,j \\ \text{voisins}}} V_{ij}(z_i, z_j) + \cdots + \sum_{\substack{i_1,\cdots,i_q \\ \text{voisins}}} V_{i_1 \cdots i_q}(z_{i_1}, \cdots, z_{i_q}) \tag{I.7}$$

Dans la plupart des cas, les potentiels sur les cliques d'ordre 1 et 2 sont considérés comme suffisants pour modéliser les dépendances spatiales. De plus, prendre en compte les cliques d'ordre > 2 complique grandement les analyses. Aussi fixe-t-on le plus souvent à zéro les potentiels sur les cliques d'ordre supérieur à 2, et l'énergie (I.7) devient :

$$H(\mathbf{z}) = \sum_i V_i(z_i) + \sum_{\substack{i,j \\ \text{voisins}}} V_{ij}(z_i, z_j) \tag{I.8}$$

Potentiels sur les singletons Les potentiels sur les singletons (les cliques d'ordre 1) $V_i(z_i)$ permettent de modéliser la probabilité d'occurence de la classe z_i au site i considéré individuellement. En mécanique statistique, ces potentiels représentent l'influence du champ magnétique externe. Lorsque les potentiels $V_i(z_i)$ dépendent de i (et non seulement de z_i), on parle de champ externe non-stationnaire. Notons que, sous l'hypothèse de non-stationnarité et sans paramétrisation particulière, l'estimation des potentiels sur les singletons est impossible puisqu'il faudrait estimer K potentiels V_i par site i. Néanmoins, de tels potentiels non-stationnaires peuvent être intéressants pour intégrer de l'information *a priori* visant à influencer les sites individuellement (voir par exemple [113]). En l'absence de connaissance particulière, l'hypothèse classique consiste à supposer que ces fonctions potentiels sont les mêmes sur l'ensemble des sites, c'est-à-dire que $V_i(z_i)$ ne dépend du site i qu'à travers la valeur de z_i. Cette hypothèse correspond à un champ magnétique externe spatialement stationnaire et peut se traduire par la notation :

$$V_i(z_i) = -\alpha_{z_i}. \tag{I.9}$$

Les fonctions potentiels sur les singletons sont alors caractérisées par le vecteur des poids $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$ associés aux K classes. En adoptant la notation vectorielle $\mathbf{z}_i = \mathbf{e}_{z_i}$ où $(\mathbf{e}_1, \cdots, \mathbf{e}_K)$ désigne la base canonique, et en notant $\mathbf{z}_i'$ la transposée du vecteur $\mathbf{z}_i$, une écriture équivalente à (I.9) est :

$$V_i(z_i) = -\mathbf{z}_i' \boldsymbol{\alpha}. \tag{I.10}$$

Si les potentiels sur les cliques de taille supérieure à 1 sont nuls, la distribution (I.3) s'écrit :

$$P_G(\mathbf{z}) = \frac{\exp(\sum_i \alpha_{z_i})}{\sum_{z_1'} \cdots \sum_{z_n'} \exp(\sum_i \alpha_{z_i'})} = \prod_i \frac{\exp(\alpha_{z_i})}{\sum_{z_i'} \exp(\alpha_{z_i'})} \tag{I.11}$$

La loi jointe se décompose alors en produit de fonctions de z_i qui peuvent s'interpréter comme la probabilité d'occurence de la classe z_i au point i. Aussi est-il commun de dire que les potentiels sur les singletons pondèrent l'importance relative des classes. En un site quelconque, la probabilité de trouver la classe k vaut alors :

$$\pi_k \;=\; \frac{e^{\alpha_k}}{\sum_{k'=1}^{K} e^{\alpha_{k'}}}$$

Notons que les fonctions potentiels sont équivalentes à une constante additive près car l'ajout d'une constante ne modifie pas la distribution de Gibbs correspondante. Pour définir les potentiels sur les singletons de manière unique, on peut par exemple imposer $\sum_k \alpha_k = 0$.

Potentiels sur les paires Les potentiels sur les paires $V_{ij}(z_i, z_j)$ permettent de modéliser la dépendance entre les classes Z_i et Z_j en des sites i et j voisins. En mécanique statistique, ces potentiels représentent la force d'interaction entre particules voisines. On suppose en général que ces fonctions potentiels sont les mêmes sur l'ensemble des sites, ce qui peut se traduire par la notation :

$$V_{ij}(z_i, z_j) = V(z_i, z_j) = -\beta_{z_i, z_j}. \tag{I.12}$$

Les fonctions potentiels sur les paires sont alors caractérisées par la matrice symétrique $\boldsymbol{\beta} = (\beta_{kk'})_{k,k' \in [\![1,K]\!]}$ associée aux $K \times K$ interactions entre classes. En adoptant la notation vectorielle $\mathbf{z}_i = \mathbf{e}_{z_i}$ où $(\mathbf{e}_1, \cdots, \mathbf{e}_K)$ désigne la base canonique, une écriture équivalente à (I.12) est :

$$V_{ij}(z_i, z_j) = -\mathbf{z}_i' \boldsymbol{\beta} \mathbf{z}_j. \tag{I.13}$$

Le terme $\beta_{kk'}$ peut s'interpréter comme le degré de compatibilité entre les classes k et k'. Les fonctions potentiels étant équivalentes à une constante additive près, pour définir les potentiels sur les paires de manière unique, on peut par exemple imposer $\beta_{11} = 0$.

Une hypothèse largement utilisée consiste à supposer de plus que la matrice $\boldsymbol{\beta}$ s'écrit $\boldsymbol{\beta} = \beta \mathbb{I}_K$ où $\mathbb{I}_K$ désigne la matrice unité de dimension $K \times K$. Son énergie est alors donnée par :

$$H(\mathbf{z}) = -\sum_{i \sim j} \mathbf{z}_i' \boldsymbol{\beta} \mathbf{z}_j = -\beta \sum_{i \sim j} 1_{z_i = z_j} = -\beta N(\mathbf{z})$$

où $N(\mathbf{z})$ désigne le nombre de paires homogènes (dans la même classe) pour la classification $\mathbf{z}$. Une telle distribution correspond à la distribution de Strauss avec classes interchangeables [120], plus communément appelé *modèle de Potts*. Elle est largement utilisée en segmentation markovienne d'image ([10] par exemple) car elle traduit de la façon la plus simple possible l'hypothèse de régularité spatiale : plus le paramètre $\beta > 0$ est grand, plus la probabilité que deux sites voisins i et j soient dans la même classe (c'est-à-dire $z_i = z_j$) est élevée. Notons que lorsque $\beta = 0$, les classes $Z_1, \cdots, Z_n$ sont indépendantes les unes des autres et toutes les classes sont équiprobables.

Exemple. En Figure I.7, se trouvent des simulations d'un modèle de Potts à 2 et 3 classes (ou couleurs), pour différentes valeurs du paramètre spatial β. Pour $\beta = 0$, les Z_i sont indépendants les uns des autres : $\forall i \in \mathcal{I},\ P(z_i) = \frac{1}{K}$. Plus β augmente, plus les classes ont tendance à se regrouper. On peut observer sur la Figure I.7 un phénomène de transition de phase : à partir d'une certaine valeur critique β_c ($\beta_c \approx 0.38$ pour $K = 2$, $\beta_c \approx 0.44$ pour $K = 3$), la réalisation générée est presque uniforme (voir section 1.4 pour plus de détails).

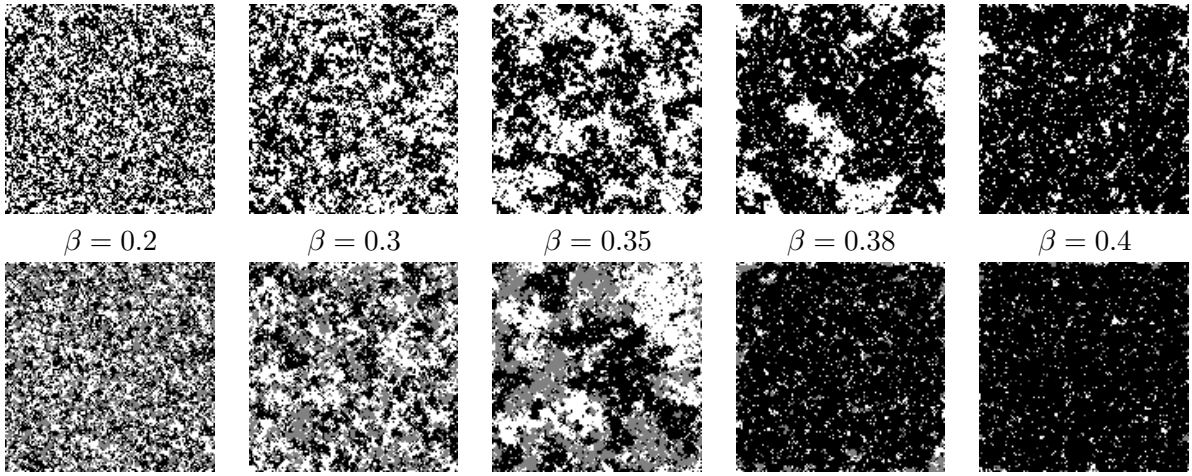


FIG. I.7 – Simulation d'un modèle de Potts à 2 couleurs (1ère ligne) et 3 couleurs (2ème ligne), pour différentes valeurs du paramètre spatial β.

1.3.3 Auto-modèles

Les auto-modèles [9] sont des modèles markoviens répandus en analyse d'image, notamment en analyse de textures. L'énergie y est définie sur les cliques d'ordre 1 et 2 uniquement. Il existe différentes variations sur les auto-modèles selon la forme donnée aux fonctions potentiels V_i et V_{ij}. Les plus utilisés sont le modèle auto-binomial [39] et le modèle auto-normal [32, 36].

Modèle auto-binomial : Soit $\mathbf{Z}$ un champ discret à valeur dans $[\![1, K]\!]$. On définit la distribution conditionnelle de Z_i connaissant la valeur du champ pour ses voisins par la loi binomiale $\mathcal{B}(K, q)$:

$$\forall \mathbf{z}, \qquad P_G(z_i|\mathbf{Z}_{N_i} = \mathbf{z}_{N_i}) = C_K^{z_i} q_{N_i}^{z_i} (1 - q_{N_i})^{K - z_i} \tag{I.14}$$

$$\text{avec} \qquad q_{N_i} = \frac{\exp(a_i + \sum_{j \in N_i} b_{ij} z_j)}{1 + \exp(a_i + \sum_{j \in N_i} b_{ij} z_j)} \tag{I.15}$$

où les a_i sont des réels quelconques et les b_{ij} sont tels que $\forall i, j,\ b_{ij} = b_{ji}$. $\mathbf{Z}$ est alors un champ de Markov d'énergie :

$$H(\mathbf{z}) = -\sum_i (\ln C_K^{z_i} + a_i z_i) - \sum_{i \sim j} b_{ij} z_i z_j \tag{I.16}$$

Ce modèle a été notamment utilisé pour la modélisation de textures [39].

Modèle auto-normal Chacune des variables Z_i est à valeur dans $\mathbb{R}$. La distribution conditionnelle de Z_i connaissant la valeur du champ pour ses voisins est donnée par :

$$\forall \mathbf{z}, \quad P_G(z_i|\mathbf{Z}_{N_i} = \mathbf{z}_{N_i}) \propto \exp(-\frac{1}{2\lambda_i}(z_i - \sum_{j \in N_i} \beta_{ij} z_j)^2) \tag{I.17}$$

Les paramètres doivent de plus vérifier : $\beta_{ij}\lambda_j = \beta_{ji}\lambda_i$ pour tous sites i et j voisins. Notons $\mathbf{Q}$ la matrice de dimension $n \times n$ telle que :

$$Q_{ij} = \begin{cases} \frac{1}{\lambda_i} & \text{si } i = j \\ -\frac{\beta_{ij}}{\lambda_i} & \text{si } i \text{ et } j \text{ voisins} \\ 0 & \text{sinon} \end{cases}$$

$\mathbf{Q}$ est une matrice symétrique, elle est de plus définie positive sous certaines conditions sur les paramètres λ_i et β_{ij} [9]. La loi jointe du champ $\mathbf{Z}$ s'écrit alors :

$$P(\mathbf{z}) \propto \exp(-\frac{1}{2}\mathbf{z}'Q\mathbf{z}) \tag{I.18}$$

Ce modèle fait partie de la famille des champs de Markov gaussiens, notamment utilisés en modélisation de textures [33]. Il s'agit d'un auto-modèle avec :

$$V_i(z_i) = -\frac{1}{2\lambda_i}z_i^2 \qquad V_{ij}(z_i, z_j) = \frac{\beta_{ij}}{\lambda_i}z_i z_j$$

1.3.4 Champ de Markov causal

Les modèles décrits jusqu'à présent font partie de la famille de champ de Markov à voisinage symétrique, ou encore *non causaux*, puisque $j \in N_i \Leftrightarrow i \in N_j$. La difficulté des modèles non causaux réside dans la difficulté du calcul de la loi jointe $P(\mathbf{z})$, et ce même pour un voisinage très simple. Les champs de Markov *causaux* (*Markov mesh models*) [2] sont construits de sorte à contourner cette difficulté. Le graphe de voisinage est supposé acyclique et orienté, la relation de voisinage n'est donc plus symétrique. La distribution du champ au pixel i ne dépend alors que d'un petit nombre des prédécesseurs $\mathcal{P}_i$ (ou *parents*) de i. Il s'en suit que la loi jointe se décompose en produit de probabilités conditionnelles facilement calculables, comme dans le cas d'une chaîne de Markov :

$$P_G(\mathbf{z}) = \prod_{i \in \mathcal{I}} P(z_i | \mathbf{z}_{\mathcal{P}_i}) \tag{I.19}$$

Les champs de Markov causaux ont notamment été utilisés en reconnaissance de caractères [62].
Un cas particulier est celui des arbres de Markov causaux dans lesquels chaque site i a un unique parent. De tels modèles ont été utilisés en segmentation d'images (voir [21, 78] par exemple). L'avantage premier des arbres de Markov est que l'estimation des paramètres peut être effectuée de manière exacte et qu'il existe des algorithmes rapides d'estimation. Néanmoins, du fait de la structure en arbre, ces modèles conduisent à des segmentations non lisses (la classification présente des "blocs", voir [49] pour illustration en segmentation d'images).

1.4 Transition de phase

Une caractéristique importante des champs de Markov est qu'ils connaissent une *transition de phase*. En physique, ce phénomène a lieu lorsqu'une petite variation des paramètres physiques (la température par exemple) produit des changements brusques et à

longue portée du système.
Considérons une distribution de Gibbs paramétrée par $\boldsymbol{\phi}$, de potentiels sur les cliques V_c :

$$P_G(\mathbf{z}|\boldsymbol{\phi}) = W(\boldsymbol{\phi})^{-1} \exp(-H(\mathbf{z};\boldsymbol{\phi})) = W(\boldsymbol{\phi})^{-1} \exp(-\sum_{c\in\mathcal{C}} V_c(\mathbf{z}_c;\boldsymbol{\phi}))$$

Lorsque les potentiels V_c sont analytiques, la fonction $\boldsymbol{\phi} \mapsto \exp(-H(\mathbf{z};\boldsymbol{\phi}))$ est analytique comme l'exponentielle d'une somme de fonctions analytiques. De même, la fonction de partition $W(\boldsymbol{\phi})$ est analytique, ainsi que son inverse $W(\boldsymbol{\phi})^{-1}$. Elle cesse néanmoins d'être analytique lorsque le nombre de sites $n \to \infty$. C'est ce caractère non-analytique de la fonction de partition lorsque $n \to \infty$ qui est responsable de la transition de phase [63, 60]. Notons ∇_ϕ le gradient par rapport à $\boldsymbol{\phi}$ et

$$f_n : \boldsymbol{\phi} \mapsto \frac{\nabla_\phi \log W(\boldsymbol{\phi})}{n}$$

Le caractère non analytique de la fonction de partition lorsque $n \to \infty$, se traduit par l'existence d'une (ou des) valeur(s) critique(s) $\boldsymbol{\phi}_c$ des paramètres tel(s) que :

$$\boldsymbol{\phi} \mapsto \lim_{n\to\infty} \nabla_\phi f_n(\boldsymbol{\phi}) \text{ est discontinue en } \boldsymbol{\phi}_c.$$

Si $\boldsymbol{\phi}$ est de dimension ν (il y a ν paramètres au modèle), alors $\nabla_\phi f_n$ est un vecteur de dimension ν et pour que $\boldsymbol{\phi}_c$ soit une valeur critique, il faut qu'au moins une des composantes $r \in [1,\nu]$ de $\boldsymbol{\phi} \mapsto \lim_{n\to\infty} \nabla_\phi f_n(\boldsymbol{\phi})$ soit discontinue :

$$\exists r \in [1,\nu] \text{ tel que } \boldsymbol{\phi} \mapsto \lim_{n\to\infty} [\nabla_\phi f_n(\boldsymbol{\phi})]_r \text{ est discontinue en } \boldsymbol{\phi}_c$$

où $[.]_r$ désigne la composante numéro r d'un vecteur. Notons $\nabla^2_\phi f_n$ la hessienne par rapport à $\boldsymbol{\phi}$. $\nabla_\phi f_n$ étant de dimension ν, cette matrice hessienne est une matrice symétrique de dimension $\nu \times \nu$. Cette discontinuité en $\boldsymbol{\phi}_c$ lorsque $n \to \infty$ s'écrit encore :

$$\exists s,t \in [1,\nu] \text{ tel que } \lim_{\phi\to\phi_c} \lim_{n\to\infty} [\nabla^2_\phi f_n(\boldsymbol{\phi})]_{st} = \infty$$

où $[.]_{st}$ désigne la composante numéro s,t d'une matrice. En utilisant les formules du gradient et de la hessienne de $\log W$ (voir Annexe 3, Proposition 18), les paramètres $\boldsymbol{\phi}_c$ de transition de phase sont ceux pour lesquels :

$$\exists r \in [1,\nu] \text{ tel que } \boldsymbol{\phi} \mapsto \lim_{n\to\infty} \left[\frac{-\langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle}{n}\right]_r \text{ est discontinue en } \boldsymbol{\phi}_c$$

ou encore tels que :

$$\exists s,t \in [1,\nu] \text{ tel que } \lim_{\phi\to\phi_c} \lim_{n\to\infty} \left[\frac{-\langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle + Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}))}{n}\right]_{st} = \infty$$

où $\langle . \rangle$ désigne l'espérance. Considérons en particulier le cas d'un modèle de Potts étendu (voir section 1.3.2), de paramètres $\boldsymbol{\phi} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ et d'énergie

$$H(\mathbf{z};\boldsymbol{\phi}) = -\sum_i \mathbf{z}'_i \boldsymbol{\alpha} - \sum_{i\sim j} \mathbf{z}'_i \boldsymbol{\beta} \mathbf{z}_j.$$

Remarquons que la hessienne $\nabla^2_\phi H(\mathbf{z};\boldsymbol{\phi})$ est alors nulle (il n'y a pas de terme en $\boldsymbol{\alpha}^2$ ou $\boldsymbol{\beta}^2$ ou $\boldsymbol{\alpha}\boldsymbol{\beta}$), si bien que la transition de phase correspond aux $\boldsymbol{\phi}_c$ tels que :

$$\exists s,t \in [1,\nu] \text{ tel que } \lim_{\phi\to\phi_c}\lim_{n\to\infty}\left[\frac{Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}))}{n}\right]_{st} = \infty$$

Lorsque $\mathbf{Z}$ est à valeur dans $[\![1,K]\!]$, $\boldsymbol{\alpha}$ est un vecteur de dimension K et $\boldsymbol{\beta}$ une matrice symétrique de dimension $K\times K$, si bien que $\boldsymbol{\phi}$ est de dimension $\nu = K + \frac{1}{2}K(K+1)$.

Exemple. Dans le cas d'un modèle de Potts de matrice $\boldsymbol{\beta} = \beta\mathbb{I}_K$, on a $\nu = 1$ et les équations :

$$\begin{aligned} H(\mathbf{z};\beta) &= \beta N(\mathbf{z}) \\ \nabla_\beta \log W(\beta) &= -\langle N(\mathbf{Z})\rangle \\ \nabla^2_\beta \log W(\beta) &= Var(N(\mathbf{Z})) \end{aligned}$$

où $N(\mathbf{Z})$ désigne le nombre de paires homogènes (voisins dans la même classe) de $\mathbf{Z}$ lorsque $\mathbf{Z}$ suit la loi $P_G(.|\beta)$. Le paramètre β_c de transition de phase est donc tel que, au voisinage de β_c et lorsque $n\to\infty$, la proportion de paires homogènes (par rapport au nombre total n d'individus) est discontinu. Physiquement, en dessous de la valeur critique β_c, les réalisations générées sont presque aléatoires. Aux environs de β_c, des groupes se forment. Enfin, au dessus de β_c, elles sont presque uniformes (d'une seule couleur). On pourra se reporter à la Figure I.7 pour illustration. La valeur de ce paramètre critique β_c est une fonction croissante du nombre de classes : dans le cas d'un voisinage d'ordre 1, la valeur du paramètre β_c critique est [60] :

$$\beta_c = \frac{1}{2}\log(1+\sqrt{K})$$

On donne en Figure I.8 les courbes correspondant à la moyenne et à la variance empirique du nombre de paires homogènes $N(\mathbf{z})$ en fonction du paramètre β pour un voisinage d'ordre 2 (l'espérance et la variance de $N(\mathbf{Z})$ ne peuvent être calculées de manière exacte). Les différentes simulations sont effectuées sur une image de taille 128×128. On observe que $\beta_c \approx 0.38$ pour $K=2$, $\beta_c \approx 0.44$ pour $K=3$. Notons que, plus la taille de l'image augmente (c'est-à-dire plus n augmente), plus la pente de $\langle N(\mathbf{Z})\rangle$ autour de β_c augmente, et plus la courbe de la variance devient pointue. A la limite ($n\to\infty$), on obtient pour la courbe $\langle N(\mathbf{Z})\rangle$ une fonction en escalier autour de β_c et un dirac pour $Var(N(\mathbf{Z}))$.

De manière générale, notons Φ_c l'ensemble de tous les paramètres critiques. Goutsias [63] remarque que l'espace Φ dans lequel vivent les paramètres peut être décomposé en 3 ensembles disjoints Φ_-, Φ_c et Φ_+ appélés respectivement région *sous-critique*, *critique* et *sur-critique*. Lorsque les paramètres $\boldsymbol{\phi}\in\Phi_-$, le modèle de Markov est soumis à des corrélations de courte portée et les réalisations comportent des régions homogènes de faible taille. Un tel modèle peut être intéressant pour modéliser des détails fins (en texture notamment). Lorsque $\boldsymbol{\phi}\in\Phi_+$, le modèle de Markov est soumis à des corrélations de longue portée et les réalisations comportent des régions homogènes de taille importante. Pour pouvoir modéliser des images aux détails plus ou moins fins, il peut être intéressant de simuler des réalisations d'un champ de Markov en jouant sur les paramètres $\boldsymbol{\phi}$ de celui-ci.

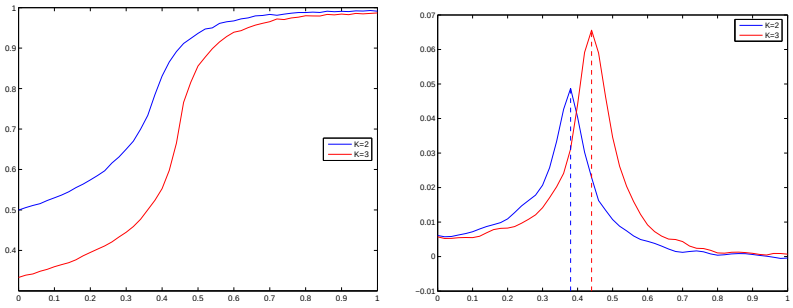


FIG. I.8 – Transition de phase pour un modèle de Potts à $K = 2$ et $K = 3$ classes, défini sur un voisinage d'ordre 2 : à gauche, proportion de paires homogènes (deux sites voisins de même classe) en fonction du paramètre β du modèle de Potts, à droite, variance de cette proportion en fonction de β.

1.5 Simuler un champ de Markov

La loi jointe d'un champ de Markov $\mathbf{Z}$ n'étant pas calculable directement, il est fondamental de pouvoir simuler des réalisations de $\mathbf{Z}$, par exemple pour approximer son espérance par des méthodes de type Monte-Carlo [105] ou pour estimer ses paramètres. Les techniques classiques de simulation (comme la méthode d'inversion ou de rejet) ne sont pas utilisables car elles nécessiteraient le calcul de la fonction de partition. Il existe d'autres méthodes permettant de simuler une distribution de Gibbs. La plus répandue est l'échantillonneur de Gibbs (*Gibbs sampler*) [58] qui est méthode de type Monte Carlo par Chaîne de Markov (MCMC). Il s'agit d'un cas particulier de l'algorithme de Metropolis-Hasting [91] à une composante. L'algorithme de l'échantillonneur de Gibbs est itératif : partant d'une réalisation initiale $\mathbf{z}^{(0)}$, il génère à l'itération (q) une réalisation $\mathbf{z}^{(q)}$. Plus précisément, lors de l'itération (q), un seul site i_q est visité, si bien que deux réalisations successives $\mathbf{z}^{(q-1)}$ et $\mathbf{z}^{(q)}$ ne peuvent différer que par la composante z_{i_q}. L'ensemble des réalisations $\{\mathbf{z}^{(q)}\}_{q\in\mathbb{N}}$ ainsi créées forme une chaîne de Markov ; la probabilité de transition entre les "états" $\mathbf{z}^{(q-1)}$ et $\mathbf{z}^{(q)}$ est donnée par :

$$P(\mathbf{Z}^{(q)} = \mathbf{z}^{(q)} | \mathbf{Z}^{(q-1)} = \mathbf{z}^{(q-1)}) = \begin{cases} 0 & \text{si } \exists i \neq i_q / z_i^{(q)} \neq z_i^{(q-1)} \\ P_G(z_{i_q}^{(q)} | \mathbf{z}_{N_{i_q}}^{(q-1)}) & \text{sinon} \end{cases} \tag{I.20}$$

Si chaque site est mis à jour de façon séquentielle (peu importe l'ordre choisi pour cela), alors la chaîne $\{\mathbf{z}^{(q)}\}_{q\in\mathbb{N}}$ est irréductible, apériodique et admet P_G comme unique mesure stationnaire, quelle que soit l'initialisation [58]. Cependant l'algorithme peut mettre longtemps à converger car à chaque itération, un seul site est mis à jour. La convergence est d'autant plus lente que les paramètres $\boldsymbol{\phi}$ du champ $\mathbf{Z}$ considéré sont critiques ou sous-critiques ($\boldsymbol{\phi} \in \Phi_- \cup \Phi_c$ avec les notation de la section 1.4). En effet, les corrélations étant de courte portée, beaucoup d'itérations sont nécessaires pour apporter l'information des sites lointains [63]. Néanmoins cet algorithme est couramment utilisé, car il ne nécessite que le calcul des probabilités conditionnelles (I.5), et non pas celui de la fonction de partition W. Notons enfin que la version synchrone de l'algorithme (mise à jour en parallèle des sites) n'est pas valide.

Une utilisation classique des algorithmes de simulation est le calcul de quantités liées au champ de Markov (son espérance $\langle \mathbf{Z} \rangle$ typiquement) par moyenne empirique. Une alternative déterministe consiste à remplacer la distribution de Gibbs par une distribution plus simple pour laquelle les calculs sont faisables. Nous précisons cette approche dans la section suivante.

2 Approximer une distribution de Gibbs

Une distribution de Gibbs n'étant pas calculable de manière exacte, des approximations sont nécessaires. Nous présentons brièvement en section 2.1 l'approche variationnelle conduisant à l'approximation en champ moyen. Pour plus de détails, on pourra se reporter en Annexe 4. Nous détaillons également en section 2.2 quelques autres approximations classiques.

2.1 Approche variationnelle : approximation en champ moyen

2.1.1 Principe

Soit P_G une distribution de Gibbs, que l'on souhaite approximer par une distribution Q ayant la propriété de factorisation :

$$Q(\mathbf{z}) = \prod_i Q_i(z_i) \tag{I.21}$$

L'approche variationnelle consiste à déterminer la distribution Q optimale au sens de la divergence de Kullback-Leibler :

Définition 4 (Divergence de Kullback-Leibler)**.** *La divergence de Kullback-Leibler entre deux distributions P et Q est définie par :*

$$KL(Q||P) = \sum_{\mathbf{z}} Q(\mathbf{z}) \log \frac{Q(\mathbf{z})}{P(\mathbf{z})} = \left\langle \log \frac{Q(\mathbf{Z})}{P(\mathbf{Z})} \right\rangle_Q$$

où on a noté $\langle . \rangle_Q$ l'espérance par rapport à la loi Q.

Notons que cette divergence n'est pas une distance mathématique : elle est non symétrique et elle ne satisfait pas l'inégalité triangulaire. Elle est néanmoins toujours non négative et s'annule si et seulement si les deux distributions P et Q sont égales, Q presque partout.

Si P_G est définie par des fonctions potentiels V_c sur les cliques $c \in \mathcal{C}$:

$$P_G(\mathbf{z}) = \frac{1}{W} \exp(-\sum_{c \in \mathcal{C}} V_c(z_c))$$

alors la distribution Q_i est donnée par (voir Annexe 2)

$$Q_i(z_i) \propto \exp \left\langle -\sum_{c \ni i} V_c(z_i, \mathbf{Z}_{c \setminus i}) \right\rangle_{Q_{\neq i}} = \prod_{c \ni i} \exp \left\langle -V_c(z_i, \mathbf{Z}_{c \setminus i}) \right\rangle_{Q_{c \setminus i}} \tag{I.22}$$

où $Q_{\neq i}$ et $Q_{c \setminus i}$ dénotent respectivement les lois définies par $Q_{\neq i}(\mathbf{z}_{\neq i}) = \prod_{j \neq i} Q_j(z_j)$ et $Q_{c \setminus i}(\mathbf{z}_{c \setminus i}) = \prod_{j \in c \setminus i} Q_j(z_j)$.

Supposons que les cliques soient d'ordre 1 et 2 uniquement et que les Z_i sont discrets, $Z_i \in [\![1, K]\!]$ par exemple. Définissons, sur les cliques c d'ordre 2, les matrices $\mathbb{V}_c$ de dimension $K \times K$ telles que $V_c(z_i, z_j) = \mathbf{e}'_{z_i} \mathbb{V}_c \, \mathbf{e}_{z_j}$ où $\mathbf{e}_k$ désigne le k-ième vecteur de la base canonique en dimension K. Alors,

$$\langle \mathbf{e}'_{z_i} \mathbb{V} \mathbf{e}_{Z_j} \rangle = \mathbf{e}'_{z_i} \mathbb{V} \langle \mathbf{e}_{Z_j} \rangle = \mathbf{e}'_{z_i} \mathbb{V} \mu_j$$

où μ_j désigne le vecteur des K probabilités $Q_j(Z_j = k)$. L'expression (I.22) revient à fixer les voisins j du site i à leur moyenne $\mu_j = \langle \mathbf{Z}_j \rangle_{Q_j}$. On retrouve alors l'approximation en champ moyen [31] de la physique statistique et l'expression (I.21) s'écrit :

$$Q(\mathbf{z}) = \prod_i \frac{\exp\left(-\sum_{c \ni i} V_c(z_i, \boldsymbol{\mu}_{c \setminus i})\right)}{\sum_{z'_i} \exp\left(-\sum_{c \ni i} V_c(z_i, \boldsymbol{\mu}_{c \setminus i})\right)} = \prod_i P(z_i | \boldsymbol{\mu}_{N_i}) \tag{I.23}$$

Cette expression (I.23) est valable de manière générale lorsque V_c est bilinéaire et lorsque sa définition peut être étendue à l'espace d'état des moyennes (voir Annexe 2 pour plus de détails).

Remarque. *Dans le cas général, minimiser la divergence de Kullback-Leibler n'est pas toujours équivalent à remplacer le champ des voisins par leur moyenne.*

C'est le cas dans beaucoup de cas courants, notamment lorsque les Z_i sont discrets et les cliques d'ordre ≥ 3 nulles. En particulier, les modèles classiques de la physique statistique (modèles d'Ising, de Potts) sont discrets et définis sur les singletons et paires uniquement. Le terme "champ moyen" [31] fait référence au remplacement des voisins du site i considéré par leurs moyennes.

Remarquons que les expressions (I.22) et (I.23) ne fournissent pas de solution explicite puisque les expressions de droite dépendent d'espérances calculées sous la loi Q. Une approche possible est alors de résoudre le problème itérativement. Dans le cas particulier où l'équation (I.23) est valable (le champ moyen "usuel"), cette approche peut être remplacée par une équation de cohérence sur les moyennes (voir Annexe 2).

Remarque. *Soit un modèle de Potts classique d'énergie paramétrée par le scalaire β (c'est-à-dire tel que $\boldsymbol{\beta} = \beta \mathbb{I}_K$ avec $\mathbb{I}_K$ la matrice identité de dimension K). Alors l'approximation en champ moyen est peu intéressante puisqu'il s'agit de la configuration uniforme en chaque site, indépendamment du paramètre spatial β. Nous verrons ultérieurement (chapitre II, section 4.2) l'utilité d'une telle approximation pour la distribution Markovienne a posteriori d'un champ de Markov caché.*

2.1.2 Généralisation du principe du champ moyen

La méthode d'approximation en champ moyen repose sur le fait qu'en fixant les voisins du site i (les moyennes μ_j pour $j \in N_i$), on supprime les interactions avec ce site i. On peut imaginer de généraliser ce principe en définissant une famille de distributions sur $\mathcal{I}$ obtenues en fixant les voisins à des constantes quelconques [29]. Ainsi, pour un champ $\tilde{\mathbf{z}}$ arbitraire, nous considérons l'approximation suivante :

$$P_G(\mathbf{z}) \approx P_{\tilde{\mathbf{z}}}(\mathbf{z}) = \prod_{i \in \mathcal{I}} P_G(z_i | \tilde{\mathbf{z}}_{N_i}) \tag{I.24}$$

Toute distribution de la forme (I.24) 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 (I.23) choisit $\tilde{\mathbf{z}} = \boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$. Nous détaillons deux autres choix possibles pour le champ voisin $\tilde{\mathbf{z}}$, conduisant à l'*approximation en champ modal* et à l'*approximation en champ simulé*.

Approximation en champ modal Plutôt que de fixer les voisins $j \in N_i$ du site cible i à leur moyenne, ils sont fixés à leur (ou un de leurs) mode(s) :

$$\forall i \in \mathcal{I},\ \tilde{z}_i = \arg\max_{z_i} P_G(z_i|\tilde{\mathbf{z}}_{N_i})$$

Comme pour l'approximation en champ moyen, le champ des voisins $\tilde{\mathbf{z}}$ est alors solution d'un problème de point fixe qui peut être résolu itérativement.

Approximation en champ simulé Une autre méthode consiste à simuler une réalisation $\tilde{\mathbf{z}}$ de la distribution de Gibbs $P_G(\mathbf{z})$ (par l'échantillonneur de Gibbs [58] par exemple, voir section 1.5).

Dans la suite du mémoire, on parlera d'*approximation de type champ moyen* pour se référer à une approximation d'une distribution de Gibbs de la forme de l'équation (I.24), c'est-à-dire pour laquelle la valeur du champ au site i ne dépend plus de celles des autres sites, ceux-ci étant fixés à des constantes $\tilde{\mathbf{z}}$.

2.2 Autres approximations

2.2.1 Cluster approximation

Il est possible d'imaginer une généralisation du principe du champ moyen en ne fixant pas tous les sites autres que le site i considéré à leurs moyennes, mais seulement ceux "suffisamment" loins, les autres étant traités de manière exacte : c'est la *cluster approximation* [131]. Plus exactement, soit i un site et $I_i \subset \mathcal{I}$ un ensemble de sites contenant i. Les interactions locales entre les sites de I_i sont traitées exactement alors que les interactions de "longue distance" entre les sites de I_i et ceux de $\mathcal{I} \setminus I_i$ sont approximées (les variables Z_j, $j \in \mathcal{I} \setminus I_i$ sont fixées à leurs moyennes μ_j). L'approximation dépend alors également d'une condition de cohérence qui se traduit par un problème de point fixe. L'approximation en champ moyen correspond au cas particulier $I_i = \{i\}$. Si $I_i = N_i$, toutes les interactions sont prises en compte exactement. Pour que cette approximation soit intéressante, il faut donc que I_i soit strictement inclus dans N_i. Néanmoins, [131] montre que les résultats ne sont pas significativement meilleurs en comparaison à l'approximation en champ moyen lorsque I_i est défini par les 4-plus proches voisins de i.

2.2.2 Approximation par un ensemble de variables indépendantes

Le principe, développé dans [57], consiste, chaque fois qu'une interaction est concernée, à négliger les fluctuations. Soit $\mathbf{Z}$ un champ de Markov de fonctions potentiels V_c sur les cliques $c \in \mathcal{C}$. L'approximation remplace V_c par :

$$\hat{V}_c(\mathbf{z}_c) = \frac{1}{|c|} \sum_{i \in c} V_c(z_i, \boldsymbol{\mu}_{c \setminus i})$$

où $\forall i,\ \mu_i$ est une approximation de $\langle Z_i \rangle_{P_G}$ et $|c|$ désigne l'ordre de la clique c. La distribution P_G est alors approximée par $\hat{P}(\mathbf{z}) \propto \prod_i \exp(-\hat{H}_i(z_i))$ ayant pour énergie locale au site i :

$$\hat{H}_i(z_i) = \frac{1}{|c|} \sum_{c \ni i} V_c(z_i, \boldsymbol{\mu}_{c \setminus i})$$

Cette méthode conduit, comme le principe du champ moyen, à approximer la distribution de Gibbs par un produit de distributions. Néanmoins, les deux méthodes diffèrent dans l'approximation des potentiels d'ordre au moins 2, le principe du champ moyen remplaçant l'énergie locale au site i par $\sum_{c \ni i} V_c(z_i, \boldsymbol{\mu}_{c \setminus i})$ (voir équation I.23) qui correspond à $|c| \times \hat{H}_i(z_i)$. L'approche variationnelle (voir Annexe 2.2) permet d'établir qu'en théorie, l'approximation en champ moyen est meilleure au sens de la divergence de Kullback-Leibler. Zhang [136] a comparé les deux méthodes et reporte que les paramètres sont mieux estimés par application du principe du champ moyen (dans le cadre d'une modélisation par champ de Markov caché avec l'algorithme EM).

2.2.3 Pseudo-vraisemblance

Les probabilité conditionnelles (I.5) étant faciles à calculer, une alternative est de remplacer la distribution de Gibbs P_G par le produit de ces probabilités conditionnelles. On obtient alors une fonction appelée *pseudo-vraisemblance* (*pseudo-likelihood*) [10] :

$$\mathcal{PL}(\mathbf{z}) = \prod_i P(z_i | \mathbf{z}_{N_i}) \tag{I.25}$$

Notons que cette fonction $\mathcal{PL}$, bien que facile à calculer, ne définit pas une loi de probabilité.

Remarque. *L'approximation de la pseudo-vraisemblance n'est pas une approximation de type champ moyen comme définie en section 2.1.2. En effet, dans l'équation (I.25), le champ* $\mathbf{z}$ *n'est pas constant, contrairement au champ* $\tilde{\mathbf{z}}$ *de l'équation (I.24).*

3 Estimation des paramètres

Jusqu'à présent, nous avons supposé les paramètres du champ de Markov connus. En pratique, ces paramètres sont en général inconnus et les estimer est une étape indispensable à l'utilisation du modèle. Cette section présente certaines méthodes les plus classiques d'estimation des paramètres $\boldsymbol{\phi}$ d'un champ de Markov.

Définition 5. *Soit* $\mathbf{z}$ *une réalisation d'un champ de Markov* $\mathbf{Z}$ *de distribution de Gibbs* $P_G(\mathbf{z}|\boldsymbol{\phi})$ *paramétrée par* ϕ*. On appelle la vraisemblance du paramètre* ϕ*, la fonction* $L(\boldsymbol{\phi}, \mathbf{z})$ *définie par :*

$$L(\boldsymbol{\phi}, \mathbf{z}) = P_G(\mathbf{z}|\boldsymbol{\phi})$$

Lorsque le champ $\mathbf{z}$ est observé, un estimateur possible du paramètre $\boldsymbol{\phi}$ est l'estimateur de maximum de vraisemblance $\boldsymbol{\phi}^{MV}$:

$$\boldsymbol{\phi}^{MV} = \arg\max_{\phi} P_G(\mathbf{z}|\boldsymbol{\phi})$$

Notons que l'estimateur du maximum de vraisemblance est également l'estimateur de *log-vraisemblance* $l(\boldsymbol{\phi}, \mathbf{z})$ défini par :

$$l(\boldsymbol{\phi}, \mathbf{z}) = \log L(\boldsymbol{\phi}, \mathbf{z}) = \log P_G(\mathbf{z}|\boldsymbol{\phi})$$

Pour une réalisation $\mathbf{z}$ donnée et des interactions d'ordre 1 et 2, l'estimateur de maximum de vraisemblance de P_G est unique (voir Annexe 3, Corollaire 20) et une simple descente de gradient converge vers cet unique estimateur. Néanmoins, cette résolution numérique n'est pas possible directement, sauf cas particuliers (par exemple les modèles *mesh* [2]). En effet, le gradient selon $\boldsymbol{\phi}$ de $P_G(\mathbf{z}|\boldsymbol{\phi})$ n'est pas calculable de manière analytique, car il nécessite l'évaluation de la constante de partition W de la distribution markovienne. Pour y remédier, deux alternatives sont possibles :

1. Faire des approximations de la distribution P_G au moyen de probabilités conditionnelles locales (pseudo-vraisemblance, approximation en champ moyen ou encore méthode de codage [9])
2. Employer des algorithmes itératifs (gradient stochastique par exemple) à partir de la vraisemblance exacte, mais dont il s'agit alors de prouver la convergence ainsi que le type d'optimum trouvé (local, global).

Par pseudo-vraisemblance. L'idée est de remplacer P_G par sa pseudo-vraisemblance $\mathcal{PL}$ (équation I.25). L'estimateur de maximum de pseudo-vraisemblance de $\boldsymbol{\phi}$ est donné par :

$$\boldsymbol{\phi}^{PL} = \arg\max_{\phi} \mathcal{PL}(\mathbf{z}|\boldsymbol{\phi})$$

La fonction de pseudo-vraisemblance, comme son logarithme, est encore concave. Contrairement à la vraisemblance $L(\boldsymbol{\phi}, \mathbf{z})$, le gradient selon $\boldsymbol{\phi}$ de $\mathcal{PL}(\mathbf{z}|\boldsymbol{\phi})$ est calculable puisque la fonction de partition de la distribution de Gibbs n'intervient plus. L'existence, l'unicité et la consistance de l'estimateur de maximum de pseudo-vraisemblance (lorsque le nombre d'individus augmente) ont été démontrées dans [59] pour une famille de champs de Markov à nombre d'états finis, incluant notamment le modèle de Potts.

Par approximation en champ moyen En utilisant l'approximation en champ moyen $P_{\tilde{\mathbf{z}}}$ (I.24) de P_G, on remplace l'estimateur $\boldsymbol{\phi}^{MV}$ par :

$$\tilde{\boldsymbol{\phi}}^{MV} = \arg\max_{\phi} P_{\tilde{\mathbf{z}}}(\mathbf{z}|\boldsymbol{\phi})$$

dont le calcul est facile. En effet, l'évaluation de la distribution $P_{\tilde{\mathbf{z}}}$, du fait de la factorisation sur les sites, ne pose pas de problème.

Par gradient stochastique. On peut également maximiser la vraisemblance $L(\boldsymbol{\phi}, \mathbf{z})$ elle-même de façon approchée au moyen de la technique de gradient stochastique de [134]. D'après la Proposition 18 de l'Annexe 3 :

$$\nabla_\phi L(\boldsymbol{\phi}, \mathbf{z}) = \nabla_\phi \log P_G(\mathbf{z}; \boldsymbol{\phi}) = -\nabla_\phi H(\mathbf{z}; \boldsymbol{\phi}) + \langle \nabla_\phi H(\mathbf{Z}; \boldsymbol{\phi}) \rangle$$

où $\mathbf{Z}$ désigne le champ markovien et $\langle . \rangle$ son espérance sous la loi $P_G(.|\boldsymbol{\phi})$. [134] propose une procédure itérative de type gradient, partant d'une valeur initiale $\boldsymbol{\phi}^{(0)}$ des paramètres.

A l'itération $(q+1)$, l'estimateur $\boldsymbol{\phi}^{(q+1)}$ est mis à jour en se déplaçant dans une direction de gradient "approximative", l'espérance $\langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}) \rangle$ étant incalculable. Pour cela, on simule une réalisation $\hat{\mathbf{z}}^{(q)}$ de la distribution $P_G(\mathbf{Z}|\boldsymbol{\phi}^{(q)})$ paramétrée par l'estimateur courant $\boldsymbol{\phi}^{(q)}$ (via l'échantillonneur de Gibbs par exemple). Les paramètres sont alors mis à jour par :

$$\boldsymbol{\phi}^{(q+1)} = \boldsymbol{\phi}^{(q)} + \tau^{(q+1)}(\nabla_\phi H(\hat{\mathbf{z}}^{(q)};\boldsymbol{\phi}) - \nabla_\phi H(\mathbf{z};\boldsymbol{\phi})) \tag{I.26}$$

où $\tau^{(q+1)}$ est le pas à l'itération $(q+1)$. Cet algorithme stochastique converge presque sûrement, quelle que soit l'initialisation, vers la valeur optimale $\boldsymbol{\phi}^{MV}$.

Exemple 1 : Cas du modèle de Potts. Dans le cas du modèle de Potts, le vecteur des paramètres $\boldsymbol{\phi}$ se réduit à l'unique scalaire β :

$$H(\mathbf{z};\beta) = -\beta \sum_{i\sim j} 1_{z_i=z_j}$$

et le gradient de l'énergie H est donné par :

$$\frac{\partial H(\mathbf{z};\beta)}{\partial \beta} = -\sum_{i\sim j} 1_{z_i=z_j} = -N(\mathbf{z})$$

où $N(\mathbf{z})$ désigne le nombre de paires de voisins homogènes, c'est-à-dire dans la même classe. La règle d'ajustement (I.26) s'écrit :

$$\beta^{(q+1)} = \beta^{(q)} + \tau^{(q+1)}(N(\mathbf{z}) - N(\hat{\mathbf{z}}^{(q+1)}))$$

Cette équation de mise à jour signifie que l'on augmente le coefficient de régularité spatiale β si le nombre de paires homogènes obtenues par simulation est inférieur à celui de la classification $\mathbf{z}$ observée, et qu'on le diminue sinon.

Exemple 2 : Cas du modèle de Potts étendu. Dans le cas du modèle de Potts étendu à K classes, le vecteur des paramètres $\boldsymbol{\phi}$ est la matrice symétrique $\boldsymbol{\beta} = (\beta_{kk'})_{k,k'\in[\![1,K]\!]}$:

$$H(\mathbf{z};\boldsymbol{\beta}) = -\sum_{i\sim j} \beta_{kk'} 1_{z_i=k} 1_{z_j=k'}$$

et le gradient de l'énergie H est donné par les K^2 dérivées :

$$\frac{\partial H(\mathbf{z};\boldsymbol{\beta})}{\partial \beta_{kk'}} = \frac{\partial H(\mathbf{z};\boldsymbol{\beta})}{\partial \beta_{k'k}} = -\sum_{i\sim j} 1_{z_i=k} 1_{z_j=k'} = -N_{kk'}(\mathbf{z})$$

où $N_{kk'}(\mathbf{z})$ désigne le nombre de paires de voisins dans les classes k et k'. La règle (I.26) revient à mettre à jour $\beta_{kk'} = \beta_{k'k}$ par :

$$\beta_{kk'}^{(q+1)} = \beta_{kk'}^{(q)} + \tau^{(q+1)}(N_{kk'}(\mathbf{z}) - N_{kk'}(\hat{\mathbf{z}}^{(q+1)}))$$

Cette équation de mise à jour signifie que l'on augmente le coefficient de compatibilité spatiale $\beta_{kk'}$ entre les classes k et k' si le nombre de paires dans les classes k et k' obtenu

en simulation est inférieur à celui de la classification $\mathbf{z}$ observée, et qu'on le diminue sinon.

Dans le chapitre suivant nous abordons le problème de la classification d'individus et décrivons le modèle de champ de Markov caché pour cette problématique. Nous développons en particulier les méthodes d'estimation précédemment décrites ainsi que la procédure ICE [99] sous une telle modélisation.

Chapitre II

Classification de variables dépendantes par champ de Markov caché

1 La classification

Classer des objets, des individus ou des concepts, c'est regrouper ceux qui se "ressemblent" dans une même classe. Une classification permet d'avoir une vue résumée d'un ensemble de données. Les groupes ainsi mis en évidence peuvent en outre servir de base à une interprétation sur les causes qui ont généré ces groupes. On distingue deux grandes familles de problèmes de classification :

- Les groupes existent déjà, on doit y affecter un nouvel individu. Il s'agit alors d'un problème de *classification supervisée*. La règle de classement se définit en général à partir d'individus étiquetés, c'est-à-dire dont la classe (le groupe) d'origine est connue.
- On veut construire des classes à partir d'individus non étiquetés. Il s'agit alors d'un processus de *classification non supervisée*, visant à détecter parmi les observations des groupements "naturels".

La classification vise donc à regrouper les observations -mono ou multidimensionnelles- en sous-ensembles. Dans beaucoup de cas pratiques, les individus sur lesquels sont effectuées ces observations sont considérés comme indépendants. Les relations pouvant exister entre eux (par exemple une proximité géographique, un lien familial, une similarité) n'ont alors pas d'influence sur leur regroupement. Dans ce travail, nous souhaitons prendre en compte ces informations de dépendance lors du processus de classification.

Dans la suite de ce chapitre, nous disposons de n observations réelles D-dimensionnelles correspondant à un ensemble d'individus $i \in \mathcal{I} = \{1, \cdots, n\}$ que l'on souhaite classer en K groupes. On note $\mathbf{x} = (x_1, \cdots, x_n)$ l'ensemble des observations, avec $\forall i \in \mathcal{I},\ x_i \in \mathbb{R}^D$. Lorsqu'il y a dépendance entre les individus, on parlera encore de sites ou noeuds d'un graphe. Dans ce travail, nous nous intéressons à prendre en compte ces dépendances. Pour cela, un cadre possible est le cadre probabiliste. Néanmoins, du fait de leur utilisation pratique, nous rappelons dans un premier temps quelques approches non probabilistes en classification. Dans ces approches, les individus sont supposés indépendants.

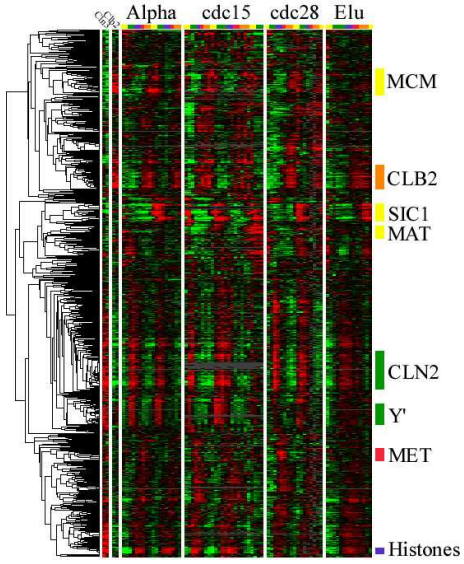


FIG. II.1 – Données d'expression de gènes de levure (*Saccharomyces cerevisiae*) [117] à différents instants. Les gènes correspondent aux lignes, le temps aux colonnes. La classification (à gauche de l'image) est une classification hiérarchique ascendante basée sur un critère d'agrégation de lien moyen [106]. A droite de la figure, sont visibles quelques groupements de gènes interprétés par les biologistes.

1.1 Classification hiérarchique

Une hiérarchie est une suite de partitions emboîtées, depuis l'ensemble de n objets $\{1, \cdots, n\}$, jusqu'aux singletons $\{1\} \cup \cdots \cup \{n\}$, en passant par des subdivisions successives de ces sous-ensembles. La procédure la plus couramment utilisée est la *classification hiérarchique ascendante* [77]. Celle-ci se base sur un tableau de similarités (ou dissimilarités) entre les objets, de taille $n \times n$, par exemple la distance euclidienne entre les objets. Lors de l'initialisation, chaque objet constitue une classe (il y en a donc n). Puis la procédure est itérative, elle détermine successivement les deux classes les plus similaires, les fusionne, puis calcule les similarités entre la nouvelle classe et les autres classes. La procédure s'arrête lorsque tous les objets sont réunis dans une même classe de taille n. Le calcul de similarité entre deux classes peut être obtenu par le critère d'*agrégation de lien minimum* (*single-link clustering*), d'*agrégation de lien moyen* (*average-link clustering*), d'*agrégation de lien maximum* [106], le *critère d'inertie de Ward* [128] (*Ward's minimum variance*) etc... Un des intérêts de la classification hiérarchique est qu'elle permet d'obtenir une classification à différents niveaux de détail : plus on descend dans la hiérarchie, plus la classification est fine. La classification hiérarchique est par exemple utilisée pour l'analyse données d'expression de gènes (voir Figure II.1 issue de [117])

1.2 Agrégation autour des centres-mobiles

Il s'agit d'une méthode produisant une partition des données, chaque donnée étant affectée à une seule des K classes possibles (le nombre K de classes est supposé connu). On parle également de *classification dure*, par opposition à la *classification floue* définie plus

loin. La partition déterminée doit être optimale du point de vue d'un critère d'homogénéité des classes formées. Dans le cas de l'agrégation autour des centres-mobiles, ce critère est la somme des inerties des classes C_k par rapport à une observation centrale $\mu_k \in \mathbb{R}^D$, pour $k = 1, \cdots, K$:

$$J(C, \mu) = \sum_{k=1}^{K} I(C_k, \mu_k) = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2 \tag{II.1}$$

Minimiser ce critère alternativement sur les centres μ et la partition C donne l'algorithme des *centres-mobiles* (*k-means*) [88]. A chaque itération, on affecte chacune des données à la classe dont le centre est le plus proche, puis on recalcule les centres de gravité μ des classes.
Cet algorithme garantit la décroissance du critère au cours des itérations :

$$\forall q,\ J(C^{(q+1)}, \mu^{(q+1)}) \leq J(C^{(q)}, \mu^{(q)}) \tag{II.2}$$

et l'on peut montrer qu'il converge vers une partition stable en un nombre fini d'itérations. L'algorithme est couramment utilisé en raison de sa simplicité de mise en œuvre et de sa rapidité de convergence. Il est néanmoins mal adaptée aux données fortement bruitées ainsi qu'aux observations atypiques. De plus, la partition obtenue dépend de la position initiale choisie. Pour y remédier, la tactique la plus simple et la plus couramment utilisée consiste à lancer l'algorithme avec plusieurs initialisations aléatoires des centres, puis à retenir la partition donnant le plus petit critère d'inertie intraclasse.

Variantes des centres-mobiles Plusieurs variantes ont été proposées. Alors que l'algorithme des centres-mobiles tend généralement à détecter des classes de forme sphérique (voir équation II.1), la méthode des nuées dynamiques [46] permet de s'adapter à des formes de classes diverses. L'algorithme des *centres-mobiles flous* (*fuzzy k-means* ou *fuzzy c-means*) [11] est une variante permettant d'obtenir une classification floue. Il est basé sur la minimisation du critère :

$$J_m(U, C, \mu) = \sum_{k=1}^{K} \sum_{i \in C_k} u_{ik}^m ||x_i - \mu_k||^2 \tag{II.3}$$

où m est un réel supérieur à 1 (fixé) et u_{ik} représente le degré d'appartenance de l'observation x_i au groupe k ($\forall i \in \mathcal{I},\ \sum_k u_{ik} = 1$ et $\forall i \in \mathcal{I}, \forall k \in [\![1, K]\!],\ u_{ik} \in [0, 1]$). L'algorithme est itératif. Au cours d'une itération, les centres sont mis à jour par :

$$\forall k \in [\![1, K]\!],\ \mu_k = \frac{\sum_{i=1}^{n} u_{ik}^m x_i}{\sum_{i=1}^{n} u_{ik}^m} \tag{II.4}$$

puis les degrés d'appartenance par :

$$\forall i \in \mathcal{I},\ \forall k \in [\![1, K]\!],\ u_{ik} = \left(\sum_{l=1}^{K} \left(\frac{||x_i - \mu_k||}{||x_i - \mu_l||} \right)^{\frac{2}{m-1}} \right)^{-1} \tag{II.5}$$

Le choix de l'exposant m reste une question délicate. Ce paramètre contrôle le degré de flou de la partition produite : plus m est grand, plus la partition sera floue. Dans le cas extrême $m = 1$, on retrouve une partition dure et l'algorithme des centres-mobiles. En pratique, la valeur 2 est souvent utilisée.

1.3 Classification supervisée par SVM

La classification supervisée suppose que l'on dispose de certaines données étiquetées (dont nous connaissons les classes), à partir desquelles une règle de décision est construite. L'application de cette règle permet ensuite de classer des données non étiquetées.
Les *Séparateurs à Vaste Marge* (*Support Vector Machine*, SVM) sont une famille de classifieurs binaires (donc ne permettant de séparer que deux groupes C_1 et C_2) introduite au milieu des années 90 [124] et qui connaît un franc succès dans la communauté de l'apprentissage (*machine learning*). Le principe des SVM est de déterminer le meilleur hyperplan séparant les deux groupes de données étiquetées, de sorte que cette frontière soit de marge maximale, c'est-à-dire telle que la distance des deux groupes à la frontière soit maximale. Soit $\mathbf{x}^{app}$ les données étiquetées (ou données d'apprentissage) et $\mathbf{x}^{test}$ les données non étiquetées à classer. Notons encore $\mathcal{I}^{app}$ l'ensemble des individus d'apprentissage (dont on connaît la classe). La règle de décision des SVM consiste à affecter une donnée x_i^{test} à l'un des deux groupes selon le signe de la quantité :

$$M(x_i^{test}|\boldsymbol{\alpha}, \beta, \gamma) = \sum_{j \in \mathcal{I}^{app}} \alpha_j \omega_j \kappa(x_i^{test}, x_j^{app}|\gamma) + \beta \tag{II.6}$$

où :

- $\boldsymbol{\alpha} = (\alpha_i)_{i \in \mathcal{I}^{app}}$ et β sont les coefficients des vecteurs supports, solution d'un problème d'optimisation convexe
- $\omega_j = 1$ si l'observation d'apprentissage est dans la classe C_1, $\omega_j = -1$ sinon
- $\kappa(.,.|\gamma)$ est une fonction noyau paramétrée par un paramètre γ contraignant le problème d'optimisation et qui doit être réglé par l'utilisateur. A titre d'exemple, le noyau RBG (*Radial Basic Function*) de paramètre $\gamma > 0$ a pour expression :

$$\kappa(x, x'|\gamma) = \exp(-\gamma||x - x'||^2).$$

Le succès de ce type de méthode est principalement dû à leurs performances. Notons en particulier que, par l'utilisation d'une fonction noyau κ dans (II.6), l'approche par SVM permet de résoudre des problèmes de classification non linéairement séparables. Néanmoins, les SVM sont des méthodes dont l'apprentissage est relativement coûteux en temps de calcul (la complexité est polynômiale en n^{app}, le nombre de données d'apprentissage) et dont la paramétrage (choix du noyau, paramètres du noyau) est souvent difficile. Enfin les SVM ne peuvent séparer que deux classes. Pour un nombre de classes plus élevé, une solution est de construire tous les classifieurs possibles entre deux groupes et d'affecter une observation non étiquetée au groupe ayant remporté le plus de matchs "un contre tous" ou "un contre un" [55]. On peut encore utiliser une formulation étendue des SVM pour générer les K classes en même temps [24, 38]. Une règle de décision M_k est alors calculée pour chaque classe k et le problème d'optimisation SVM considère toutes les règles de décision en même temps. Les contraintes sont également relâchées : au lieu de forcer les règles de décision à être nulles sur la frontière de décision, il est désormais suffisant que, pour toute donnée d'apprentissage de la classe k, la quantité M_k soit supérieure à toutes les autres valeurs $M_{k'}$ calculées sur les classes $k' \neq k$. Néanmoins, les résultats de la classification ne sont pas nettement améliorées et l'optimisation devient très compliquée.

Nous avons présenté les approches non probabiliste les plus utilisées en classification. Une approche alternative pour une telle problématique est l'approche probabiliste sous

laquelle observations et classes sont supposées être des réalisations de variables aléatoires. C'est ce type d'approche que nous adoptons dans cette thèse. Nous présentons dans les prochaines sections les principes d'une classification probabiliste, puis développons le modèle de mélange indépendant sous lequel les individus sont supposés indépendants, avant de présenter le modèle par champ de Markov caché permettant la prise en compte des dépendances.

2 Classification probabiliste

2.1 Notion de variable cachée

Soit n observations réelles D-dimensionnelles, notées $\mathbf{x} = (x_1, \cdots, x_n)$ que l'on souhaite classer. Notons $\mathcal{K} = [\![1, K]\!]$ l'ensemble de classes (*labels*) dans lesquelles on désire ranger ces observations. Dans une approche probabiliste, les observations $\mathbf{x} = (x_1, \cdots, x_n)$ sont supposées être des réalisations de n vecteurs aléatoires $\mathbf{X} = (X_1, \cdots, X_n)$. Il s'agit d'associer à chacune des observations x_i une classe notée $z_i \in \mathcal{K}$ qui peut être vue comme la réalisation d'une variable aléatoire discrète $Z_i \in \mathcal{K}$. Les classes $\mathbf{Z} = (Z_1, \cdots, Z_n)$ étant inconnues (le but de la classification est de les déterminer), on les appelle *variables cachées*. La distribution des observations est alors donnée par :

$$P(\mathbf{x}) = \sum_{\mathbf{z}} P(\mathbf{z}) P(\mathbf{x}|\mathbf{z}) \tag{II.7}$$

Lorsque les variables cachées $\mathbf{Z} = (Z_1, \cdots, Z_n)$ sont supposées indépendantes les unes des autres et sous l'hypothèse de bruit indépendant ($P(\mathbf{x}|\mathbf{z}) = \prod_i P(x_i|z_i)$), la distribution (II.7) est celle d'un mélange indépendant (voir section 3). Lorsqu'elles suivent une distribution markovienne, on obtient un champ de Markov caché (voir section 4).

2.2 Classification optimale - notion de coût

Lorsque les classes des données sont inconnues, la notion de "bonne classification" reste subjective. La quantification de l'erreur et de la qualité de la classification estimée dépendent de la fonction de coût d'erreur choisie. Nous détaillons ici deux méthodes de classification correspondant à la minimisation de deux coûts moyens différents.

2.2.1 Coût de classification

Il faut dans un premier temps définir une fonction qui indique le coût d'une classification $\mathbf{z}$ lorsque la vraie classification est $\mathbf{z}^*$.

Définition 6. *Une fonction de coût est une fonction* $c : \mathcal{K}^n \times \mathcal{K}^n \to \mathbb{R}^+$ *telle que :*

$$\begin{aligned} \forall \mathbf{z}, \mathbf{z}^* \in \mathcal{K}^n,\ c(\mathbf{z}, \mathbf{z}^*) &\geq 0 \\ c(\mathbf{z}, \mathbf{z}^*) &= 0 \Leftrightarrow \mathbf{z} = \mathbf{z}^* \end{aligned}$$

La vraie classification $\mathbf{z}^*$ étant inconnue, pour estimer une classification à partir des données $\mathbf{x}$, on s'appuie sur le coût ou risque conditionnel $\bar{c}(.|\mathbf{x})$:

Définition 7. *Le coût ou risque conditionnel est la fonction* $\bar{c}(.|\mathbf{x}) : \mathcal{K}^n \rightarrow \mathbb{R}^+$ *égale au coût moyen de l'estimation de la vraie classification par la configuration* $\mathbf{z}$ *pour les observations* $\mathbf{x}$ *:*

$$\forall \mathbf{z} \in \mathcal{K}^n, \bar{c}(\mathbf{z}|\mathbf{x}) = \langle c(\mathbf{z}, \mathbf{Z})|\mathbf{x}\rangle = \sum_{\mathbf{z}^*} c(\mathbf{z}, \mathbf{z}^*) P(\mathbf{Z} = \mathbf{z}^*|\mathbf{X} = \mathbf{x})$$

Du point de vue de la classification, une *règle de décision* ou *stratégie* est une fonction s associant, à un ensemble d'observations $\mathbf{x} \in (\mathbb{R}^D)^n$, une classification $s(\mathbf{x}) \in \mathcal{K}^n$.

Définition 8. *Etant donnée une stratégie* $s : (\mathbb{R}^D)^n \rightarrow \mathcal{K}^n$*, le risque global est défini comme la moyenne du risque conditionnel :*

$$c \;=\; \langle \bar{c}(s(\mathbf{X})|\mathbf{X})\rangle = \int_{\mathbf{x}} \bar{c}(s(\mathbf{x})|\mathbf{x}) P(\mathbf{x}) d\mathbf{x}$$

Pour les observations $\mathbf{x}$, la classification optimale consiste alors à choisir la valeur de $\mathbf{z}$ qui minimise le coût conditionnel $\bar{c}(\mathbf{z}|\mathbf{x})$. En sections 2.2.2 et 2.2.3, nous décrivons deux stratégies associées à des coûts c différents, définissant les estimateurs les plus utilisés dans la littérature : l'estimateur du *maximum a posteriori* (MAP) et l'estimateur du *maximum des probabilités marginales* (MPM).

2.2.2 Maximum a posteriori

La fonction de coût la plus simple et la plus utilisée est appelée coût $0-1$. Elle associe un coût 0 à la bonne décision, 1 à la mauvaise :

$$c(\mathbf{z}, \mathbf{z}^*) = \mathbb{1}_{\mathbf{z} \neq \mathbf{z}^*} = 1 - \mathbb{1}_{\mathbf{z} = \mathbf{z}^*} = \begin{cases} 1 & \text{si } \exists i \text{ tel que } z_i \neq z_i^* \\ 0 & \text{sinon} \end{cases}$$

Le coût conditionnel est alors :

$$\bar{c}(\mathbf{z}|\mathbf{x}) = \sum_{\mathbf{z}^*} c(\mathbf{z}, \mathbf{z}^*) P(\mathbf{z}^*|\mathbf{x}) = 1 - P(\mathbf{z}|\mathbf{x}) = P(\mathbf{Z} \neq \mathbf{z}|\mathbf{x})$$

Cette quantité s'interprète comme la probabilité de se tromper en choisissant la classification $\mathbf{z}$ au lieu de la vraie classification. La règle de décision du coût $0-1$ revient alors à choisir la configuration la plus probable conditionnellement aux données :

$$\mathbf{z}^{map} = \arg\max_{\mathbf{z}} P(\mathbf{z}|\mathbf{x})$$

On appelle cette règle celle du *maximum a posteriori* (MAP).

2.2.3 Maximum des probabilités marginales

La méthode du MAP est une méthode globale, elle associe le même coût $c(\mathbf{z}, \mathbf{z}^*)$ à une configuration $\mathbf{z}$ différant en un seul site de la vraie classification $\mathbf{z}^*$ qu'à une configuration quelconque. Il peut être plus intéressant de maximiser la proportion moyenne de pixels bien classés et de définir la fonction coût comme une somme de coûts locaux :

$$c(\mathbf{z}, \mathbf{z}^*) = \sum_{i \in \mathcal{I}} \mathbb{1}_{z_i \neq z_i^*}$$

La règle de classification optimale consiste alors à choisir pour le site $i \in \mathcal{I}$ la classe :

$$z_i^{mpm} = \arg\max_{z_i} P(z_i|\mathbf{x}) \tag{II.8}$$

L'estimateur $\mathbf{z}^{mpm} = (z_1^{mpm}, \cdots, z_n^{mpm})$ correspondant porte le nom d'estimateur du *maximum des probabilités marginales* (MPM) [10].

Remarque. *Lorsque $P(\mathbf{z}|\mathbf{x}) = \prod_{i\in\mathcal{I}} P(z_i|\mathbf{x})$, les règles du MAP et du MPM sont équivalentes. C'est en particulier le cas sous l'hypothèse de mélange indépendant (section 3). Dans le cas général, en particulier sous un modèle markovien (section 4), ces deux règles ne sont pas équivalentes.*

3 Modèle à variables indépendantes - Mélange indépendant

3.1 Distribution de mélange

Du point de vue de la classification, les règles définies en section 2.2 nécessistent de définir la loi *a posterori* $P(\mathbf{z}|\mathbf{x})$. Pour cela, le modèle le plus simple est celui de mélange indépendant, largement utilisé dans les problèmes de classification.

Définition 9. *On parlera de mélange indépendant si le couple $(\mathbf{X}, \mathbf{Z})$ suit une loi définie par :*

$$P(\mathbf{z}) = \prod_{i\in\mathcal{I}} P(z_i) \tag{II.9}$$

$$\text{et } P(\mathbf{x}|\mathbf{z}) = \prod_{i\in\mathcal{I}} P(x_i|z_i) \tag{II.10}$$

Le modèle de mélange suppose que les classes $\mathbf{Z} = \{Z_i, i \in \mathcal{I}\}$ sont indépendantes (équation II.9). L'équation (II.10) est appelée *hypothèse de bruit indépendant*. Comme souligné dans [6], cette hypothèse de bruit indépendant, se décompose en :

(B1) $P(x_i|\mathbf{z}) = P(x_i|z_i)$ pour tout $i \in \mathcal{I}$

(B2) $P(\mathbf{x}|\mathbf{z}) = \prod_{i\in\mathcal{I}} P(x_i|\mathbf{z})$

L'hypothèse (B1) signifie que, conditionnellement à Z_i, l'observation X_i au site $i \in \mathcal{I}$ est indépendante des classes Z_j, $j \neq i$. L'hypothèse (B2) revient à dire que les observations $\mathbf{X}$ sont indépendantes conditionnellement aux classes $\mathbf{Z}$.

Sous les hypothèses (II.9) et (II.10), la loi des observations est :

$$P(\mathbf{x}) = \prod_{i\in\mathcal{I}} P(x_i) = \prod_{i\in\mathcal{I}} \sum_{z_i\in\mathcal{K}} P(x_i|z_i)P(z_i).$$

Les $x_1, \cdots, x_n$ peuvent alors être vues comme des réalisations indépendantes suivant chacune une loi de mélange :

$$P(x_i) = \sum_{z_i\in\mathcal{K}} P(x_i|z_i)P(z_i).$$

Sous les hypothèses (II.9) et (II.10), la distribution *a posterori* $P(\mathbf{z}|\mathbf{x})$ est alors donnée par :

$$P(\mathbf{z}|\mathbf{x}) = \prod_{i\in\mathcal{I}} P(z_i|x_i). \tag{II.11}$$

Pour retrouver la définition classique du mélange indépendant, il faut encore supposer que les classes Z_i sont identiquement distribuées, c'est-à-dire que $P(z_i)$ ne dépend pas de i et on peut alors noter π_k la probabilité $P(Z_i = k)$ (avec, pour tout $k \in \mathcal{K}$, $\pi_k \in [0,1]$ et $\sum_{k\in\mathcal{K}} \pi_k = 1$). De même, $P(x_i|z_i)$ est supposée égale à $f(x_i|\theta_{z_i})$ avec θ_k le paramètre de la loi dans la classe $Z_i = k$. L'hypothèse de bruit indépendant (II.10) s'écrit alors :

$$P(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) = \prod_{i\in\mathcal{I}} f(x_i|\theta_{z_i}) \tag{II.12}$$

où $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ désigne les paramètres de la loi $P(.|\mathbf{z})$ des observations conditionnellement aux classes $\mathbf{z}$. Notons encore $\boldsymbol{\psi} = (\psi_1, \cdots, \psi_K) = (\pi_1, \theta_1, \cdots, \pi_k, \theta_k)$ l'ensemble des paramètres du mélange. La loi (II.7) s'écrit alors :

$$P(\mathbf{x}|\boldsymbol{\psi}) = \prod_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} \pi_k f(x_i|\theta_k) = \prod_{i\in\mathcal{I}} P(x_i|\boldsymbol{\psi}). \tag{II.13}$$

La loi jointe des variables $(\mathbf{X}, \mathbf{Z})$ est donnée par :

$$P(\mathbf{x},\mathbf{z}|\boldsymbol{\psi}) = \prod_{i\in\mathcal{I}} \pi_{z_i} f(x_i|\theta_{z_i}) = \prod_{i\in\mathcal{I}} P(x_i, z_i|\psi_{z_i}) \tag{II.14}$$

3.2 Modèle gaussien

Lorsque les données sont à valeurs réelles continues, et en l'absence de connaissances particulières, il est courant de supposer que chaque classe suit une loi gaussienne. La loi gaussienne modélise en effet de façon adéquate un grand nombre de phénomènes aléatoires. Conditionnellement à l'appartenance à la classe k, la densité en un point $\mathbf{x} \in \mathbb{R}^D$ est alors donnée par

$$f(\mathbf{x}|\theta_k) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} \exp(-\frac{1}{2}(\mathbf{x}-\mu_k)\Sigma_k^{-1}(\mathbf{x}-\mu_k)') \tag{II.15}$$

où $|.|$ désigne le déterminant. La distribution de la classe k est donc la loi normale $\mathcal{N}(\mu_k, \Sigma_k)$ paramétrée par son vecteur moyenne μ_k de dimension D et sa matrice de covariance Σ_k (symétrique définie positive) de dimension $D \times D$.

D'autres choix sont possibles concernant la forme des distributions $f(.|\theta_k)$. Par exemple lorsque les données sont binaires ($x_i \in \{0,1\}^D$), un modèle simple consiste à supposer que $f(.|\theta_k)$ est le produit de D loi de Bernoulli indépendantes [64]. En présence de données aberrantes, $f(.|\theta_k)$ peut être le produit de D lois de Laplace. Notons que, de manière générale, la très grande majorité des articles traitant des données multidimensionnelles supposent l'indépendance des D dimensions, c'est-à-dire que $f(\mathbf{x}|\theta_k) = \prod_{d=1}^{D} f(x_i|\theta_k)$. La loi gaussienne présente l'avantage de relâcher cette hypothèse d'indépendance des dimensions lorsque la matrice de covariance Σ_k est non diagonale. Citons encore la loi de Laplace multivariée généralisée de [48].

3.3 Classification à paramètres connus - Règle du MAP/MPM

Cas général. Dans le cas d'un mélange indépendant, la factorisation (II.11) implique que les règles de classement du MAP (2.2.2) et du MPM (2.2.3) sont équivalentes. Une telle règle revient à choisir en chaque site i la classe la plus probable connaissant l'observation x_i :

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{z_i} P(z_i|x_i, \psi_{z_i}) = \arg\max_{z_i} \pi_{z_i} f(x_i|\theta_{z_i}) \tag{II.16}$$

Cas gaussien. Dans le cas du modèle gaussien (II.15), en passant au logarithme, la règle de classement (II.16) devient ([109], p.468)

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{k} \log(\pi_k) - \frac{1}{2}(x_i - \mu_k)\Sigma_k^{-1}(x_i - \mu_k)' - \frac{1}{2}\log(|\Sigma_k|) \tag{II.17}$$

Lorsque les Σ_k sont différents, cette règle est quadratique et il faut comparer K fonctions quadratiques de x_i. Si $\Sigma_1 = \cdots = \Sigma_K = \Sigma$, en éliminant $x_i\Sigma_k^{-1}x_i'$ et $\log(|\Sigma_k|)$ qui ne dépendent pas de la classe, la règle (II.16) devient linéaire en x_i :

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{k} \log(\pi_k) + x_i\Sigma^{-1}\mu_k' - \frac{1}{2}\mu_k\Sigma^{-1}\mu_k' \tag{II.18}$$

3.4 Estimation par l'algorithme EM

En général les paramètres $\boldsymbol{\psi}$ de la loi jointe (II.14) ne sont pas connus, seules sont disponibles les observations $\mathbf{x}$. Cette section présente l'algorithme EM [44] pour l'estimation des paramètres $\boldsymbol{\psi}$ à partir des données $\mathbf{x}$.

3.4.1 Condition d'identifiabilité d'un mélange

On ne peut estimer les paramètres d'un modèle que si celui-ci est *identifiable* c'est-à-dire si deux jeux de paramètres différents donnent lieu à deux distributions différentes. Dans le cas d'un modèle de mélange indépendant, la condition d'identifiabilité classique s'écrit :

Deux mélanges ont même fonction densité, *i.e* $\displaystyle\sum_{k=1}^{K} \pi_k f(.|\theta_k) = \sum_{h=1}^{K'} \pi'_h f(.|\theta'_h)$,
si et seulement si $K = K'$ et $\forall k \in [\![1, K]\!],\ \pi_k = \pi'_k$ et $\theta_k = \theta'_k$

Il est facile de voir qu'un modèle de mélange de lois appartenant à la même famille paramétrique de distributions (par exemple des lois gaussiennes) est non-identifiable puisqu'une telle loi de mélange est invariante par renumérotation des indices : si σ est une permutation de $[\![1, K]\!]$,

$$\sum_{k=1}^{K} \pi_k f(.|\theta_k) = \sum_{k=1}^{K} \pi_{\sigma(k)} f(.|\theta_{\sigma(k)}).$$

Dans un cadre bayésien, ce problème est connu sous le nom de *label-switching* [132]. Pour y remédier, des contraintes artificielles d'identifiabilité peuvent être imposées, par exemple ordonner les classes selon l'ordre des proportions ($\pi_1 \leq \cdots \leq \pi_K$) [3] ou des moyennes

($\mu_1 \leq \cdots \leq \mu_K$), ce qui permet de briser la symétrie de la loi de mélange. D'autres solutions plus sophistiquées ont été proposées, notamment dans [118], [26] ou [119] et permettant d'ordonner les classes au cours de l'algorithme d'estimation (méthodes *on-line*).

Notons néanmoins que ce problème de *label-switching* est en réalité un faux problème dans le cas d'une estimation non-supervisée des paramètres par maximum de vraisemblance. En effet, une permutation des classes n'a pas d'effet sur la valeur du maximum de vraisemblance ni sur l'objectif final de classification. Il s'agit de trouver la "meilleure" distribution de mélange par rapport aux données et tous les $K!$ jeux de paramètres solutions donnent la même distribution. Reste à s'assurer que, à l'ordre près des classes, le modèle de mélange admet bien une seule décomposition. A titre d'exemple, cette propriété est vraie pour les mélanges gaussiens (et la plupart des mélanges), mais faux pour les mélanges de lois uniformes [132]. On pourra également se reporter en Annexe 1.

3.4.2 Maximum de vraisemblance

Le principe du maximum de vraisemblance est l'un des plus utilisés pour estimer les paramètres $\boldsymbol{\psi}$ d'une distribution à partir d'une réalisation $\mathbf{x}$ d'un échantillon $\mathbf{X}$. Il consiste à déterminer les paramètres $\boldsymbol{\psi}^{MV}(\mathbf{x})$ maximisant la vraisemblance $L(\boldsymbol{\psi}, \mathbf{x}) = P(\mathbf{x}|\boldsymbol{\psi})$, ou encore la log-vraisemblance $l(\boldsymbol{\psi}, \mathbf{x}) = \log P(\mathbf{x}|\boldsymbol{\psi})$:

$$\boldsymbol{\psi}^{MV}(\mathbf{x}) = \arg\max_{\boldsymbol{\psi}} \log P(\mathbf{x}|\boldsymbol{\psi})$$

L'estimateur $\boldsymbol{\psi}^{MV}(\mathbf{x})$, s'il existe, est consistant, sans biais et asymptotiquement gaussien. Dans le cas d'un modèle de mélange, d'après (II.13), la log-vraisemblance s'écrit :

$$\log P(\mathbf{x}|\boldsymbol{\psi}) = \sum_{i \in \mathcal{I}} \log(\sum_{k \in \mathcal{K}} \pi_k f(x_i|\theta_k)). \tag{II.19}$$

L'équation (II.19) n'ayant pas une forme simple, il peut être plus judicieux de s'intéresser à la log-vraisemblance $l_c(\boldsymbol{\psi}, \mathbf{x}, \mathbf{z}) = \log P(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi})$ des données complétées $(\mathbf{x}, \mathbf{z})$:

$$\log P(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}) = \sum_{i \in \mathcal{I}} \log(\pi_{z_i} f(x_i|\theta_{z_i}))$$

et de chercher les paramètres $\boldsymbol{\psi}^{MVC}(\mathbf{x}, \mathbf{z})$ maximisant cette log-vraisemblance complétée :

$$\boldsymbol{\psi}^{MVC}(\mathbf{x}, \mathbf{z}) = \arg\max_{\boldsymbol{\psi}} \log P(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}).$$

3.4.3 L'algorithme EM - Cas général

L'algorithme EM [44] est un algorithme d'estimation de paramètres, et non un algorithme de classification. Son cadre est celui des modèles à données incomplètes, c'est-à-dire pour lesquels certaines observations sont manquantes. Le problème de la classification de données $\mathbf{x}$ dans des classes $\mathbf{z}$ fait partie des problèmes à données manquantes : seuls sont observés les $\mathbf{x}$, les classes $\mathbf{z}$ sont manquantes. L'algorithme EM n'étant pas un algorithme de classification, la configuration recherchée $\mathbf{z}$ doit être obtenue dans un second temps à partir des paramètres estimés par EM (voir section 3.4.5). Son principe repose sur les deux idées suivantes :

1. Il est en général plus facile de calculer l'estimateur $\boldsymbol{\psi}^{MVC}(\mathbf{x}, \mathbf{z})$ sur les données complétées que l'estimateur de maximum de vraisemblance $\boldsymbol{\psi}^{MV}(\mathbf{x})$ (comme c'est le cas pour les modèles de mélange).
2. Les classes $\mathbf{z}$ étant inconnues, la log-vraisemblance complète $\log P(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi})$ est remplacée par son espérance conditionnellement aux observations, $\langle \log P(\mathbf{x}, \mathbf{Z}|\boldsymbol{\psi})|\mathbf{x}, \boldsymbol{\psi}\rangle$.

L'algorithme EM est une procédure itérative partant d'une valeur initiale $\boldsymbol{\psi}^{(0)}$ des paramètres. L'itération $(q+1)$ consiste alors à calculer les nouveaux paramètres $\boldsymbol{\psi}^{(q+1)}$ à partir de ceux $\boldsymbol{\psi}^{(q)}$ de l'itération précédente de façon à maximiser la fonction $Q(.|\boldsymbol{\psi}^{(q)})$:

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)}) = \langle \log P(\mathbf{x}, \mathbf{Z}|\boldsymbol{\psi})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle \tag{II.20}$$

L'itération $(q+1)$ de l'algorithme EM se décompose alors en deux étapes :

(E) Calcul des termes de $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)})$ ne faisant pas intervenir $\boldsymbol{\psi}$

(M) Mise à jour des paramètres par :

$$\boldsymbol{\psi}^{(q+1)} = \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)}).$$

Cette procédure a pour propriété fondamentale de faire croître la vraisemblance au cours des itérations :

$$\forall q \in \mathbb{N},\ L(\boldsymbol{\psi}^{(q+1)}, \mathbf{x}) \geq L(\boldsymbol{\psi}^{(q)}, \mathbf{x})$$

Convergence. Sous des conditions suffisantes de régularité, l'estimateur obtenu converge vers un maximum local de la vraisemblance [130]. Néanmoins, la valeur de l'estimateur à la convergence peut dépendre fortement de la position initiale. De plus, l'algorithme peut se retrouver bloqué dans un point selle ou un plateau de la vraisemblance $L(\boldsymbol{\psi}, \mathbf{x})$. Des perturbations stochastiques de EM ont été proposées, comme l'algorithme *Stochastic EM* [27] ou *Simulated Annealing EM* [28], permettant de réduire le risque de tomber dans un maximum local de vraisemblance.

Stratégie d'initialisation. Il n'y a pas de solution universelle pour palier la limitation de la dépendance vis à vis de l'initialisation. Néanmoins, plusieurs stratégies d'initialisation ont été proposées. Une stratégie courante est d'effectuer un nombre fixe (petit) d'itérations de r algorithmes EM initialisés aléatoirement et de choisir comme initialisation celle, parmi les r initialisations effectuées, associée à la plus grande vraisemblance. MacLachlan et Peel [89] proposent une autre stratégie d'initialisation dans le cas gaussien : les proportions sont supposées égales et les moyennes sont générées selon une loi normale de moyenne et covariance empiriques calculées sur l'ensemble des observations $\mathbf{x}$. Une autre solution est de déterminer une pré-classification $\mathbf{z}^{pre}$ (par les centres mobiles par exemple) et d'initialiser les paramètres par l'estimateur de maximum de vraisemblance complétée $\boldsymbol{\psi}^{MVC}(\mathbf{x}, \mathbf{z}^{pre})$.

3.4.4 L'algorithme EM - Cas du mélange

Dans le cas d'un mélange indépendant de loi jointe (II.14), à l'itération $(q+1)$, la fonction Q (II.20) s'écrit :

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)}) = \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} t_{ik}^{(q)} \log(\pi_k f(x_i|\theta_k))$$

où le coefficient $t_{ik}^{(q)}$ désigne la probabilité *a posteriori* $P(Z_i = k|x_i, \boldsymbol{\psi}^{(q)})$. Les deux étapes de cette itération $(q+1)$ sont alors :

(E) Calcul des probabilités *a posteriori* pour tout $i \in \mathcal{I}$ et $k \in \mathcal{K}$:

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} f(x_i|\theta_k^{(q)})}{\sum\limits_{k'\in\mathcal{K}} \pi_{k'}^{(q)} f(x_i|\theta_{k'}^{(q)})}$$

(M) Mise à jour des proportions $(\pi_k)_{k\in\mathcal{K}}$ et des paramètres $\boldsymbol{\theta} = (\theta_k)_{k\in\mathcal{K}}$ des densités $f(.|\theta_k)$:

$$\begin{aligned} \pi_k^{(q+1)} &= \frac{\sum\limits_{i\in\mathcal{I}} t_{ik}^{(q)}}{n} \\ \boldsymbol{\theta}^{(q+1)} &= \arg\max_{\boldsymbol{\theta}} \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} t_{ik}^{(q)} \log f(x_i|\theta_k) \end{aligned} \tag{II.21}$$

Cas gaussien Dans le cas du modèle gaussien (II.15), la mise à jour des paramètres θ_k des densités $f(.|\theta_k)$ est explicite :

$$\mu_k^{(q+1)} = \frac{\sum\limits_{i\in\mathcal{I}} t_{ik}^{(q)} x_i}{\sum\limits_{i\in\mathcal{I}} t_{ik}^{(q)}} \tag{II.22}$$

$$\Sigma_k^{(q+1)} = \frac{\sum\limits_{i\in\mathcal{I}} t_{ik}^{(q)} (x_i - \mu_k^{(q+1)})(x_i - \mu_k^{(q+1)})'}{\sum\limits_{i\in\mathcal{I}} t_{ik}^{(q)}} \tag{II.23}$$

$\mu_k^{(q+1)}$ et $\Sigma_k^{(q+1)}$ s'interprètent alors comme les moyenne et variance empiriques calculées sur les observations $(x_i)_{i\in\mathcal{I}}$ affectées des poids $(t_{ik})_{i\in\mathcal{I}}$.

3.4.5 Classer les données suite à l'algorithme EM

L'algorithme EM est un algorithme général d'estimation des paramètres en présence de données non observées. Une fois les paramètres estimés, la classification $\mathbf{z}$ recherchée peut être obtenue par application des règles du MAP ou MPM (voir section 2.2) sans aucun calcul supplémentaire. En effet, avec les notations de la section 3.4.4, l'équation (II.16) s'écrit :

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{z_i\in\mathcal{K}} P(z_i|x_i, \psi_{z_i}) = \arg\max_{k\in\mathcal{K}} t_{ik}$$

Les valeurs des t_{ik} en sortie de l'algorithme EM permettent donc directement de restaurer les classes $\mathbf{z}^{map}$.

4 Modèle à variables dépendantes - champ de Markov caché

4.1 Distribution de champ de Markov caché

Dans un modèle de champ de Markov caché, la classification non observée $\mathbf{z}$ est supposée être la réalisation d'un champ de Markov $\mathbf{Z}$.

Définition 10. *On dit que* $(\mathbf{X}, \mathbf{Z})$ *est un champ de Markov caché si le champ caché* $\mathbf{Z}$ *est markovien.*

Pour définir entièrement le modèle, il reste alors à préciser la distribution $P(\mathbf{x}|\mathbf{z})$.

Définition 11. *On dit que* $(\mathbf{X}, \mathbf{Z})$ *est un champ de Markov caché à bruit indépendant si le champ caché* $\mathbf{Z}$ *est markovien et si le bruit est indépendant (équation II.10).*

Notons $P_G(.|\boldsymbol{\phi})$ la distribution de Gibbs de $\mathbf{Z}$, $H(.;\boldsymbol{\phi})$ son énergie et $W(\boldsymbol{\phi})$ sa fonction de partition (voir chapitre I, section 1.2), toutes trois paramétrées par un ensemble de paramètres $\boldsymbol{\phi}$:

$$P_G(\mathbf{z}|\boldsymbol{\phi}) = W(\boldsymbol{\phi})^{-1} \exp(-H(\mathbf{z};\boldsymbol{\phi})) \tag{II.24}$$

Sous l'hypothèse de bruit indépendant (II.10), la loi (II.7) s'écrit alors :

$$P(\mathbf{x}|\boldsymbol{\psi}) = \sum_{\mathbf{z}} \left(P_G(\mathbf{z}|\boldsymbol{\phi}) \prod_{i \in \mathcal{I}} f(x_i|\theta_{z_i}) \right) \tag{II.25}$$

où le vecteur $\boldsymbol{\psi} = (\psi_1, \cdots, \psi_K) = (\theta_1, \cdots, \theta_K, \boldsymbol{\phi})$ dénote les paramètres du champ de Markov caché. La loi jointe du couple $(\mathbf{X}, \mathbf{Z})$ est donnée par :

$$P(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}) = P_G(\mathbf{z}|\boldsymbol{\phi}) \prod_{i \in \mathcal{I}} f(x_i|\theta_{z_i}) \propto \exp(-H(\mathbf{z};\boldsymbol{\phi}) + \sum_{i \in \mathcal{I}} \log f(x_i|\theta_{z_i}))$$

D'après le théorème d'Hammersley-Clifford (Théorème 3, chapitre I, section 1.2), on en déduit la proposition :

Proposition 12. *Soient* $\mathbf{X} = (X_i)_{i \in \mathcal{I}}$ *et* $\mathbf{Z} = (Z_i)_{i \in \mathcal{I}}$ *deux champs. Il y a équivalence entre les affirmations :*

(1) Le couple $(\mathbf{X}, \mathbf{Z})$ *est un champ de Markov caché à bruit indépendant*

(2) Le couple $(\mathbf{X}, \mathbf{Z})$ *est un champ de Markov d'énergie* $H(\mathbf{z};\boldsymbol{\phi}) - \sum_{i \in \mathcal{I}} \log f(x_i|\theta_{z_i})$

Enfin, en appliquant la règle de Bayes, le champ $\mathbf{Z}|\mathbf{x}$ des classes $\mathbf{Z}$ conditionnellement aux observations $\mathbf{X} = \mathbf{x}$ est également un champ de Markov d'énergie :

$$H(\mathbf{z}|\mathbf{x};\boldsymbol{\psi}) = H(\mathbf{z};\boldsymbol{\phi}) - \sum_{i \in \mathcal{I}} \log f(x_i|\theta_{z_i}) \tag{II.26}$$

et de fonction de partition :

$$W(\mathbf{x}, \boldsymbol{\psi}) = \sum_{\mathbf{z}} \left(\exp(-H(\mathbf{z};\boldsymbol{\phi})) \prod_{i \in \mathcal{I}} f(x_i|\theta_{z_i}) \right)$$

C'est sur cette distribution $P(\mathbf{z}|\mathbf{x})$ que s'appliquent les règles de classification du MAP et du MPM (voir section 2.2). Se placer sous l'hypothèse de champ de Markov caché à bruit indépendant nous assure alors que la distribution *a posteriori* $P(\mathbf{z}|\mathbf{x})$ est markovienne. Notons néanmoins que, pour avoir $P(\mathbf{z}|\mathbf{x})$ markovienne, cette hypothèse est suffisante mais non nécessaire. Une hypothèse moins forte est alors de supposer directement que le couple $(\mathbf{X}, \mathbf{Z})$ est markovien (sans que $\mathbf{Z}$ soit nécessairement markovien). En effet, il est facile de voir que :

Proposition 13. *Si le couple* $(\mathbf{X}, \mathbf{Z})$ *est un champ de Markov, alors les champs conditionnels* $\mathbf{Z}|\mathbf{x}$ *et* $\mathbf{X}|\mathbf{z}$ *sont également des champs de Markov.*

On parle alors de champ de Markov couple [100]. Ce modèle, ainsi que son extension par champ de Markov triplet [6] seront étudiés au chapitre IV. D'ici là, **nous nous plaçons sous l'hypothèse de champ de Markov caché à bruit indépendant**.

Simulation. La simulation de données $\mathbf{x}$ selon un modèle de champ de Markov caché à bruit indépendant s'effectue en deux étapes :

1. On génère une partition $\mathbf{z}$ selon une distribution *a priori* markovienne $P_G(\mathbf{z}|\boldsymbol{\phi})$. On peut utiliser pour cela un échantillonneur de Gibbs (voir chapitre I, section 1.5).
2. On se base ensuite sur cette partition simulée pour générer des observations selon la loi $P(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$: pour tout site $i \in \mathcal{I}$, x_i est généré selon la loi $f(.|\theta_{z_i})$.

4.2 Approximation de type champ moyen - Choix des voisins

Dans le modèle de champ de Markov caché à bruit indépendant, $\mathbf{Z}$ et $\mathbf{Z}|\mathbf{x}$ sont tous deux markoviens. Leurs distributions respectives $P_G(\mathbf{z}|\boldsymbol{\phi})$ et $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi})$ ne peuvent donc être calculées de manière exacte (voir la remarque du chapitre I, section 1.2). Néanmoins, elles peuvent être approximées par application d'un principe de type champ moyen comme décrite au chapitre I, section 2.1.2. Or, connaissant la loi conditionnelle des observations $f(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$, ces deux distributions se déduisent l'une de l'autre par l'application de la règle de Bayes. L'approximation de type champ moyen ne peut donc être appliquée qu'à une seule de ces deux distributions de Gibbs, l'autre approximation en découlant. La question est alors : est-il préférable d'appliquer l'approximation de type champ moyen à la distribution marginale $P_G(\mathbf{z}|\boldsymbol{\phi})$ ou conditionnelle $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi})$? Les auteurs de [97] (p. 78) recommandent de l'appliquer à la loi conditionnelle. En effet, ce choix présente l'avantage de tenir compte des données : le champ des voisins est déterminé conditionnellement aux données $\mathbf{x}$. De plus, l'étude du modèle de Potts simple (voir chapitre I, section 1.3.2) dissuade d'utiliser l'approximation en champ moyen sur la loi *a priori* $P_G(\mathbf{z}|\boldsymbol{\phi})$. En effet, cette approximation conduit à fixer le champ des voisins à la configuration uniforme en chaque site, indépendamment du paramètre spatial β.

Notons $\tilde{\mathbf{z}}^{\mathbf{x}}$ le champ des voisins, déterminé conditionnellement aux observations $\mathbf{x}$, dans l'approximation de type champ moyen de la loi conditionnelle $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi})$:

$$P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi}) \approx \prod_{i \in \mathcal{I}} P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(z_i|\mathbf{x}, \boldsymbol{\psi}) = \prod_{i \in \mathcal{I}} P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(z_i|x_i, \boldsymbol{\psi}) = \prod_{i \in \mathcal{I}} P_G(z_i|x_i, \tilde{\mathbf{z}}^{\mathbf{x}}_{N_i}, \boldsymbol{\psi}) \qquad \text{(II.27)}$$

La loi *a priori* $P_G(\mathbf{z}|\boldsymbol{\phi})$ est alors approximée par :

$$P_G(\mathbf{z}|\boldsymbol{\phi}) \approx \prod_{i\in\mathcal{I}} P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(z_i|\boldsymbol{\phi}) = \prod_{i\in\mathcal{I}} P_G(z_i|\tilde{\mathbf{z}}^{\mathbf{x}}_{N_i}, \boldsymbol{\phi})$$

et donc la loi du couple $(\mathbf{X}, \mathbf{Z})$ par :

$$P_G(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}) \approx \prod_{i\in\mathcal{I}} P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(x_i, z_i|\boldsymbol{\psi}) = \prod_{i\in\mathcal{I}} P_G(z_i|\tilde{\mathbf{z}}^{\mathbf{x}}_{N_i}, \boldsymbol{\phi}) f(x_i|\theta_{z_i}) \tag{II.28}$$

Sous cette approximation, on est ramené à un modèle de mélange indépendant (Définition 9) pour lequel, à la différence avec la section 3, la probabilité $\tilde{\pi}_{ik} = P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(Z_i = k|\boldsymbol{\phi})$ d'occurrence de la classe k au site i dépend de i.

De manière analogue au chapitre I, section 2.1.2, nous précisons trois approximations, correspondant à différents choix pour le champ des voisins $\tilde{\mathbf{z}}^{\mathbf{x}}$:

- *Approximation en champ moyen :* fixer $\tilde{\mathbf{z}}^{\mathbf{x}}$ à l'estimation en champ moyen de l'espérance de la distribution conditionnelle $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi})$

$$\forall i \in \mathcal{I},\ \tilde{z}^{\mathbf{x}}_i = \langle Z_i \rangle_{P_G(.|x_i, \tilde{\mathbf{z}}^{\mathbf{x}}_{N_i}, \boldsymbol{\psi})}$$

- *Approximation en champ modal :* fixer $\tilde{\mathbf{z}}^{\mathbf{x}}$ à l'estimation en champ modal du mode de la distribution conditionnelle $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi})$

$$\forall i \in \mathcal{I},\ \tilde{z}^{\mathbf{x}}_i = \arg\max_{z_i} P_G(z_i|x_i, \tilde{\mathbf{z}}^{\mathbf{x}}_{N_i}, \boldsymbol{\psi})$$

- *Approximation en champ simulé :* simuler $\tilde{\mathbf{z}}^{\mathbf{x}}$ selon la distribution conditionnelle $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi})$ (par l'échantillonneur de Gibbs [58] par exemple, voir chapitre I, section 1.5).

4.3 Classification à paramètres connus - Règle du MAP/MPM

Dans cette section, nous supposons les paramètres $\boldsymbol{\psi}$ du modèle de champ de Markov caché connus. Par soucis de clarté, nous les omettons dans les écritures des distributions de cette section.

Dans le cas d'un modèle de champ de Markov caché, la règle de classement du MAP (section 2.2.2) revient à choisir la configuration $\mathbf{z}$ minimisant l'énergie du champ conditionnel $H(\mathbf{z}|\mathbf{x})$. Contrairement au cas du mélange indépendant (voir section 3.3), du fait de la non factorisation de cette expression sur les site $i \in \mathcal{I}$, cet estimateur du MAP ne peut en pratique pas être calculé car il nécessiterait le calcul de l'énergie $H(\mathbf{z}|\mathbf{x})$ pour les K^n configurations $\mathbf{z} = (z_i, \cdots, z_n)$ possibles. De même l'estimateur du MPM (voir section 2.2.3) nécessite le calcul de la loi marginale $P(z_i|\mathbf{x}) = \sum_{\mathbf{z}_{\neq i}} P_G(\mathbf{z}|\mathbf{x})$ et ne peut par conséquent être obtenu sans approximation. Nous décrivons dans les paragraphes suivants deux méthodes très utilisées pour approximer le calcul du MAP, le recuit simulé [58] et l'algorithme *Iterated Conditional Modes* (ICM) [10]. Nous présentons également les approximations de type champ moyen pour le calcul du MAP et du MPM. Enfin, nous décrivons une procédure par simulation pour le calcul du MPM.

Calcul du MAP par l'algorithme de recuit simulé L'algorithme de recuit simulé [58] est une méthode de relaxation stochastique pour le calcul du ou de l'un des modes d'une distribution de Gibbs. Dans le cadre de la recherche du MAP, il s'agit de déterminer une configuration $\mathbf{z}^{map}$ maximisant la distribution $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi})$. L'algorithme du recuit simulé introduit la distribution de probabilité suivante, dépendante d'un paramètre $\mathcal{T}$ interprétable comme une température dans un formalisme physique :

$$P_{\mathcal{T}}(\mathbf{z}|\mathbf{x}) = (P_G(\mathbf{z}|\mathbf{x}))^{1/\mathcal{T}} \propto \exp(-\frac{H(\mathbf{z}|\mathbf{x})}{\mathcal{T}}) \tag{II.29}$$

Le principe de l'algorithme est alors de simuler une chaîne suivant la distribution (II.29) en faisant décroître la température au cours des itérations. Tant que la température reste élevée, l'algorithme autorise des passages dans des configurations qui font augmenter la fonction d'énergie $H(.|\mathbf{x})$, ce qui permet d'éviter qu'il ne reste bloqué dans un minimum local. Puis lorsque la température diminue, la chaîne évolue vers le ou les minima globaux de $H(.|\mathbf{x})$. La chaîne est simulée via l'échantillonneur de Gibbs (voir chapitre I, section 1.5). Sous des conditions peu contraignantes, quelle que soit l'initialisation, la distribution de la configuration $\mathbf{Z}^{(q)}$ à l'itération (q) converge vers la distribution uniforme sur l'ensemble des minimiseurs globaux de $H(.|\mathbf{x})$ [58].

Calcul du MAP par l'algorithme ICM Le principe de l'algorithme *Iterated Conditional Modes* (ICM) [10] est de mettre à jour les sites de manière itérative et les uns après les autres à partir des observations $\mathbf{x}$ et de la configuration courante : à l'itération (q), seul le site i_q est mis à jour par :

$$z_{i_q} = \arg\max_{k \in \mathcal{K}} P(Z_{i_q} = k|\mathbf{x}, \mathbf{z}_{N_{i_q}}^{(q-1)})$$

Le calcul de ces probabilités locales ne pose pas de problème car il ne fait intervenir que les K états possibles du site considéré. Lorsque le parcours se fait de manière séquentielle, l'algorithme converge vers un maximum local de la distribution de Gibbs conditionnelle $P_G(\mathbf{z}|\mathbf{x})$. Néanmoins, la classification obtenue dépend fortement de l'initialisation choisie.

Calcul du MAP et du MPM par approximation de type champ moyen Soit $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}$ l'approximation de type champ moyen de $P_G(\mathbf{z}|\mathbf{x})$ (voir section 4.2). Du fait de la factorisation de $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}$ (équation II.27), les règles du MAP et du MPM appliquées à $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}$ sont équivalentes. Elles conduisent à choisir en chaque site i la classe la plus probable connaissant l'observation x_i :

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{z_i \in \mathcal{K}} P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(z_i|x_i) = \arg\max_{z_i \in \mathcal{K}} P_G(z_i|x_i, \tilde{z}_{N_i}^{\mathbf{x}}, \boldsymbol{\psi}) f(x_i|\theta_{z_i})$$

L'approximation en champ moyen nous fournit donc une formule explicite pour approximer le calcul du MAP et du MPM.

Remarque. *A paramètres connus, le calcul du MAP par l'algorithme ICM et par approximation en champ modal (section 4.2) sont équivalents.*

Calcul du MPM par simulation Le calcul direct des $P(z_i|\mathbf{x})$ est impossible compte tenu du gigantisme de l'espace des configurations. Néanmoins, le fait de pouvoir simuler des réalisations du champ de Markov $\mathbf{Z}|\mathbf{x}$ permet leur estimation. La procédure pour le calcul du MPM consiste alors à simuler M réalisations $(\mathbf{z}^{[1]}, \cdots, \mathbf{z}^{[M]})$ de $\mathbf{Z}$ selon la loi *a posteriori* $P_G(\mathbf{z}|\mathbf{x})$ puis à retenir, en chaque site i, la classe dont le nombre d'apparitions dans ces M simulations est le plus grand.

4.4 Estimation des paramètres

En général les paramètres $\boldsymbol{\psi}$ de la loi jointe (II.14) ne sont pas connus, seules sont disponibles les observations $\mathbf{x}$. Il est donc important de disposer de méthodes d'estimation de ces paramètres à partir des données $\mathbf{x}$. Pour estimer $\boldsymbol{\phi}$ par maximum de vraisemblance, une première idée serait, comme dans le cas du mélange indépendant (voir section 3.4), d'utiliser l'algorithme EM. Malheureusement, sous modélisation markovienne, cet algorithme ne peut être utilisé sans approximation. Nous détaillons l'algorithme NREM [29] (section 4.4.2) qui est un algorithme de type EM sous approximation de type champ moyen. Nous présentons également en section 4.4.4 d'autres algorithmes très répandus pour l'estimation des paramètres d'un champ de markov caché à bruit indépendant : l'algorithme Gradient Stochastique [135] et la procédure ICE [99].

4.4.1 Algorithme EM

L'itération $(q+1)$ de l'algorithme EM [44] (voir section 3.4.3) appliqué au champ de Markov caché $(\mathbf{X}, \mathbf{Z})$ consiste à calculer les nouveaux paramètres $\boldsymbol{\psi}^{(q+1)}$ à partir de ceux $\boldsymbol{\psi}^{(q)}$ de l'itération précédente de façon à maximiser la fonction $Q(.|\boldsymbol{\psi}^{(q)})$ définie par :

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)}) = \langle \log P_G(\mathbf{x}, \mathbf{Z}|\boldsymbol{\psi})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle \tag{II.30}$$

Puisque $P_G(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}) = P_G(\mathbf{z}|\boldsymbol{\phi}) f(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$, l'équation (II.30) se décompose en :

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)}) = Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\psi}^{(q)}) + Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\boldsymbol{\psi}^{(q)})$$

où, en notant V_c les potentiels sur les clique $c \in \mathcal{C}$ du champ de Markov $\mathbf{Z}$,

$$\begin{aligned} Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\psi}^{(q)}) &= \langle \log f(\mathbf{x}|\mathbf{Z}, \boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle = \sum_{i\in\mathcal{I}} \sum_{z_i} P_G(z_i|\mathbf{x}, \boldsymbol{\psi}^{(q)}) \log f(x_i|z_i, \boldsymbol{\theta}) \\ Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\boldsymbol{\psi}^{(q)}) &= \langle \log P_G(\mathbf{Z}|\boldsymbol{\phi})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle = -\log W(\boldsymbol{\phi}) - \sum_{c\in\mathcal{C}} \sum_{\mathbf{z}_c} P_G(\mathbf{z}_c|\mathbf{x}, \boldsymbol{\psi}^{(q)}) V_c(\mathbf{z}_c|\boldsymbol{\phi}) \end{aligned}$$

L'étape (E) requiert alors le calcul des termes de $Q_{\boldsymbol{\theta}}$ et $Q_{\boldsymbol{\phi}}$ ne faisant pas intervenir $\boldsymbol{\psi}$, à savoir respectivement $P_G(z_i|\mathbf{x}, \boldsymbol{\psi}^{(q)})$ et $P_G(\mathbf{z}_c|\mathbf{x}, \boldsymbol{\psi}^{(q)})$. Or ces expressions ne sont pas calculables dans le cas d'une distribution de Gibbs (voir la remarque du chapitre I, section 1.2). De même, l'étape (M) nécessiterait pour l'estimation des paramètres $\boldsymbol{\psi}$ de $Q_{\boldsymbol{\phi}}$ le calcul de la fonction de partition $W(\boldsymbol{\phi})$ qui ne peut être effectué. De nombreuses solutions ont été proposées pour rendre les étapes (E) et (M) réalisables. Nous détaillons dans le paragraphe suivant l'algorithme NREM [29] inspiré de l'algorithme de Zhang [136] et fondé sur une approximation de type champ moyen (section 4.2). D'autres solutions, que nous ne détaillerons pas ici, sont l'EM gibbsien [30], Monte-Carlo EM [129] ou encore les solutions proposées dans [102] et [101].

4.4.2 L'algorithme NREM

G. Celeux, F. Forbes et N. Peyrard [29] proposent de remplacer le modèle de champ de Markov caché de loi $P_G(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi})$ par une approximation de type champ moyen définie par l'équation (II.28). Sous cette approximation, on est ramené à un modèle de mélange indépendant sur lequel peut donc être appliqué l'algorithme EM. Néanmoins, le champ des voisins $\tilde{\mathbf{z}}^{\mathbf{x}}$ dépendant dans le cas le plus général des paramètres $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\phi})$, celui-ci doit également être mis à jour au cours de la procédure d'estimation. Le principe de l'algorithme NREM consiste alors à alterner une étape (NR) de choix du voisinage (*neighborhood restoration*), puis une étape (EM) d'estimation des paramètres du modèle par application de l'algorithme EM sur le mélange indépendant défini par l'approximation. Partant de valeurs initiales $\tilde{\mathbf{z}}^{\mathbf{x}}$ du champ des voisins et $\boldsymbol{\psi}^{(0)}$ des paramètres, l'itération $(q+1)$ de l'algorithme est la suivante :

(EM) **Estimation :** Mettre à jour les estimateurs $\boldsymbol{\psi}^{(q+1)}$ des paramètres en appliquant l'algorithme EM sur le modèle de mélange indépendant défini par la loi jointe

$$P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}) = \prod_{i \in \mathcal{I}} \tilde{\pi}_{iz_i} f(x_i|\theta_{z_i}) \qquad \text{où } \tilde{\pi}_{iz_i} = P_G(z_i|\tilde{\mathbf{z}}^{\mathbf{x}}_{N_i}, \boldsymbol{\phi}) \tag{II.31}$$

avec $\boldsymbol{\psi}^{(q)}$ comme valeur initiale des paramètres.

(NR) **Choix des voisins :** créer, à partir des observations $\mathbf{x}$ et de l'estimation courante $\boldsymbol{\psi}^{(q+1)}$ des paramètres, un nouveau champ des voisins $\tilde{\mathbf{z}}^{\mathbf{x}}$.

L'étape (EM) en pratique En pratique, un seul pas de l'algorithme EM est suffisant à chaque itération [29]. L'étape (EM) se résume alors à une étape (E) et une étape (M). A l'itération $(q+1)$, l'espérance Q (II.30) à maximiser, sous l'hypothèse (II.31), s'écrit :

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)}) = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \tilde{t}_{ik}^{(q)} \log(\tilde{\pi}_{ik} f(x_i|\theta_k))$$

où le coefficient $\tilde{t}_{ik}^{(q)}$ désigne la probabilité *a posteriori* $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(Z_i = k|x_i, \boldsymbol{\psi}^{(q)})$. Les deux étapes de cette itération $(q+1)$ sont alors :

(E) Calcul des probabilités *a posteriori* pour tout $i \in \mathcal{I}$ et $k \in \mathcal{K}$:

$$\tilde{t}_{ik}^{(q)} = \frac{\tilde{\pi}_{ik}^{(q)} f(x_i|\theta_k^{(q)})}{\sum_{l \in \mathcal{K}} \tilde{\pi}_{il}^{(q)} f(x_i|\theta_l^{(q)})} \qquad \text{où } \tilde{\pi}_{iz_i}^{(q)} = P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(z_i|\boldsymbol{\phi}^{(q)}). \tag{II.32}$$

(M) Mise à jour des paramètres $\boldsymbol{\phi} = (\phi_k)_{k \in \mathcal{K}}$ de la distribution $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z}|\phi)$ et des paramètres $\boldsymbol{\theta} = (\theta_k)_{k \in \mathcal{K}}$ des densités $f(.|\theta_k)$:

$$\boldsymbol{\phi}^{(q+1)} = \arg\max_{\boldsymbol{\phi}} \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \tilde{t}_{ik}^{(q)} \log \tilde{\pi}_{ik} \tag{II.33}$$

$$\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}} \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \tilde{t}_{ik}^{(q)} \log f(x_i|\theta_k) \tag{II.34}$$

Notons que l'équation (II.34) est la même que dans le cas du mélange indépendant (voir section 3.4.3), en remplaçant t_{ik} par $\tilde{t}_{ik}$. En particulier, dans le cas de densités $f(.|\theta_k) \sim \mathcal{N}(\mu_k, \Sigma_k)$, la mise à jour des paramètres est explicite et donnée par les équations (II.22) et (II.23). Par contre, même dans le cas simple du modèle de Potts (voir chapitre I, section 1.3.2), il n'y a pas de formule explicite pour la mise à jour des paramètres $\boldsymbol{\psi}$ de l'équation (II.33). Néanmoins, dans le cas du modèle de Potts étendu, ces paramètres peuvent être obtenus par une descente de gradient (voir annexe 4).

L'étape (NR) en pratique Trois choix sont naturels pour la mise à jour du champ des voisins, conduisant aux algorithmes en champ moyen, en champ modal et en champ simulé. Plus précisément, à l'itération $(q+1)$, les voisins sont fixés à $\tilde{\mathbf{z}}^{\mathbf{x}}$ définis comme décrit en section 4.2 en remplaçant dans les formules les paramètres $\boldsymbol{\psi}$ par leur estimation courante $\boldsymbol{\psi}^{(q+1)}$.

- *Algorithme en champ moyen* : fixer $\tilde{\mathbf{z}}^{\mathbf{x}}$ à l'estimation en champ moyen de l'espérance de la distribution conditionnelle $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi}^{(q+1)})$
- *Algorithme en champ modal* : fixer $\tilde{\mathbf{z}}^{\mathbf{x}}$ à l'estimation en champ modal du mode de la distribution conditionnelle $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi}^{(q+1)})$
- *Algorithme en champ simulé* : simuler $\tilde{\mathbf{z}}^{\mathbf{x}}$ selon la loi conditionnelle $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi}^{(q+1)})$, via l'échantillonneur de Gibbs.

Stratégie d'initialisation. Notons que, en plus des paramètres, le champ des voisins $\tilde{\mathbf{z}}^{\mathbf{x}}$ doit être initialisé. Des stratégies similaires à l'initialisation de l'algorithme EM (voir section 3.4.3) peuvent être utilisées. On peut par exemple effectuer un nombre fixe (petit) d'itérations de r algorithmes NREM initialisés aléatoirement et choisir comme initialisation (champ $\tilde{\mathbf{z}}^{\mathbf{x}}$ et paramètres $\boldsymbol{\psi}$) celle, parmi les r initialisations effectuées, associée à la plus grande vraisemblance. Une autre solution est de déterminer une pré-classification $\mathbf{z}^{pre}$ (par les centres mobiles par exemple), d'initialiser $\tilde{\mathbf{z}}^{\mathbf{x}}$ à cette pré-classification et les paramètres par l'estimateur de maximum de vraisemblance complétée $\boldsymbol{\psi}^{MVC}(\mathbf{x}, \mathbf{z}^{pre})$.

Convergence. Il n'existe pas de preuve de convergence de l'algorithme NREM. Notons néanmoins que, lorsque le champ $\tilde{\mathbf{z}}^{\mathbf{x}}$ ne bouge plus à l'étape (NR), l'algorithme appliqué est exactement l'algorithme EM sur la distribution (II.31), algorithme dont nous avons des preuves de convergence sous des conditions suffisantes de régularité [130]. De plus, l'approche variationnelle nous assure que l'approximation en champ moyen de $P_G(.|\mathbf{x})$ par (II.31) est optimale au sens de la divergence de Kullback-Leibler (voir Annexe 4). Notons néanmoins qu'en pratique, c'est l'algorithme en champ simulé qui donne les résultats les meilleurs [29]. On peut en effet penser que, du fait des simulations lors de l'étape (NR), l'algorithme en champ simulé arrive à s'échapper des minimas locaux de la vraisemblance. Dans ce même cadre markovien, des résultats de convergence ont été démontrés récemment pour un algorithme combinant approximation en champ moyen et simulations de Monte-Carlo (voir [50]).

4.4.3 Classer les données suite à l'algorithme NREM.

L'algorithme NREM permet donc d'estimer les paramètres d'un champ de Markov caché sous approximation de type champ moyen. Dans un second temps, comme dans le

cas de l'algorithme EM pour mélange indépendant (voir section 3.4.5), la classification par MAP ou MPM (voir section 4.3) peut-être restaurée sans calcul supplémentaire. En effet, sous l'approximation en champ moyen (II.31), on est ramené à un modèle de mélange indépendant. Le MAP, comme le MPM, conduisent à choisir en chaque site i la classe la plus probable connaissant l'observation x_i :

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{z_i \in \mathcal{K}} P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(z_i|x_i) = \arg\max_{z_i \in \mathcal{K}} \tilde{\pi}_{iz_i} f(x_i|\theta_{z_i}) \tag{II.35}$$

Avec les notations de 4.4.2, l'équation (II.35) s'écrit encore :

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{z_i \in \mathcal{K}} P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(z_i|x_i) = \arg\max_{k \in \mathcal{K}} \tilde{t}_{ik}$$

Les $\tilde{t}_{ik}$ en sortie de l'algorithme permettent donc d'obtenir directement la classification recherchée.

4.4.4 Autres algorithmes

D'autres approches ont été proposées pour estimer les paramètres d'un champ de Markov caché à partir d'observations non étiquetées. Nous décrivons plus particulièrement l'algorithme de Gradient Stochastique [135] et la procédure ICE [99] pour leur utilisation pratique. Citons encore les algorithme EM gibbsien [30], ainsi que l'algorithme MCEM [129] et ses généralisations proposées par [103].

Gradient Stochastique La procédure de gradient stochastique [135] est un algorithme d'estimation des paramètres $\boldsymbol{\psi}$ maximisant la log-vraisemblance $l(\boldsymbol{\psi}; \mathbf{x}) = \log P(\mathbf{x}|\boldsymbol{\psi})$ des données observées. Notons ∇_ψ le gradient par rapport à $\boldsymbol{\psi}$. On a :

$$\nabla_\psi l(\boldsymbol{\psi}, \mathbf{x}) = \nabla_\psi \log P(\mathbf{x}|\boldsymbol{\psi}) = \frac{\nabla_\psi P(\mathbf{x}|\boldsymbol{\psi})}{P(\mathbf{x}|\boldsymbol{\psi})} = \frac{\sum_{\mathbf{z}} \nabla_\psi P_G(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi})}{P(\mathbf{x}|\boldsymbol{\psi})}$$

Or, en utilisant la Proposition 18 de l'Annexe 3 et puisque $(\mathbf{X}, \mathbf{Z})$ est un champ de Markov (Proposition 12, section 4),

$$\nabla_\psi P_G(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}) = P_G(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi})(-\nabla_\psi H(\mathbf{x}, \mathbf{z}; \boldsymbol{\psi}) + \langle H(\mathbf{X}, \mathbf{Z}; \boldsymbol{\psi})\rangle)$$

Le gradient de la log-vraisemblance des données observées s'écrit donc :

$$\begin{aligned} \nabla_\psi l(\boldsymbol{\psi}, \mathbf{x}) &= \sum_{\mathbf{z}} P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi})(-\nabla_\psi H(\mathbf{x}, \mathbf{z}; \boldsymbol{\psi}) + \langle H(\mathbf{X}, \mathbf{Z}; \boldsymbol{\psi})\rangle) \\ &= -\langle \nabla_\psi H(\mathbf{x}, \mathbf{Z}; \boldsymbol{\psi})|\mathbf{x}\rangle + \langle \nabla_\psi H(\mathbf{X}, \mathbf{Z}; \boldsymbol{\psi})\rangle \end{aligned} \tag{II.36}$$

où $\langle .|\mathbf{x}\rangle$ désigne l'espérance conditionnellement aux observations $\mathbf{x}$.

La procédure de gradient stochastique [135] est une procédure itérative, partant d'une valeur donnée $\boldsymbol{\psi}^{(0)}$ des paramètres. A l'itération $(q+1)$, on ajuste les paramètres en se déplaçant dans une direction de gradient "approximative", la direction de gradient (II.36) ne pouvant être calculée de façon exacte. Pour approcher $\langle \nabla_\psi H(\mathbf{x}, \mathbf{Z}; \boldsymbol{\psi})|\mathbf{x}\rangle$, on simule une réalisation $\hat{\mathbf{z}}^{(q)}$ de la distribution à posteriori $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi}^{(q)})$. D'autre part, pour

approcher $\langle H(\mathbf{X}, \mathbf{Z}; \boldsymbol{\psi})\rangle$, on simule une réalisation $(\overline{\mathbf{z}}^{(q)}, \overline{\mathbf{x}}^{(q)})$ de la loi jointe $P_G(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}^{(q)})$. On met ensuite à jour les paramètres par un déplacement dans la direction du gradient ainsi approché :

$$\boldsymbol{\psi}^{(q+1)} = \boldsymbol{\psi}^{(q)} + \tau^{(q+1)} \left(-\nabla_\psi H(\mathbf{x}, \hat{\mathbf{z}}^{(q)}; \boldsymbol{\psi}^{(q)}) + \nabla_\psi H(\overline{\mathbf{x}}^{(q)}, \overline{\mathbf{z}}^{(q)}; \boldsymbol{\psi}^{(q)})\right)$$

où $\tau^{(q+1)}$ est le pas à l'itération $(q+1)$.

Procédure ICE La procédure ICE (*Iterated Conditional Expectation*) [99] est un algorithme général d'estimation en présence de données cachées. L'hypothèse de base est que l'on dispose d'une technique estimer les paramètres $\boldsymbol{\psi}(\mathbf{x}, \mathbf{z})$ sur les données complétées. Nous noterons $\hat{\boldsymbol{\psi}}(\mathbf{x}, \mathbf{z})$ un tel estimateur. Il est important de remarquer que, contrairement à l'algorithme EM, $\hat{\boldsymbol{\psi}}(\mathbf{x}, \mathbf{z})$ n'est pas nécessairement l'estimateur de maximum de vraisemblance $\boldsymbol{\psi}^{MVC}(\mathbf{x}, \mathbf{z})$ (défini en section 3.4.2). La configuration $\mathbf{z}$ n'étant pas disponible, le principe de la procédure consiste à approcher l'estimateur $\hat{\boldsymbol{\psi}}(\mathbf{x}, \mathbf{z})$ à partir des seules observations $\mathbf{x}$. Notons $(\psi_1, \cdots, \psi_R)$ l'ensemble des paramètres $\boldsymbol{\psi}$. Partant d'une initialisation $\boldsymbol{\psi}^{(0)}$ des paramètres, l'itération $(q+1)$ met à jour l'estimateur $\boldsymbol{\psi}^{(q+1)}$. Pour cela, on pose :

$$\psi_r^{(q+1)} = \langle \hat{\psi}_r(\mathbf{x}, \mathbf{Z})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle$$

pour tous les $r \in [\![1, R]\!]$ pour lesquels cette espérance est calculable. Pour les autres, on approche cette quantité en simulant M réalisations $(\mathbf{z}^{[1]}, \cdots, \mathbf{z}^{[M]})$ de $\mathbf{Z}$ selon la loi *a posteriori* $P_G(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi}^{(q)})$ puis en faisant la moyenne empirique des estimateurs de maximum de vraisemblance sur les données ainsi complétées :

$$\psi_r^{(q+1)} = \frac{1}{M} \sum_{m=1}^{M} \hat{\psi}_r(\mathbf{x}, \mathbf{z}^{[m]}).$$

Remarque. *Lorsque la loi jointe $P(\mathbf{x}, \mathbf{z})$ appartient à la famille exponentielle (par exemple gaussienne), on peut montrer que l'algorithme EM est un cas particulier de la procédure ICE [43].*

5 Sélection de modèle

Nous avons présenté un ensemble de modèles pour la classification de données. Se pose alors le problème de savoir quel modèle choisir pour modéliser et classer au mieux un jeu de données spécifique.

5.1 Cadre supervisé

En classification supervisée, on dispose d'un certain nombre de données étiquetées (la base d'apprentissage). L'approche la plus simple est alors de choisir le modèle par validation croisée : le modèle retenu est celui fournissant la meilleure classification sur les données d'apprentissage. Il est également possible de sélectionner ce "meilleur" modèle sur un critère probabiliste prenant en compte à la fois l'objectif de classification et celui

de la modélisation probabiliste, comme le *Bayesian Entropy Criterion* [20]. Enfin on peut choisir le modèle le plus adapté aux données sur l'unique critère de modélisation probabiliste et utiliser les techniques proposées dans un cadre non supervisé.

5.2 Cadre non supervisé

Le "meilleur" modèle choisi devra présenter un bon compromis entre complexité et adéquation aux données. De nombreux critères ont été proposés pour choisir entre différents modèles dans un cadre non-supervisé. Le *Bayesian Information Criterion* (BIC) [114] est certainement le plus répandu. Le principe du critère BIC est de sélectionner parmi l'ensemble des modèles $\mathcal{M}$ étudiés, celui maximisant la quantité :

$$BIC_{\mathcal{M}} = 2 \log L(\boldsymbol{\psi}_{\mathcal{M}}^{MV}; \mathbf{x}) - \nu_{\mathcal{M}} \log n \tag{II.37}$$

où $\nu_{\mathcal{M}}$ est le nombre de paramètres du modèle $\mathcal{M}$ et L est la valeur de la vraisemblance, calculée sur les paramètres de maximimum de vraisemblance $\boldsymbol{\psi}_{\mathcal{M}}^{MV}$ et les observations $\mathbf{x} = (x_1, \cdots, x_n)$. Ce critère BIC se décompose donc en deux termes : le terme de vraisemblance $2 \log L(\boldsymbol{\psi}_{\mathcal{M}}^{MV}; \mathbf{x})$ favorisant la sélection d'un modèle complexe et le terme de pénalité $\nu_{\mathcal{M}} \log n$, fonction croissante du nombre de paramètres, favorisant la sélection d'un modèle parcimonieux. Des expériences dans le cas de l'estimation du nombre de classes d'un modèle de mélange ont montré l'intérêt pratique de ce critère [53]. On observe cependant qu'il a tendance à surestimer le nombre de classes lorsque le vrai modèle ne fait pas partie de la famille considérée, notamment dans le cas de données réelles [13].

Le critère BIC est à préférer au *Akaike Information Criterion* (AIC) [4] :

$$AIC_{\mathcal{M}} = 2 \log L(\boldsymbol{\psi}_{\mathcal{M}}^{MV}; \mathbf{x}) - 2\nu_{\mathcal{M}}$$

qui ne pénalise pas suffisamment la complexité des modèles. Notons que le critère *Integrated Completed Likelihood* (ICL) [13] permet de tenir compte de la pertinence de la classification obtenue. La vraisemblance $L(\boldsymbol{\psi}_{\mathcal{M}}; \mathbf{x}) = P(\mathbf{x}|\boldsymbol{\psi}_{\mathcal{M}})$ du critère BIC y est remplacée par la vraisemblance complète $P(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}_{\mathcal{M}}^{MV})$:

$$ICL_{\mathcal{M}} = 2 \log P(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}_{\mathcal{M}}^{MV}) - \nu_{\mathcal{M}} \log n$$

Lorsque le champ $\mathbf{z}$ est inconnu, il est remplacé par son estimateur $\mathbf{z}^{map}$ du Maximum A Posteriori (voir section 2.2.2) :

$$\begin{aligned} ICL_{\mathcal{M}} &\approx 2 \log P(\mathbf{x}, \mathbf{z}^{map}|\boldsymbol{\psi}_{\mathcal{M}}^{MV}) - \nu_{\mathcal{M}} \log n & \text{(II.38)} \\ &= BIC_{\mathcal{M}} + 2 \log P(\mathbf{z}^{map}|\mathbf{x}, \boldsymbol{\psi}_{\mathcal{M}}^{MV}) & \text{(II.39)} \end{aligned}$$

L'espérance conditionnellement aux observations $\mathbf{x}$ du critère ICL, nous donne le critère EICL :

$$\begin{aligned} EICL_{\mathcal{M}} &= 2\langle \log P(\mathbf{x}, \mathbf{Z}|\boldsymbol{\psi}_{\mathcal{M}}^{MV})|\mathbf{x}\rangle - \nu_{\mathcal{M}} \log n \\ &= 2 \log P(\mathbf{x}|\boldsymbol{\psi}_{\mathcal{M}}^{MV}) + \langle \log P(\mathbf{Z}|\mathbf{x}, \boldsymbol{\psi}_{\mathcal{M}}^{MV})|\mathbf{x}\rangle - \nu_{\mathcal{M}} \log n \\ &= BIC_{\mathcal{M}} - 2S(P_G(.|\mathbf{x}, \boldsymbol{\psi}_{\mathcal{M}}^{MV})) \end{aligned}$$

où $S(P) = -\sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z})$ désigne l'entropie associée à la distribution P. Cette entropie $S(P)$ est d'autant plus grande que la distribution P est proche de la loi uniforme. La pénalité $2S(P(.|\mathbf{x}, \boldsymbol{\psi}_{\mathcal{M}}^{MV}))$ favorise donc les classes bien séparées.

Remarque. *Pour tout modèle $\mathcal{M}$, les critère BIC, ICL et EICL sont liés par les inégalités :*

$$BIC_{\mathcal{M}} > ICL_{\mathcal{M}} > EICL_{\mathcal{M}}$$

Dans les applications des chapitres VI, section 2.2 et VIII, section 2.3.2, nous utiliserons le critère BIC qui est, de manière générale, souvent préféré aux critère ICL, EICL et AIC. En pratique, l'estimateur de maximum de vraisemblance $\boldsymbol{\psi}_{\mathcal{M}}^{MV}$ est souvent inconnu. On le remplace alors dans l'équation (II.37) par son estimation (en sortie des algorithmes EM, NREM, Gradient Stochastique, ICE,...).

Cas du mélange indépendant Lorsque le modèle $\mathcal{M}$ est celui d'un mélange indépendant de distribution définie en section 3.1, le critère BIC est donné par :

$$BIC_{\mathcal{M}} = 2\sum_{i\in\mathcal{I}} \log\left(\sum_{k\in\mathcal{K}} \pi_k f(x_i|\theta_k)\right) - \nu_{\mathcal{M}} \log n$$

Cas d'une distribution de Gibbs Lorsque le modèle $\mathcal{M}$ est celui d'un champ de Markov caché de distribution définie en section 4.1, le critère BIC ne peut être calculé sans approximation. Nous décrivons dans les paragraphes suivants deux approximations du critère BIC définis dans [51], l'une utilisant l'approximation en champ moyen de la distribution de Gibbs P_G, l'autre l'approximation en champ moyen de la fonction de partition W.

5.2.1 Critère BIC par approximation de la distribution de Gibbs

Soit $\mathcal{M}$ le modèle de champ de Markov caché à bruit indépendant, de distribution définie en section 4.1. En utilisant l'approximation en champ moyen $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}$ de la distribution $P_G(\mathbf{z}|\mathbf{x})$ (voir section 4.2), la vraisemblance des paramètres du modèle est donnée par :

$$L(\boldsymbol{\psi}_{\mathcal{M}};\mathbf{x}) \approx P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x}|\boldsymbol{\Psi}_{\mathcal{M}}) = \prod_{i\in S}\sum_{k\in\mathcal{K}} f(x_i|\theta_k) P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(Z_i = k|\boldsymbol{\phi})$$

et le critère BIC est alors approximé par le critère $\text{BIC}_{\mathcal{M}}^{p}$:

$$BIC_{\mathcal{M}}^{p} = 2\sum_{i\in S} \log\left(\sum_{k\in\mathcal{K}} f(x_i|\theta_k) P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(Z_i = k|\boldsymbol{\phi})\right) - \nu_{\mathcal{M}} \log n \qquad \text{(II.40)}$$

5.2.2 Critère BIC par approximation de la fonction de partition

Soit $\mathcal{M}$ le modèle de champ de Markov caché de distribution définie en section 4.1. Remarquons que :

$$\begin{aligned} P(\mathbf{x}|\boldsymbol{\psi}) &= \frac{P_G(\mathbf{x},\mathbf{z}|\boldsymbol{\psi})}{P_G(\mathbf{z}|\mathbf{x},\boldsymbol{\psi})} \\ &= \frac{f(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})P_G(\mathbf{z}|\boldsymbol{\phi})}{P_G(\mathbf{z}|\mathbf{x},\boldsymbol{\psi})} \\ &= \underbrace{\frac{f(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})\exp(-H(\mathbf{z}|\boldsymbol{\phi}))}{\exp(-H(\mathbf{z}|\mathbf{x},\boldsymbol{\psi}))}}_{=1}\frac{W(\mathbf{x},\boldsymbol{\psi})}{W(\boldsymbol{\phi})} \\ &= \frac{W(\mathbf{x},\boldsymbol{\psi})}{W(\boldsymbol{\phi})} \end{aligned} \tag{II.41}$$

et donc :

$$BIC_{\mathcal{M}} = 2\log W(\mathbf{x},\boldsymbol{\psi}^{MV}) - 2\log W(\boldsymbol{\phi}^{MV}) - \nu_{\mathcal{M}}\log n \tag{II.42}$$

Pour approximer $BIC_{\mathcal{M}}$, une autre alternative à $BIC^p_{\mathcal{M}}$ est donc d'approximer les constantes de partitions $W(\mathbf{x},\boldsymbol{\psi}^{MV})$ et $W(\boldsymbol{\phi}^{MV})$. Pour clarifier les notations, nous omettons provisoirement les paramètres $\boldsymbol{\psi}^{MV}$ et $\boldsymbol{\phi}^{MV}$. Notons que :

$$\begin{aligned} W(\mathbf{x}) &= \sum_{\mathbf{z}}\exp(-H(\mathbf{z}|\mathbf{x})) \\ &= W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x})\sum_{\mathbf{z}}\underbrace{W^{-1}_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x})\exp(-H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z}|\mathbf{x}))}_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(.|\mathbf{x})}\exp(-H(\mathbf{z}|\mathbf{x}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z}|\mathbf{x})) \\ &= W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x})\left\langle\exp(-H(\mathbf{Z}|\mathbf{x}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z}|\mathbf{x}))\right\rangle_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(.|\mathbf{x})} \end{aligned}$$

où on a noté $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(.|\mathbf{x})$ l'approximation en champ moyen de $P_G(.|\mathbf{x})$, $H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(.|\mathbf{x})$ son énergie et $W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x})$ sa constante de normalisation. Dans une approche en champ moyen, on suppose que la variation $H(.|\mathbf{x}) - H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(.|\mathbf{x})$ est petite. En utilisant le développement à l'ordre 1 de l'exponentielle au voisinage de 0 ($e^x = 1 + x + o(x)$), $\langle\exp(-H(\mathbf{Z}|\mathbf{x}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z}|\mathbf{x}))\rangle$ peut être approximé par $\exp\langle -H(\mathbf{Z}|\mathbf{x}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z}|\mathbf{x})\rangle$. On en déduit que :

$$W(\mathbf{x}) \approx W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x})\exp\left\langle -H(\mathbf{Z}|\mathbf{x},\boldsymbol{\psi}^{MV}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z}|\mathbf{x})\right\rangle_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(.|\mathbf{x})} \tag{II.43}$$

Or, sous l'hypothèse d'indépendance conditionnelle,

$$H(\mathbf{z}|\mathbf{x}) = H(\mathbf{z}) + \sum_{i\in\mathcal{I}}\log f(x_i|\theta_{z_i}) \qquad H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z}|\mathbf{x}) = H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z}) + \sum_{i\in\mathcal{I}}\log f(x_i)$$

L'équation (II.43) s'écrit donc :

$$W(\mathbf{x}) \approx W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x})\exp\langle -H(\mathbf{Z}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z})\rangle_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(.|\mathbf{x})} \tag{II.44}$$

Par une démarche analogue, on peut voir que :

$$\begin{aligned} W &= \sum_{\mathbf{z}} \exp(-H(\mathbf{z})) \\ &= W_{\tilde{\mathbf{z}}^{\mathbf{x}}} \sum_{\mathbf{z}} \underbrace{W_{\tilde{\mathbf{z}}^{\mathbf{x}}}^{-1} \exp(-H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z}))}_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}} \exp(-H(\mathbf{z}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z})) \\ &= W_{\tilde{\mathbf{z}}^{\mathbf{x}}} \left\langle \exp(-H(\mathbf{Z}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z})) \right\rangle_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}} \end{aligned}$$

où on a noté $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z})$ l'approximation en champ moyen de $P_G(\mathbf{z})$ en fixant les voisins au champ $\tilde{\mathbf{z}}^{\mathbf{x}}$, $H_{\tilde{\mathbf{z}}^{\mathbf{x}}}$ son énergie et $W_{\tilde{\mathbf{z}}^{\mathbf{x}}}$ sa constante de normalisation.

De même, en omettant provisoirement les paramètres $\boldsymbol{\phi}^{MV}$,

$$\begin{aligned} W &= \sum_{\mathbf{z}} \exp(-H(\mathbf{z})) \\ &= W_{\tilde{\mathbf{z}}^{\mathbf{x}}} \sum_{\mathbf{z}} \underbrace{W_{\tilde{\mathbf{z}}^{\mathbf{x}}}^{-1} \exp(-H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z}))}_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}} \exp(-H(\mathbf{z}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z})) \\ &= W_{\tilde{\mathbf{z}}^{\mathbf{x}}} \left\langle \exp(-H(\mathbf{Z}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z})) \right\rangle_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}} \end{aligned}$$

En utilisant une démarche similaire à l'approximation de $W(\mathbf{x}, \boldsymbol{\psi})$, on trouve :

$$W \approx W_{\tilde{\mathbf{z}}^{\mathbf{x}}} \exp \left\langle -H(\mathbf{Z}) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z}) \right\rangle_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}} \tag{II.45}$$

En utilisant les approximations (II.43) et (II.45), le critère BIC (équation II.42) est donc approximé par $\text{BIC}_{\mathcal{M}}^{w}$:

$$\begin{aligned} BIC_{\mathcal{M}}^{w} &= 2 \log W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\boldsymbol{\phi}) + 2 \left\langle -H(\mathbf{Z}, \phi) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z}, \boldsymbol{\phi}) \right\rangle_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}} \\ &- 2 \log W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x}, \boldsymbol{\psi}) - 2 \left\langle -H(\mathbf{Z}, \phi) + H_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z}, \boldsymbol{\phi}) \right\rangle_{P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(.|\mathbf{x}, \boldsymbol{\psi})} \\ &- \nu_{\mathcal{M}} \log n \end{aligned}$$

Lien entre les deux approximations. Puisque :

$$P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x}) = \frac{W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x})}{W_{\tilde{\mathbf{z}}^{\mathbf{x}}}}$$

l'approximations $BIC_{\mathcal{M}}^{p}$ (II.40) s'écrit encore :

$$BIC_{\mathcal{M}}^{p} = 2 \log W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{x}, \boldsymbol{\psi}) - 2 \log W_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\boldsymbol{\phi}) - \nu_{\mathcal{M}} \log n$$

ce qui correspond au développement à l'ordre 1 de $BIC_{\mathcal{M}}^{w}$. Les auteurs de [51] remarquent qu'en théorie comme en pratique, l'approximation $BIC_{\mathcal{M}}^{w}$ est plus fine que l'approximation $BIC_{\mathcal{M}}^{p}$.

Partie B

Modèles de bruit non standards

Problématique

On se place toujours dans un contexte de classification. On dispose d'observations $\mathbf{x} = (x_i)_{i\in\mathcal{I}}$ réelles, D-dimensionnelles, associées à des individus $i \in \mathcal{I}$ que l'on souhaite grouper en K classes. Ces individus sont supposés en interaction les uns avec les autres. Le problème est d'associer à chacun des individus $i \in \mathcal{I}$, une classe $z_i \in \mathcal{K} = [\![1, K]\!]$. Les observations $\mathbf{x} = (x_i)_{i\in\mathcal{I}}$ et classes $\mathbf{z} = (z_i)_{i\in\mathcal{I}}$ peuvent être vues comme des réalisations de variables aléatoires $\mathbf{X} = (X_i)_{i\in\mathcal{I}}$ à valeur dans $\mathbb{R}^D$ et classes $\mathbf{Z} = (Z_i)_{i\in\mathcal{I}}$ discrètes, à valeur dans $\mathcal{K}$.

L'hypothèse classiquement utilisée concernant la loi des classes $P(\mathbf{x}|\mathbf{z})$ est celle de bruit indépendant :

$$P(\mathbf{x}|\mathbf{z}) = \prod_{i\in\mathcal{I}} P(x_i|z_i)$$

Les distributions $P(.|z_i)$, $z_i \in \mathcal{K}$, sont en général des distributions standards, typiquement des gaussiennes $\mathcal{N}(.|\theta_{z_i})$. Dans le cas unidimensionnel ($\forall i \in \mathcal{I},\ \mathbf{x}_i \in \mathbb{R}$), $P(.|z_i) \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i})$ et la distribution conditionnelle s'écrit :

$$P(\mathbf{X} = \mathbf{x}|\mathbf{Z} = \mathbf{z}) \propto \exp(-\frac{1}{2}\sum_{i\in\mathcal{I}} \sigma^{-2}_{z_i}(x_i - \mu_{z_i})^2)$$

Dans le cas D-dimensionnel ($\forall i \in \mathcal{I},\ \mathbf{x}_i \in \mathbb{R}^D$), $P(.|z_i) \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$, la distribution conditionnelle s'écrit :

$$P(\mathbf{X} = \mathbf{x}|\mathbf{Z} = \mathbf{z}) \propto \exp(-\frac{1}{2}\sum_{i\in\mathcal{I}} (x_i - \mu_{z_i})\Sigma^{-1}_{z_i}(x_i - \mu_{z_i})')$$

Concernant ces hypothèses standards, deux limitations importantes peuvent être soulevées :

1. Les données modernes sont en dimension de plus en plus importante. Or le modèle gaussien, sans autre hypothèse, nécessite l'estimation d'un très grand nombre de paramètres (essentiellement pour l'estimation de la matrice de covariance). D'un autre côté, la plupart des données de grande dimension vivent en réalité dans des espaces intrinsèques de dimension beaucoup plus faible. C'est le phénomène de l'*espace vide*.
2. Quelle que soit la dimension D, la loi gaussienne $\mathcal{N}(\mu, \Sigma)$ est unimodale : son unique mode correspond au vecteur moyenne D-dimensionnel μ. Or ces deux hypothèses de bruit indépendant et d'uni-modalité sont trop restrictives pour modéliser certains problèmes, notamment celui de la segmentation d'images en différentes textures.

Les prochains chapitres s'intéressent alors à surmonter les 2 limitations des hypothèses classiques de bruit indépendant gaussien. Le chapitre III traite de la classification d'observations de grande dimension. Nous développons pour cela le modèle proposé dans [23] et l'adaptons au cadre d'une classification markovienne. Le chapitre IV présente différentes méthodes développées pour relâcher l'hypothèse de bruit indépendant unimodal et notamment les modèles de Markov triplets introduits de manière générale dans [6] et dont certains cas particuliers ont été étudiés et illustrés sur des applications, notamment dans [6] et [7]. L'étude de ces modèles nous a cependant permis de mettre en évidence des problèmes possibles d'identifiabilité des paramètres pour des modèles de bruit trop généraux. Enfin, nous proposons dans le chapitre V un nouveau modèle de champ de Markov triplet adapté à la classification supervisée de données, sous lequel le modèle de bruit n'est ni indépendant, ni unimodal dans le cas général. Nous verrons néanmoins que les étapes d'apprentissage et de classification peuvent être effectuées par les méthodes bayésiennes classiques. Enfin, des simulations de notre modèle de Markov triplet ainsi qu'une application à des données réelles de grande dimension (issues d'images de textures) sont présentées.

Chapitre III

Classification d'observations de grande dimension

1 Traiter des données de grande dimension

1.1 Le fléau de la dimension

La spécificité de beaucoup de données modernes est certainement leur grande dimension. Dans de nombreux domaines, là où on ne faisait qu'une poignée de mesures il y a quelques années, nous en faisons parfois aujourd'hui des dizaines, des centaines, voire des milliers. Cette évolution a été rendue possible principalement grâce aux progrès des capacités de stockage. Citons par exemple le cas de l'imagerie hyperspectrale. Une image hyperspectrale est composée d'une serie d'images de la même scène, mais prises dans plusieurs dizaines, voire centaines, de longueurs d'ondes - qui correspondent à autant de "couleurs". Un pixel n'est alors plus décrit par une couleur mais par un vecteur de couleurs, pouvant être de grande dimension. Les images de Mars fournies par le satellite OMEGA par exemple correspondent à 256 longueurs d'onde.
Pour des données de si grande dimension, se pose alors le problème de la qualité des estimateurs obtenus. Dans le cas d'un modèle gaussien $\mathcal{N}(\mu, \Sigma)$ par exemple, la moyenne μ est un vecteur de dimension D et la matrice de covariance Σ est une matrice de dimension $D \times D$. Leur estimation requiert alors celle de $D + \frac{1}{2}D(D+1)$ paramètres.
Sans autre hypothèse, c'est la matrice de covariance Σ qui contient la plus grande part des paramètres à estimer. Si le nombre d'observations n'est pas suffisant, cette matrice sera mal conditionnée (son calcul numérique sera difficile), voire singulière (son inversion sera numériquement impossible). De manière générale, on parle du *fléau de la dimension* (*curse of dimensionality*) [84]. Pour pallier les problèmes liés au mauvais conditionnement ou à la singularité des estimations de matrices de covariance, plusieurs alternatives ont été proposées, notamment de régulariser la matrice de covariance, c'est-à-dire à remplacer son estimateur $\hat{\Sigma}$ par $\hat{\Sigma} + c^2\mathbb{I}_D$. Dans le cadre de la classification de données, c'est le type de méthodes employées par l'Analyse Discriminante Régularisée (*Regularized Discriminant Analysis*) [54] et l'Analyse Discriminante Pénalisée (*Penalized Discriminant Analysis*) [67]. Cependant, de telles approches introduisent un biais dans l'estimation de la matrice de covariance, biais qui influe sur la règle de décision de la classification. Une seconde solution est l'utilisation de modèles parcimonieux, c'est-à-dire de modèles qui requièrent l'estimation d'un nombre "raisonnable" de paramètres. A partir de la décomposition spec-

trale de Σ dans le cas gaussien, les auteurs de [5] mettent en évidence 14 modèles particuliers allant du modèle le plus parcimonieux (le modèle sphérique) au modèle le moins parcimonieux (le modèle gaussien classique à matrice de covariance pleine). Dans le cas de la classification, cette paramétrisation donne naissance à l'Analyse Discriminante par décomposition selon les valeurs propres (*Eigenvalue Decomposition Discriminant Analysis*) [8]. Les modèles parcimonieux ont l'avantage de n'être décrits que par un nombre raisonnable de paramètres. Néanmoins, ces modèles n'ont pas été spécifiquement créés pour les données de grande dimension et ne prennent notamment pas en compte le phénomène de l'espace vide caractéristique de la grande dimension.

1.2 Le phénomène de l'espace vide

La plupart des données de grande dimension vivent en réalité dans des espaces intrinsèques de dimension beaucoup plus faible. C'est le phénomène de l'espace vide : l'espace de dimension D est presque vide puisque la plupart des observations appartiennent à un espace de dimension inférieure. Une des solutions pouvant être mis en œuvre pour faire face à ce phénomène est alors de réduire la dimension des données, et ce sans perte d'information. L'enjeu est alors d'identifier les axes porteurs d'information redondante. Cette réduction de dimension peut être effectuée en pré-traitement, par analyse en composante principale ou sélection de variables par exemple. Une limitation de ces méthodes est qu'elles ne prennent pas en compte l'objectif de classification lors de la réduction de dimension. Une alternative est l'analyse factorielle discriminante combinant réduction de dimension et discrimination. Le principe est projeter les données sur les axes maximisant le rapport de la variance inter-classe et de la variance intra-classe. Dans la même idée, citons la méthode de poursuite de projection visant à rechercher un sous-espace dans lequel un indice de projection est maximisé. L'intérêt de cette méthode est que l'indice de projection peut être adapté selon le traitement visé. En particulier, pour le problème de la classification d'images hyperspectrale, les auteurs de [76] et de [104] utilisent la poursuite de projection pour déterminer le sous-espace maximisant la distance de Bhattacharyya entre les classes, distance qui est liée à une borne supérieure de la probabilité d'erreur de classification. Néanmoins, de manière générale en classification, ces méthodes de réduction de dimension se font souvent au prix d'une perte d'information car, certes, les D variables ne sont peut être pas toutes nécessaires pour décrire les observations, mais l'ensemble des variables est souvent utile (voire nécessaire) pour discriminer les classes les unes par rapport aux autres. De plus, on peut supposer que les données de classes différentes vivent dans des sous-espaces différents, hypothèse qu'une approche globale de réduction de dimension ne peut prendre en compte.

2 Modèle gaussien pour la classification en grande dimension

2.1 Modèle gaussien de grande dimension

L'idée est d'utiliser le fait que les données de grande dimension vivent dans des sous-espaces dont les dimensions intrinsèques sont faibles. Pour cela [23] propose une reparamétrisation du modèle gaussien prenant en compte le fait que les données de chaque classe vivent dans des sous-espaces différents dont les dimensions intrinsèques peuvent varier. Soit donc Σ_k la matrice de covariance de la classe k. Considérons la décomposition

spectrale :

$$\Sigma_k = Q_k \Delta_k Q_k'$$

où Q_k est la matrice orthogonale de taille $D \times D$ des vecteurs propres de Σ_k, Q_k' sa transposée et Δ_k est la matrice diagonale des valeurs propres. [23] propose de modéliser le fait que les données de chacune des classes vivent dans des sous-espaces de dimensions inférieures en écrivant Δ_k sous la forme :

$$\Delta_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{kD_k} \end{matrix}} & & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} D_k \\ & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ 0 & & b_k \end{matrix}} & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} (D - D_k) \end{pmatrix}$$

où, pour tout $d = 1, ..., D_k$, $a_{kd} > b_k$, et $D_k < D$. Notons que cela revient à supposer que les $D - D_k$ plus petites valeurs propres sont égales. Il est toujours possible de faire cette hypothèse quitte à prendre $D_k = D - 1$. Néanmoins en pratique, et c'est tout l'intérêt de cette modélisation, on a $D_k << D$.

Calcul de $\mathbf{\Sigma_k}$ et $\mathbf{\Sigma_k^{-1}}$ Notons Q_k^a la matrice $D \times D$ composée des D_k premières colones de Q_k complétée par des zéros et $Q_k^b = Q_k - Q_k^a$ la matrice composée des $D - D_k$ dernières colones de Q_k complétée par des zéros. Notons que :

$$Q_k^a \Delta_k (Q_k^b)' = Q_k^b \Delta_k (Q_k^a)' = 0_D$$

où 0_D désigne la matrice nulle. De plus, en notant que les lignes de Q_k forment un système orthonormé ($Q_k Q_k' = \mathbb{I}_D$, la matrice identité en dimension D) et que les $D - D_k$ premières colonnes de Q_k^b sont nulles, on a :

$$Q_k^b \Delta_k (Q_k^b)' = b_k Q_k^b (Q_k^b)' = b_k (\mathbb{I}_D - Q_k^a (Q_k^a)').$$

On en déduit :

$$\begin{aligned} \Sigma_k &= Q_k^a \Delta_k (Q_k^a)' + b_k (\mathbb{I}_D - Q_k^a (Q_k^a)') \\ \Sigma_k^{-1} &= Q_k^a \Delta_k^{-1} (Q_k^a)' + \frac{1}{b_k} (\mathbb{I}_D - Q_k^a (Q_k^a)') \end{aligned} \quad \text{(III.1)}$$

Notons que les calculs de Σ_k et Σ_k^{-1} ne mettent en jeu que les D_k vecteurs propres de Σ_k associées aux D_k plus grandes valeurs propres. La loi $\mathcal{N}(\mu_k, \Sigma_k)$ requiert alors l'estimation de $\nu = D + 2 + D_k + D_k(D - \frac{1}{2}(D_k + 1))$ paramètres. A titre d'exemple, lorsque $D = 100$ et $D_k = 10$, $\nu = 1057$ alors que pour le modèle à matrice de covariance pleine, $\nu = D + \frac{1}{2}D(D + 1) = 5150$.

2.2 Règle de classification

Les règles de classification (II.17) pour mélange indépendant ou (II.35) pour champ de Markov caché avec approximation de type champ moyen, nécessitent le calcul, en chaque site i et pour chaque classe k, de :

$$\log f(x_i | \mu_k, \Sigma_k) = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k) \Sigma_k^{-1} (x_i - \mu_k)'$$

L'expression de Σ_k^{-1} est donnée par l'équation (III.1). Le déterminant $|\Sigma_k|$ est égal au déterminant de Δ_k, c'est-à-dire :

$$\log|\Sigma_k| = \sum_{d=1}^{D_k} \log a_{kd} + (D - D_k)\log b_k$$

3 Estimation par algorithme de type EM

Que ce soit lors de l'application de l'algorithme EM pour mélange indépendant ou de l'algorithme NREM pour champ de Markov caché à bruit indépendant, l'estimation des paramètres μ_k et Σ_k s'effectue au cours de l'étape (M), qui est la même pour ces deux algorithmes (quitte à remplacer t_{ik} par son approximation en champ moyen $\tilde{t}_{ik}$). Les moyennes μ_k sont mises à jour par l'équation (II.22). La mise à jour de Σ_k, quant à elle, diffère du cas classique donné par l'équation (II.23). L'obtention de Σ_k nécessite de connaître :

- la dimension intrinsèque D_k
- les valeurs propres a_{kd}, $d \in [\![1, D_k]\!]$
- la valeur propre b
- les vecteurs propres q_{kd} associés aux valeurs propres a_{kd}, $d \in [\![1, D_k]\!]$.

C. Bouveyron montre dans [23] que les estimateurs de ces paramètres au cours de l'algorithme EM s'appuient sur l'équation (II.23) de mise à jour de la matrice de covariance dans le cas classique, que nous noterons Σ^{cla} :

$$\Sigma^{cla} = \frac{\sum_{i \in \mathcal{I}} t_{ik}(x_i - \mu_k)(x_i - \mu_k)'}{\sum_{i \in \mathcal{I}} t_{ik}}$$

Estimation de $\mathbf{D}_k$ Dans le cas supervisé, D_k peut être obtenu par validation croisée sur le jeu d'apprentissage. Dans le cas non supervisé, [23] propose d'utiliser le scree-test de Catell [25] consistant à déterminer un coude dans l'éboulis des valeurs propres de la matrice de covariance Σ^{cla}.

Estimation des valeurs propres $\mathbf{a}_{kd}$ et des vecteurs propres associés L'estimateur $\hat{a}_{kd}$ de la valeur propres a_{kd}, $d \in [\![1, D_k]\!]$ correspond à la dième plus grande valeur propre de Σ^{cla}. L'estimateur du vecteur propre q_{kd} correspondant est le vecteur propre associé à $\hat{a}_{kd}$ dans la décomposition de Σ^{cla}.

Estimation de la valeur propre $\mathbf{b}_k$ L'estimateur $\hat{b}_k$ de la valeur propre b_k est la moyenne des $(D - D_k)$ plus petites valeurs propres de Σ^{cla}.

L'intérêt d'une telle méthode est de permettre d'aller au delà du cas diagonal généralement supposé et donc de prendre en compte les dépendances entre les différentes dimensions. Nous donnerons un exemple d'utilisation de cette technique pour la classification de données en grande dimension (les descripteurs d'images de textures) au chapitre VI, section 2. Néanmoins, la distribution gaussienne, quelque soit sa dimension, reste unimodale et l'hypothèse de bruit indépendant gaussien ne peut modéliser les classes complexes,

multi-modales par exemple. De plus, l'hypothèse de bruit indépendant est trop restrictive pour certaines applications (la modélisation de textures par exemple). Le chapitre suivant présente les modèles proposés pour surmonter l'hypothèse de bruit indépendant et d'unimodalité, notamment les modèles de champs de Markov triplet de [6].

Chapitre IV

Variations autour des modèles de bruit standards

Ce chapitre présente différentes méthodes développées pour relâcher l'hypothèse de bruit indépendant unimodal et notamment les modèles de Markov triplets proposés par [6] et permettant de s'affranchir, dans un contexte markovien, à la fois de l'hypothèse de bruit indépendant, et de celle d'unimodalité des classes.

1 Relâcher l'hypothèse de bruit indépendant

De manière générale, les champs de Markov cachés que nous avons utilisés jusqu'à présent (et classiquement utilisés) font les deux hypothèses suivantes :

(A) $\mathbf{Z}$ est un champ de Markov : $P_G(\mathbf{z}) = W^{-1} \exp(-H(\mathbf{z}))$

(B) Le bruit est indépendant : $P(\mathbf{x}|\mathbf{z}) = \prod_{i \in \mathcal{I}} P(x_i|z_i)$

Comme souligné dans [6], cette hypothèse (B) de bruit indépendant, se décompose en :

(B1) $P(x_i|\mathbf{z}) = P(x_i|z_i)$ pour tout $i \in \mathcal{I}$

(B2) $P(\mathbf{x}|\mathbf{z}) = \prod_{i \in \mathcal{I}} P(x_i|\mathbf{z})$

L'hypothèse (B1) signifie que, conditionnellement à la classe Z_i, l'observation X_i au site $i \in \mathcal{I}$ est indépendante des classes Z_j, $j \neq i$. L'hypothèse (B2) revient à dire que les observations $\mathbf{X}$ sont indépendantes conditionnellement aux classes $\mathbf{Z}$. Sous ces hypothèse (A) et (B), la distribution *a posteriori* $P(\mathbf{Z}|\mathbf{x})$ du champ $\mathbf{Z}$ est alors une distribution de Gibbs :

$$
\begin{aligned}
P(\mathbf{z}|\mathbf{x}) &= W(\mathbf{x})^{-1} \exp\left(-H(\mathbf{z}) + \sum_{i \in \mathcal{I}} \log P(x_i|z_i) \right) \\
\text{où } W(\mathbf{x}) &= \sum_{\mathbf{z}} \left(\exp(-H(\mathbf{z})) \prod_{i \in \mathcal{I}} P(x_i|z_i) \right)
\end{aligned}
$$

Cette markovianité du champ conditionnel $\mathbf{Z}|\mathbf{x}$ rend possible la simulation de $\mathbf{Z}$ selon cette loi *a posteriori*, et permet ainsi d'utiliser toutes les approximations bayésiennes pour le calcul du MAP, comme l'algorithme de recuit simulé [58] ou l'algorithme ICM [10] (voir chapitre II, section 4.3). De même, les algorithmes d'estimation comme l'EM gibbsien [30], EM de type champ moyen, MCEM [129] et ses généralisations [103], le Gradient

Stochastique [135] ou encore la procédure ICE [99] (voir chapitre II, section 4.4) utilisent tous la markovianité de la loi *a posteriori*.

Nous présentons dans cette partie certains modèles proposés dans le cadre des champs de Markov cachés pour relâcher cette hypothèse de bruit indépendant (section 1.1), puis le modèle plus général de champ de Markov couple (section 1.2) proposé par [6].

1.1 Pour les champs de Markov cachés

Nous supposons ici que le champ caché $\mathbf{Z}$ est markovien :

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

et présentons plusieurs alternatives pour relâcher l'hypothèse de bruit indépendant, c'est-à-dire l'hypothèse (B2) et/ou (B1).

Champs de Markov gaussiens. Dans le cadre de la modélisation de textures, une alternative pour relâcher (B2) est de supposer que les textures (c'est-à-dire les classes) sont des réalisations d'un champ de Markov gaussien [39]. Dans le cas 1-dimensionnel, la distribution des observations conditionnellement aux classes est alors de la forme :

$$\begin{aligned} P_G(\mathbf{x}|\mathbf{z}) &= W(\mathbf{z})^{-1} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{z}})\mathbf{Q}_{\mathbf{z}}(\mathbf{x}-\mu_{\mathbf{z}})'\right) \\ \text{où } W(\mathbf{z}) &= (2\pi)^{\frac{n}{2}} (\log|Q_{\mathbf{z}}|)^{-\frac{1}{2}} \end{aligned} \tag{IV.1}$$

où la matrice $\mathbf{Q}_{\mathbf{z}} = (Q_{z_i z_j})_{i,j\in\mathcal{I}}$, de dimension $n \times n$, est une matrice symétrique telle :

$$Q_{z_i z_j} = Q_{z_j z_i} \neq 0 \Leftrightarrow i \text{ et } j \text{ sont voisins.}$$

Pour que cette distribution soit définie, il faut de plus que $\mathbf{Q}_{\mathbf{z}}$ soit définie positive. En développant le terme $(\mathbf{x}-\mu_{\mathbf{z}})\mathbf{Q}_{\mathbf{z}}(\mathbf{x}-\mu_{\mathbf{z}})'$, la distribution (IV.1) s'écrit encore :

$$P_G(\mathbf{x}|\mathbf{z}) = W(\mathbf{z})^{-1} \exp\left(-\frac{1}{2}\sum_{i\sim j} Q_{z_i z_j}(x_i-\mu_{z_i})(x_j-\mu_{z_j}) - \frac{1}{2}\sum_{i\in\mathcal{I}} Q_{z_i z_i}(x_i-\mu_{z_i})^2\right)$$

Notons le terme additionnel croisé $\sum_{i\sim j} Q_{z_i z_j}(x_i-\mu_{z_i})(x_j-\mu_{z_j})$ en comparaison avec la distribution d'un bruit gaussien indépendant :

$$P(\mathbf{x}|\mathbf{z}) \propto \exp(-\frac{1}{2}\sum_{i\in\mathcal{I}} \sigma_{z_i}^{-2}(x_i-\mu_{z_i})^2) \tag{IV.2}$$

Plus précisément, le cas gaussien avec bruit indépendant défini par (IV.2) correspond au cas où $\mathbf{Q}_{\mathbf{z}}$ est une matrice diagonale avec, pour tout $i \in \mathcal{I}$, $Q_{z_i z_i} = \sigma_{z_i}^{-2}$. Sous la forme (IV.1), l'hypothèse (B2) est relâchée puisque les observations $\mathbf{X}$ ne sont plus indépendantes conditionnellement aux classes $\mathbf{Z}$. En revanche, l'hypothèse (B1) est toujours vérifiée puisque $P(x_i|\mathbf{z}) \sim \mathcal{N}(\mu_{z_i}, Q_{z_i z_i}^{-1})$ ne dépend que de z_i.
Sous l'hypothèse de markovianité de $\mathbf{Z}$, la loi du couple $(\mathbf{X}, \mathbf{Z})$ est alors :

$$P(\mathbf{x},\mathbf{z}) = W^{-1}W(\mathbf{z})^{-1} \exp(-H(\mathbf{z}) - \frac{1}{2}(\mathbf{x}-\mu_{\mathbf{z}})\mathbf{Q}_{\mathbf{z}}(\mathbf{x}-\mu_{\mathbf{z}})')$$

Bien que $\mathbf{Z}$ et $\mathbf{X}|\mathbf{z}$ soient markoviens, ni le couple $(\mathbf{X}, \mathbf{Z})$, ni $\mathbf{Z}$ conditionnellement à $\mathbf{X}$ ne sont nécessairement markoviens. En effet, $W(\mathbf{z})$, sauf cas particulier, ne peut être écrit sous la forme d'une distribution markovienne en $\mathbf{z}$. Ce modèle présente donc le désavantage de ne pouvoir utiliser les traitement bayésiens classiques pour l'estimation des paramètres (MCEM, EM Gibbsien...) ou le calcul du MAP (ICM, Recuit simulé...) car toutes ces procédures nécessitent la markovianité de $P(\mathbf{z}|\mathbf{x})$. Enfin, notons que (IV.1) est la distribution d'une loi gaussienne dans $\mathbb{R}^n$ de moyenne $\mu_{\mathbf{z}}$ et de matrice de covariance $\mathbf{Q}_{\mathbf{z}}^{-1}$. Si elle permet de relâcher l'hypothèse de bruit indépendant, elle ne permet pas de relâcher celle d'unimodalité.

Champs de Markov cachés à bruit corrélé. Pour relâcher (B2) dans le cas général, [6] propose de définir le modèle de la manière suivante. Soit $\mathbf{T} = (T_i)_{i\in\mathcal{I}}$ un champ markovien sur $\mathbb{R}^D$, d'énergie H', indépendant du champ $\mathbf{Z}$. Soit encore K fonctions dérivables et bijectives $f_1, \cdots, f_K$ de $\mathbb{R}^D \to \mathbb{R}^D$. Considérons le champ $\mathbf{X} = (X_i)_{i\in\mathcal{I}}$ tel que, pour tout $i \in \mathcal{I}$, $X_i = f_{Z_i}(T_i)$. La loi du couple $(\mathbf{X}, \mathbf{Z})$ est alors donnée par :

$$P(\mathbf{x}, \mathbf{z}) \propto \exp\left(-H(\mathbf{z}) - H'(f_{\mathbf{z}}^{-1}(\mathbf{x})) + \sum_{i\in\mathcal{I}} \log|\frac{\partial f_{z_i}^{-1}(x_i)}{\partial x_i}|\right) \tag{IV.3}$$

où $f_{\mathbf{z}}^{-1}(\mathbf{x}) = (f_{z_1}^{-1}(x_1), \cdots, f_{z_n}^{-1}(x_n))$. Le couple $(\mathbf{X}, \mathbf{Z})$ est donc un champ de Markov, si bien que les champ conditionnels $\mathbf{Z}|\mathbf{x}$ et $\mathbf{X}|\mathbf{z}$ sont également markoviens. Sous ces hypothèses, (B2) est relâchée puisque, d'après (IV.3), $P(\mathbf{x}|\mathbf{z}) \neq \prod_{i\in\mathcal{I}} P(x_i|\mathbf{z})$. Par contre, on a toujours $P(x_i|\mathbf{z}) = P(x_i|z_i)$ car $X_i = f_{Z_i}(T_i)$ ne dépend pas des Z_j pour $j \neq i$.

Pour relâcher à la fois (B1) et (B2), [6] propose de partitionner $\mathcal{I}$ en un ensemble de cliques $\{c_1, \cdots, c_Q\}$ de cardinal $a \geq 2$ ($c_1 \cup \cdots \cup c_Q = \mathcal{I}$ et pour tout $q \neq q'$, $c_q \cap c_{q'} = \varnothing$). Considérons alors le champ $\mathbf{X} = (X_i)_{i\in\mathcal{I}}$ tel que, pour tout $q \in [\![1, Q]\!]$, $X_{c_q} = f_{Z_{c_q}}(T_{c_q})$. La loi du couple $(\mathbf{X}, \mathbf{Z})$ est alors donnée par :

$$P(\mathbf{x}, \mathbf{z}) \propto \exp\left(-H(\mathbf{z}) - H'(f_{\mathbf{z}}^{-1}(\mathbf{x})) + \sum_{q=1}^{Q} \log|\nabla_{x_{c_q}} f_{z_{c_q}}^{-1}(x_{c_q})|\right) \tag{IV.4}$$

$(\mathbf{X}, \mathbf{Z})$ est donc un champ de Markov, si bien que les champ conditionnels $\mathbf{Z}|\mathbf{x}$ et $\mathbf{X}|\mathbf{z}$ sont également markoviens. Cette formulation est plus générale que (IV.3) car les hypothèses (B1) et (B2) sont relâchées. En effet, puisque $X_{c_q} = f_{Z_{c_q}}(T_{c_q})$, X_i dépend des Z_j aux sites j situés dans la même clique c_q que i et donc $P(x_i|\mathbf{z}) = P(x_i|z_{c_q}) \neq P(x_i|z_i)$. De même, d'après (IV.4), $P(\mathbf{x}|\mathbf{z}) \neq P(x_i|\mathbf{z})$.

Les modèles (IV.3) et (IV.4) permettent donc de relâcher l'hypothèse de bruit indépendant, (B2) et/ou (B1), tout en gardant la markovianité du couple $(\mathbf{X}, \mathbf{Z})$ et donc des champs conditionnels $\mathbf{Z}|\mathbf{x}$ et $\mathbf{X}|\mathbf{z}$. Il est alors aisé de simuler des données suivant ces modèles (par l'échantionneur de Gibbs par exemple). Néanmoins, ils peuvent difficilement être utilisés dans la pratique car, sans information supplémentaire, ils nécessiteraient d'estimer les **fonctions** $f_1, \cdots, f_K$ (ou $f_{c_1}, \cdots, f_{c_q}$), ce qui est en général plus complexe que d'estimer des paramètres.

1.2 Champs de Markov couples

Notons que, pour garantir la markovianité du champ *a posteriori* (et donc pouvoir utiliser les traitements bayésiens classiques), les hypothèses (A) et (B) sont suffisantes mais non nécessaires. [100] propose alors un modèle plus général, les champs de Markov couples. Ce modèle est défini, non pas à partir des lois de $P(\mathbf{Z})$ et $P(\mathbf{X}|\mathbf{z})$, mais en supposant directement que le couple $(\mathbf{X}, \mathbf{Z})$ est markovien. Soient donc $\mathcal{I}$ un ensemble de sites et $\mathbf{X} = (X_i)_{i\in\mathcal{I}}$ un champ aléatoire réel ($\forall i \in \mathcal{I},\ X_i \in \mathbb{R}^D$). On suppose $\mathcal{I}$ muni d'un système de voisinage défini par un ensemble de cliques $c \in \mathcal{C}$. Soit $\mathbf{Z} = (Z_i)_{i\in\mathcal{I}}$ un champ aléatoire discret ($\forall i \in \mathcal{I},\ Z_i \in \mathcal{K} = [\![1, K]\!]$) **quelconque**.

Définition 14. *Le couple* $(\mathbf{X}, \mathbf{Z})$ *est un champ de Markov couple si sa distribution est markovienne, c'est-à-dire de la forme :*

$$P_G(\mathbf{x}, \mathbf{z}) \propto \exp(-H(\mathbf{x}, \mathbf{z})) = \exp(-\sum_{c\in\mathcal{C}} V_c(\mathbf{x}_c, \mathbf{z}_c)) \tag{IV.5}$$

Le fait de supposer directement la markovianité de $(\mathbf{X}, \mathbf{Z})$ impose celle de $\mathbf{Z}|\mathbf{x}$ et celle de $\mathbf{X}|\mathbf{z}$. En particulier, lorsque $D = 1$, une énergie de la forme :

$$H(\mathbf{x}, \mathbf{z}) = H(\mathbf{z}) - \frac{1}{2}(\mathbf{x} - \mu_{\mathbf{z}})\mathbf{Q}_{\mathbf{z}}(\mathbf{x} - \mu_{\mathbf{z}})', \tag{IV.6}$$

rend à la fois possible la modélisation de textures et les traitements bayésiens sur le modèle. Remarquons que, sous l'hypothèse (IV.5) et en notant W_2 la fonction de partition de la distribution markovienne $P_G(\mathbf{x}, \mathbf{z})$ et $W(\mathbf{z})$ la constante de normalisation de la distribution gaussienne $P(\mathbf{x}|\mathbf{z})$, le champ $\mathbf{Z}$ n'est plus nécessairement markovien puisque

$$\begin{aligned} P(\mathbf{z}) &= \frac{W(\mathbf{z})}{W_2}\exp(-H(\mathbf{z})) \\ \text{où } W(\mathbf{z}) &= (2\pi)^{\frac{n}{2}}(\log|Q_{\mathbf{z}}|)^{-\frac{1}{2}} \end{aligned}$$

et $W(\mathbf{z})$ ne s'écrit pas nécessairement sous forme markovienne. Le modèle de champ de Markov couple (*Pairwise Markov Field*) défini par (IV.5) se distingue donc des champs de Markov cachés habituels du fait de la non-markovianité du champ caché $\mathbf{Z}$. [6] propose de plus une procédure d'estimation des paramètres du modèle basée sur le principe d'ICE [99] et la méthode des moindres carrés de [45]. Remarquons néanmoins que, si l'énergie (IV.6) permet de modéliser simplement le caractère texturé d'une image, elle ne peut modéliser le fait qu'une image soit composée de plusieurs textures. Un modèle plus général est celui des champs de Markov triplets, proposés dans [6]. De tels modèles permettent en effet de modéliser des bruits plus complexes et entre autre de segmenter des images texturées [6] ou encore de modéliser des images non stationnaires [7].

2 Relâcher l'unimodalité : champs de Markov triplets

Soient $\mathcal{I}$ un ensemble de sites et $\mathbf{X} = (X_i)_{i\in\mathcal{I}}$ un champ aléatoire à valeur dans $\mathbb{R}^D$. On suppose $\mathcal{I}$ muni d'un système de voisinage défini par un ensemble de cliques $c \in \mathcal{C}$. Soit $\mathbf{Z} = (Z_i)_{i\in\mathcal{I}}$ un champ aléatoire discret ($\forall i \in \mathcal{I},\ Z_i \in \mathcal{K} = [\![1, K]\!]$). La modélisation par champ de Markov triplet est une extension du champ de Markov couple, dans laquelle on introduit un troisième processus $\mathbf{Y} = (Y_i)_{i\in\mathcal{I}}$ discret, avec pour tout $i \in \mathcal{I}, Y_i \in \mathcal{L} = [\![1, L]\!]$.

Définition 15. *Le triplet* $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ *est un champ de Markov triplet si sa distribution est markovienne, c'est-à-dire de la forme :*

$$P(\mathbf{x}, \mathbf{y}, \mathbf{z}) \propto \exp(-H(\mathbf{x}, \mathbf{y}, \mathbf{z})) = \exp(-\sum_{c \in \mathcal{C}} V_c(\mathbf{x}_c, \mathbf{y}_c, \mathbf{z}_c))$$

Notons que, pour tout $\mathbf{x} \in (\mathbb{R}^D)^n$ et $\mathbf{z} \in \mathcal{K}^n$, nous pouvons toujours écrire :

$$P(\mathbf{x}|\mathbf{z}) = \sum_{\mathbf{y} \in \mathcal{L}^n} P(\mathbf{y}|\mathbf{z}) P(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \sum_{\mathbf{y} \in \mathcal{L}^n} \Pi_{\mathbf{yz}} \ \ f(\mathbf{x}|\theta_{\mathbf{yz}}) \tag{IV.7}$$

La distribution de $\mathbf{X}$ conditionnellement aux classes $\mathbf{Z} = \mathbf{z}$ peut alors être vue comme un mélange de $\mathcal{L}^n$ distributions pour lequel les proportions du mélange, notées $\Pi_{\mathbf{yz}}$, sont les probabilités $P(\mathbf{y}|\mathbf{z})$ et les distributions du mélange sont les lois $P(\mathbf{x}|\mathbf{y}, \mathbf{z}) = f(\mathbf{x}|\theta_{\mathbf{yz}})$. Dans le cas le plus général, la loi du bruit (IV.7) ne vérifie ni (B1), ni (B2). Les couples $\{(l, k), l \in \mathcal{L}\}$ peuvent alors être vus comme les sous-classes de la classe k. L'ensemble des $\{(l, k), l \in \mathcal{L}, k \in \mathcal{K}\}$ désigne alors l'ensemble de toutes les sous-classes du modèle. Sous cette forme, la loi du bruit $P(\mathbf{x}|\mathbf{z})$ n'est alors ni indépendante, ni unimodale.

Le triplet $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ étant markovien, il en est de même des lois conditionnelles $(\mathbf{Y}, \mathbf{Z})|\mathbf{x}$, $(\mathbf{X}, \mathbf{Y})|\mathbf{z}$. En particulier, notons $\mathbf{U} = (\mathbf{Y}, \mathbf{Z}) = (Y_i, Z_i)_{i \in \mathcal{I}}$ le champ aléatoire discret à KL classes ($\forall i \in \mathcal{I}, \ U_i \in \mathcal{K} \times \mathcal{L}$). Par définition, $(\mathbf{X}, \mathbf{U})$ est un champ de Markov couple, donc $\mathbf{U}$ conditionnellement à $\mathbf{X} = \mathbf{x}$ est markovien. Sa distribution s'écrit :

$$\begin{aligned} P(\mathbf{y}, \mathbf{z}|\mathbf{x}) &= P(\mathbf{u}|\mathbf{x}) = W(\mathbf{x})^{-1} \exp(-\sum_{c \in \mathcal{C}} V_c(u_c, x_c)) \\ \text{où} \quad W(\mathbf{x}) &= \sum_{\mathbf{u} \in \mathcal{K}^n \times \mathcal{L}^n} \exp(-\sum_{c \in \mathcal{C}} V_c(u_c, x_c)) \end{aligned}$$

Estimer les paramètres du triplet $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ revient alors à estimer ceux d'un champ de Markov couple à LK classes. Une fois les paramètres estimés, la classification $\mathbf{Z} = \mathbf{z}$ peut être obtenue par la règle du MPM à partir des marginales $P(u_i|\mathbf{x}) = P(y_i, z_i|\mathbf{x})$:

$$z_i^{mpm} = \arg\max_{k \in \mathcal{K}} \sum_{l \in \mathcal{L}} P(Y_i = l, Z_i = k|\mathbf{x}) \tag{IV.8}$$

Notons que lorsque les paramètres ne sont pas connus, sans hypothèse supplémentaire, on se heurte à un problème d'identifiabilité du modèle. En effet, le codage des LK sous-classes n'a pas de signification et est arbitraire de sorte que les numéros attribués aux sous-classes $(u_1, \cdots, u_n)$ en sortie de l'algorithme d'estimation n'ont pas de signification et chacune des $(LK)!$ possibilités conduit à la même vraisemblance. Dès lors, bien que les LK probabilités $P(u_i|\mathbf{x})$ soient calculables, on ne peut retrouver les $P(z_i|\mathbf{x})$ correspondants puisqu'on ne sait pas sur quels numéros u_i il faut sommer.

A titre d'exemple, si les u_i sont classés selon l'ordre lexicographique des (z_i, y_i) comme en Figure IV.1, la règle du MPM (IV.8) s'écrit :

$$z_i^{mpm} = \arg\max_{k \in \mathcal{K}} \sum_{u_i = L(k-1)+1}^{Lk} P(u_i|\mathbf{x})$$

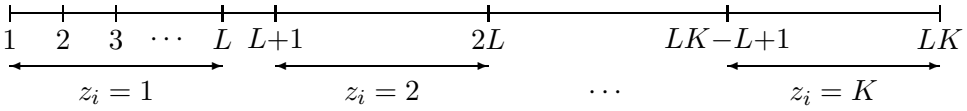


FIG. IV.1 – Lorsque les LK sous-classes sont agencées selon un ordre préétabli, on peut associer à un $u_i \in [\![1, LK]\!]$, la classe $z_i \in \mathcal{K}$ correspondante.

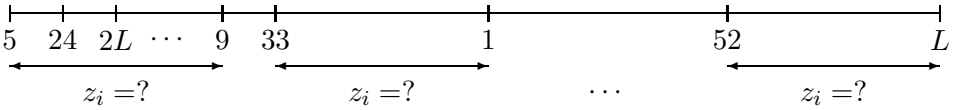


FIG. IV.2 – Lorsque les LK sous-classes sont agencées selon un ordre inconnu, on ne sait pas à associer à un $u_i \in [\![1, LK]\!]$, la classe $z_i \in \mathcal{K}$ correspondante.

car l'on sait que les sous-classes de la classe k sont $L(k-1)+1, \cdots, Lk$. En Figure IV.2, on ne sait *a priori* pas comment ont été numérotées les sous-classes et on ne sait pas retrouver les classes z_i à partir des u_i. Notons que ce problème de *label-switching* disparaît dès lors que, pour chaque classe $k \in \mathcal{K}$, on connaît les L lois $P(\mathbf{x}|Y_i = l, Z_i = k)$ associées à ses L sous-classes $(l, k), l \in \mathcal{L}$. En effet, connaître ces distributions revient à connaître les numéros des u_i correspondant à la classe k (par exemple $5, 24, 2L, \cdots, 9$ pour la classe $k = 1$ de la Figure IV.2). C'est en particulier le cas en classification supervisée, lorsque l'apprentissage permet d'estimer, pour chaque classe k, les lois associées à ses sous-classes $P(\mathbf{x}|Y_i = l, Z_i = k)$.

Signalons que dans les applications de [6] et [7], les champs de Markov triplets considérés sont construits sous l'hypothèse supplémentaire :

$$P(\mathbf{x}|\mathbf{z}, \mathbf{y}) = \prod_{i \in \mathcal{I}} P(x_i|z_i) \tag{IV.9}$$

Remarquons que la dépendance en $\mathbf{y}$ a disparu dans le terme de droite. L'interprétation qui est faite dans [7] est de voir le champ auxiliaire $\mathbf{Y}$ comme indiquant les différentes stationnarités possibles d'une image observée $\mathbf{x}$, le but étant de la segmenter en régions indiquées par les variables de classe Z_i. Dans le cadre de la segmentation d'images texturées, les valeurs des Z_i s'interprètent comme les différentes couleurs de l'image alors que les valeurs des Y_i peuvent être vues comme les textures. Ainsi, dans la Figure 9 de [7] (p. 1377), on distingue 3 couleurs (classes) pour les variables Z_i et 2 valeurs ou textures pour les variables Y_i. Pour ce qui est de l'interprétation en termes de textures, nous verrons dans le chapitre V que nous n'adoptons pas exactement le même point de vue, ce qui peut générer des confusions avec les travaux de [7]. Dans notre interprétation, les classes sont les textures alors que les variables auxiliaires représentent des sous-classes. Ainsi, pour une image comme celle de la Figure 9 de [7], les variables Z_i pourraient prendre 2 valeurs et par exemple 3 couleurs par texture, soit 6 couleurs au total.
Pour revenir à la signification de l'hypothèse (IV.9), le fait que le bruit en chaque site ne dépende pas du régime implique que $P(\mathbf{x}|\mathbf{z}, \mathbf{y}) = P(\mathbf{x}|\mathbf{z}) = \prod_i P(x_i|z_i)$. Cette hypothèse lève le problème de la non-identifiabilité rencontré dans l'équation IV.8 et permet de faire effectivement de la segmentation non supervisée. Si on veut relâcher (IV.9) et ne pas avoir de problème d'identifiabilité, une solution est de se placer dans un cadre supervisé.

Modèle markovien	hypothèse (B1)	hypothèse (B2)	$\mathbf{Z}$ est markovien	$(\mathbf{X}, \mathbf{Z})$ est markovien	$\mathbf{Z}\|\mathbf{x}$ est markovien
caché à bruit indépendant	X	X	X	X	X
gaussien (équation IV.1)	X		X		
à bruit corrélé (équation IV.3)	X		X	X	X
à bruit corrélé (équation IV.4)			X	X	X
couple				X	X
triplet					

FIG. IV.3 – Caractéristiques des différents modèles présentés. Un "X" indique que cette caractéristique est vérifiée par le modèle. L'absence de "X" indique que la propriété n'est pas nécessairement vérifiée par le modèle.

Remarque. *Le champ de Markov Triplet est un modèle plus large que celui de champ de Markov caché car* $\mathbf{Z}$ *n'est pas nécessairement markovien, et plus large que celui par champ de Markov couple car* $(\mathbf{X}, \mathbf{Z})$ *n'est pas nécessairement markovien. Par contre, un champ de Markov caché est un champ de Markov couple (équation IV.4). De plus, si* $(\mathbf{X}, \mathbf{Z})$ *est un champ de Markov couple, le triplet* $(\mathbf{X}, \mathbf{Z}, \mathbf{Z})$ *est un champ de Markov triplet. Les caractéristiques des modèles exposés précédemment sont résumées dans la Table IV.3.*

Chapitre V

Classification supervisée par champ de Markov triplet

1 Problématique

On se place dans le cadre d'une segmentation supervisée, en ce sens que l'on dispose de deux ensembles d'observations, situés en des sites $\mathcal{I}^1$ et $\mathcal{I}^2$. Les observations $\mathbf{x}^1 = (x_i)_{i\in\mathcal{I}^1}$ de $\mathcal{I}^1$ sont étiquetées, nous connaissons donc leurs classes $\mathbf{z}^1 = (z_i)_{i\in\mathcal{I}^1}$. Elle constituent ce qu'on appelle la *base d'apprentissage* (*learning database*). Les données $\mathbf{x}^2 = (x_i)_{i\in\mathcal{I}^2}$ de $\mathcal{I}^2$ sont non-étiquetées. Elles constituent *la base de test* (*test database*). L'objectif est alors, à partir des observations d'apprentissage $\mathbf{x}^1$ et $\mathbf{z}^1$ d'apprendre certains paramètres du modèle, de manière à pouvoir classer dans un second temps les observations de test $\mathbf{x}^2$. On suppose que les données d'apprentissage et de test suivent le même modèle et dans les deux cas, on notera $\mathbf{X}$ les observations et $\mathbf{Z}$ les classes.

Nous nous plaçons dans un cadre où le bruit $P(\mathbf{x}|\mathbf{z})$ n'est ni indépendant, ni assez simple pour être modélisé par une distribution unimodale (gaussienne par exemple). Une idée naturelle pour classer de telles données est alors de décomposer chaque classe $k \in \mathcal{K}$ en sous-classes. Supposons par exemple que chacune des K classes puisse être décomposée en L sous-classes. Pour cela, introduisons un champ auxilliaire $\mathbf{Y} = (Y_i)_{i\in\mathcal{I}}$ discret à valeurs dans $\mathcal{L}^n$ ($\forall i \in \mathcal{I},\ Y_i \in \mathcal{L} = [\![1, L]\!]$). Les sous-classes de la classe $k \in \mathcal{K}$ sont alors les couples (l, k), $l \in \mathcal{L}$. L'ensemble des LK sous-classes correspond alors l'ensemble des couples $(l, k) \in \mathcal{L} \times \mathcal{K}$. Rappelons que la distribution de $\mathbf{X}$ conditionnellement aux classes $\mathbf{Z} = \mathbf{z}$ peut alors être vue comme un mélange de $\mathcal{L}^n$ distributions (voir équation IV.7) et que, dans le cas le plus général, la distribution (IV.7) n'est pas celle d'un bruit indépendant.

Enfin, nous désirons tenir compte des dépendances entre sites, à la fois pour l'apprentissage et pour le test. Pour cela, nous désirons pouvoir utiliser les traitements bayésiens classiques d'estimation des paramètres lors de l'apprentissage et du test. Le modèle adopté devra donc à la fois vérifier :

1. $P(\mathbf{Y}|\mathbf{x}, \mathbf{z})$ est markovien (utilisé lors de la phase d'apprentissage)
2. $P(\mathbf{Y}, \mathbf{Z}|\mathbf{x})$ est markovien (utilisé lors de la phase de test)

Notons que ces deux exigences ne sont pas incompatibles. Plus que cela, supposer $P(\mathbf{Y}, \mathbf{Z}|\mathbf{x})$

markovien impose $P(\mathbf{Y}|\mathbf{x},\mathbf{z})$ markovien. L'unique exigence que nous souhaitons vérifiée par notre modèle est donc de garantir la markovianité de $(\mathbf{Y},\mathbf{Z})$ conditionnellement à $\mathbf{X}=\mathbf{x}$. Pour cela, le modèle génératif le plus général est de considérer que $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ est un champ de Markov triplet, c'est-à-dire, en notant $U_i=(Y_i,Z_i)$, que $(\mathbf{X},\mathbf{U})$ est un champ de Markov couple. Nous détaillons dans la section suivante le modèle adopté, celui d'un champ de markov caché pour $(\mathbf{X},\mathbf{U})$, qui est un cas particulier du champ de Markov couple.

2 Modèle de Markov triplet pour la classification supervisée

Les sites $i\in\mathcal{I}$ considérés ($\mathcal{I}=\mathcal{I}^1$ ou $\mathcal{I}^2$) sont supposés en interaction via un graphe défini par des cliques $c\in\mathcal{C}$. Nous proposons de définir la distribution du triplet $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ de la manière suivante :

$$P_G(\mathbf{x},\mathbf{y},\mathbf{z}) \propto \exp(-\sum_{c\in\mathcal{C}} V_c(\mathbf{y}_c,\mathbf{z}_c) + \sum_{i\in\mathcal{I}} \log f(x_i|\theta_{y_i z_i})) \tag{V.1}$$

$(\mathbf{X},\mathbf{Y},\mathbf{Z})$ est alors un champ de Markov triplet. De plus, (V.1) nous assure que le couple $(\mathbf{Y},\mathbf{Z})$ est markovien :

$$P_G(\mathbf{y},\mathbf{z}) \propto \exp(-\sum_{c\in\mathcal{C}} V_c(\mathbf{y}_c,\mathbf{z}_c)) \tag{V.2}$$

et que la distribution $P(.|\mathbf{y},\mathbf{z})$ vérifie :

$$P(\mathbf{x}|\mathbf{y},\mathbf{z}) = \prod_{i\in\mathcal{I}} P(x_i|y_i,z_i) = \prod_{i\in\mathcal{I}} f(x_i|\theta_{y_i z_i}) \tag{V.3}$$

où, pour tout $l\in\mathcal{L}$ et $k\in\mathcal{K}$, $f(.|\theta_{lk})$ sont des distributions sur $\mathbb{R}^D$ paramétrées par θ_{lk} (par exemple $\theta_{lk}=(\mu_{lk},\Sigma_{lk})$ dans le cas gaussien). Notons $\mathbf{U}=(U_i)_{i\in\mathcal{I}}$ le champ défini par $U_i=(Y_i,Z_i)$. Les équations (V.2) et (V.3) s'écrivent encore :

$$P_G(\mathbf{u}) \propto \exp(-\sum_{c\in\mathcal{C}} V_c(\mathbf{u}_c)) \tag{V.4}$$

$$P(\mathbf{x}|\mathbf{u}) = \prod_{i\in\mathcal{I}} f(x_i|\theta_{u_i}) \tag{V.5}$$

Le champ $(\mathbf{X},\mathbf{U})$ est alors un champ de Markov caché à bruit indépendant. Notons que l'équation (V.1) est plus générale que l'hypothèse faite dans les exemples de [7] qui suppose que $P(\mathbf{x}|\mathbf{y},\mathbf{z})=\prod_i P(x_i|z_i)$ (voir aussi chapitre IV, section 2).

Si l'on se restreint aux interactions d'ordre 2, l'équation (V.2) s'écrit :

$$P_G(\mathbf{y},\mathbf{z}) \propto \exp(-\sum_{i\sim j} V_{ij}(y_i,z_i,y_j,z_j)) \tag{V.6}$$

où V_{ij} sont les potentiels sur les paires. Si de plus ces potentiels sont supposés être les mêmes sur l'ensemble des sites, nous pouvons écrire sans perte de généralité :

$$V_{ij}(y_i,z_i,y_j,z_j) = -B_{z_i z_j}(y_i,y_j) - C(z_i,z_j)$$

où les $B_{kk'}$ sont K^2 fonctions de $\mathcal{L} \times \mathcal{L} \to \mathbb{R}$ et C est une fonction de $\mathcal{K} \times \mathcal{K} \to \mathbb{R}$. En utilisant la notation vectorielle $z_i = k \Leftrightarrow \mathbf{z}_i = e_k$ (le k-ième vecteur canonique en dimension K) et $y_i = l \Leftrightarrow \mathbf{y}_i = e'_l$ (le l-ième vecteur canonique en dimension L), on peut encore écrire :

$$V_{ij}(\mathbf{y}_i, \mathbf{z}_i, \mathbf{y}_j, \mathbf{z}_j) = -\mathbf{y}'_i \mathbb{B}_{z_i z_j} \mathbf{y}_j - \mathbf{z}'_i \mathbb{C} \mathbf{z}_j \tag{V.7}$$

où les $\mathbb{B}_{kk'}$ sont K^2 matrices de taille $L \times L$ et $\mathbb{C}$ est une matrice de dimension $K \times K$ indépendante des z_i. Par symétrie des intéractions, on a alors, pour tout $k, k' \in \mathcal{K}$ et $l, l' \in \mathcal{L}$:

$$B^{ll'}_{kk'} + c_{kk'} = B^{l'l}_{k'k} + c_{k'k} \tag{V.8}$$

où on a noté $B^{ll'}_{kk'}$ la composante (l, l') de la matrice $\mathbb{B}_{kk'}$ et $c_{kk'}$ la composante (k, k') de la matrice $\mathbb{C}$. Notons que l'équation (V.8) implique que, pour tout $k \in \mathcal{K}$, $\mathbb{B}_{kk}$ est symétrique. Nous supposons de plus que, pour tout $k \neq k'$, $\mathbb{B}_{kk'} = \mathbb{B}_{k'k}$ (il y a donc $\frac{1}{2}K(K+1)$ matrices $\mathbb{B}_{kk'}$) et que $\mathbb{C}$ est symétrique. Sous ces hypothèses, les matrices $\mathbb{B}_{kk'}$ sont alors symétriques.
Remarquons que l'écriture (V.7) est toujours possible. Cela revient simplement à voir V_{ij} sous la forme d'une matrice $\mathbb{V}$ de dimension $LK \times LK$, elle-même décomposée comme $L \times L$ blocs de matrices de dimension $K \times K$. En notant $(c_{kk'})_{k,k' \in \mathcal{K}}$ les coefficients de $\mathbb{C}$:

$$-\mathbb{V} = \begin{array}{|c|c|c|} \hline \mathbb{B}_{11}+c_{11} & \cdots & \mathbb{B}_{1K}+c_{1K} \\ \hline \vdots & \ddots & \vdots \\ \hline \mathbb{B}_{K1}+c_{K1} & \cdots & \mathbb{B}_{KK}+c_{KK} \\ \hline \end{array}$$

avec LK au-dessus, L au-dessus du dernier bloc, L à droite du premier bloc et LK à droite de l'ensemble.

Signalons que le terme $\mathbf{z}'_i \mathbb{C} \mathbf{z}_j$ de l'équation (V.7) aurait pu être intégré directement dans celui $\mathbf{y}'_i \mathbb{B}_{z_i z_j} \mathbf{y}_j$ mais l'intérêt d'une telle modélisation apparaîtra ultérieurement. La composante (l, l') de la matrice $\mathbb{B}_{kk'}$ s'interprète comme un coefficient de compatibilité entre les sous-classes l et l' des classes k et k'. Plus ce terme est grand, plus il est vraisemblable que deux sites voisins soient dans les sous-classes l et l' des classes k et k'. De même, le coefficient $c_{kk'}$ de la matrice $\mathbb{C}$ s'interprète comme un coefficient de compatibilité entre les classes k et k'. Plus ce terme est grand, plus il est vraisemblable que les classes k et k' soient voisines. Lorsque $\mathbb{C} = c\mathbb{I}_K$ est paramétré par un unique coefficient spatial

$c \in \mathbb{R}^+$, le terme $\sum_{i\sim j} \mathbf{z}_i' \mathbb{C} \mathbf{z}_j = c \sum_{i\sim j} \mathbb{1}_{z_i=z_j}$ agit alors comme un terme de régularisation favorisant les régions homogènes (c'est-à-dire les régions de même classe).

En utilisant la notation vectorielle $u_i = (l,k) \Leftrightarrow \mathbf{u}_i = e''_{lk}$, le $(k-1)L + l$-ième vecteur de la base canonique en dimension LK, la loi de $\mathbf{U}$ est donnée par :

$$P(\mathbf{u}) \propto \exp(\sum_{i\sim j} \mathbf{u}_i' \mathbb{V} \mathbf{u}_j) \tag{V.9}$$

$(\mathbf{X}, \mathbf{U})$ étant un champ de Markov caché à bruit indépendant, les procédures décrites au chapitre II, section 4, pour estimer les paramètres et/ou classer les données peuvent toutes être appliquées au couple $(\mathbf{X}, \mathbf{U})$. En particulier, la procédure d'estimation de NREM de [29] (voir chapitre II, section 4.4.2) peut être utilisée pour estimer les paramètres $\psi = \{\mathbb{B}_{kk'}, \mathbb{C}, \theta_{lk}, l \in \mathcal{L}, k, k' \in \mathcal{K}\}$ du modèle.

Notons que, sous (V.1), $P(\mathbf{y}|\mathbf{z})$ est également markovien :

$$\Pi_{\mathbf{yz}} = P_G(\mathbf{y}|\mathbf{z}) = \frac{1}{W(\mathbf{z})} \exp(\sum_{i\sim j} \mathbf{y}_i' \mathbb{B}_{z_i z_j} \mathbf{y}_j) \tag{V.10}$$

où la constante de normalisation $W(\mathbf{z})$ dépend de $\mathbf{z}$. Notons que la matrice $\mathbb{C}$ a disparu de (V.10) en comparaison à (V.7). Cette matrice ne peut donc être apprise conditionnellement à $\mathbf{z}$, et ne pourra être estimée que lors de la phase de test. D'après (V.3) et (V.10),

$$P_G(\mathbf{x}, \mathbf{y}|\mathbf{z}) = \frac{1}{W(\mathbf{z})} \exp(\sum_{i\sim j} \mathbf{y}_i' \mathbb{B}_{z_i z_j} \mathbf{y}_j + \log f(x_i|\theta_{y_i z_i})), \tag{V.11}$$

si bien que le couple $(\mathbf{X}, \mathbf{Y})$ est, conditionnellement à $\mathbf{Z} = \mathbf{z}$, un champ de Markov caché à bruit indépendant, sur lequel peuvent donc être appliquées toutes les méthodes bayésiennes d'estimation et/ou classification décrites au chapitre II, section 4. Notons par ailleurs que l'expression (V.11) ne se factorise pas dans le cas général et est donc différente du modèle proposé dans [6] (p. 483).

En résumé, on a donc :

1. $\mathbf{Y}|\mathbf{Z} = \mathbf{z}$, est un champ de Markov et $\mathbf{X}, \mathbf{Y}|\mathbf{Z} = \mathbf{z}$ est un champ de Markov caché à bruit indépendant. Il sera utilisé pour la phase d'apprentissage.
2. $\mathbf{U} = (\mathbf{Y}, \mathbf{Z})$ est un champ de Markov et $(\mathbf{X}, \mathbf{U})$ est un champ de Markov caché à bruit indépendant. Il sera utilisé pour la phase de test.

Notre modèle est donc bien adapté au cadre de la classification supervisée. Les traitements bayésiens classiques pourront être appliqués sur $(\mathbf{X}, \mathbf{Y})$ conditionnellement à $\mathbf{Z} = \mathbf{z}$ pour l'apprentissage. La classification d'observations non étiquetées pourra ensuite être obtenue par application de ces mêmes méthodes sur $(\mathbf{X}, \mathbf{U})$. Notons néanmoins que dans le cas le plus général, ni $\mathbf{Y}$ ni $\mathbf{Z}$ ne sont markoviens. En effet,

$$P(\mathbf{y}) \propto \sum_{\mathbf{z}\in\mathcal{K}^n} \exp(\sum_{i\sim j} \mathbf{y}_i' \mathbb{B}_{z_i z_j} \mathbf{y}_j + \mathbf{z}_i' \mathbb{C} \mathbf{z}_j) \tag{V.12}$$

$$P(\mathbf{z}) \propto \sum_{\mathbf{y}\in\mathcal{L}^n} \exp(\sum_{i\sim j} \mathbf{y}_i' \mathbb{B}_{z_i z_j} \mathbf{y}_j + \mathbf{z}_i' \mathbb{C} \mathbf{z}_j) \tag{V.13}$$

ne se factorisent pas comme l'exponentielle d'une somme sur les cliques.

Faire varier le nombre de sous-classes. Dans le modèle présenté précédemment, chaque $Y_i \in [\![1, L]\!]$, ce qui signifie implicitement que chacune des K classes se décompose en L sous-classes. En pratique, il est important de pouvoir modéliser le fait que les distributions des classes puissent être de formes très diverses, et en particulier être composées d'un nombre variable de sous-classes. Prendre en compte une telle variabilité entre les classes requiert quelques modifications du modèle mais ne change pas fondamentalement la procédure.

Soit $L_1, \cdots, L_K$ le nombre de sous-classes désiré respectivement dans chacune des classes $1, \cdots, K$. Notons $\xi = \{(l,k), l \in [\![1, L_k]\!], k \in [\![1, K]\!]\}$. Remarquons que l'ensemble ξ est inclu dans $[1, \max_k L_k] \times \mathcal{K}$. En gardant la même définition (V.7) pour les potentiels, il est suffisant de remplacer l'équation (V.6) par :

$$\begin{aligned} P_G(\mathbf{y}, \mathbf{z}) &\propto \exp(-\sum_{i \sim j} V_{ij}(y_i, z_i, y_j, z_j)), \quad \text{si } (\mathbf{y}, \mathbf{z}) \in \xi^n \\ &= 0 \quad \text{sinon} \end{aligned}$$

La propriété de factorisation sur les cliques caractéristique des distributions de Gibbs est conservée par une telle définition et le champ $(\mathbf{Y}, \mathbf{Z})$ reste markovien [81].

3 Schéma de classification

Plus qu'un algorithme, nous décrivons un schéma général pour traiter des données issues de classes complexes au sens défini précédemment (bruit non indépendant, distributions des classes non unimodales). Pour estimer les paramètres, nous utilisons l'algorithme d'estimation "flou" NREM (voir chapitre II, section 4.4.2) mais d'autres algorithmes (Gradient Stochastique, ICE ...) auraient pu être utilisés. La classification supervisée est effectuée en deux étapes, l'une d'apprentissage (section 3.1), l'autre de classification (section 3.2).

3.1 Etape d'apprentissage

Nous nous plaçons dans un contexte supervisé où une partie de l'information est disponible via des données d'apprentissage. Nous supposons que, pour un certain nombre de sites $i \in \mathcal{I}^1$, nous observons à la fois x_i et sa classe z_i. Avec le modèle introduit en section 2, seul y_i est donc manquant. Puisque, conditionnellement à $\mathbf{Z} = \mathbf{z}$, $(\mathbf{X}, \mathbf{Y})$ est un champ de Markov caché à bruit indépendant, nous pouvons appliquer une des méthodes du chapitre II, section 4.4, pour estimer les paramètres du modèle conditionnellement aux classes (équations V.3 et V.7), à savoir les matrices $\mathbb{B}_{kk'}$, $k, k' \in \mathcal{K}$ et les θ_{lk}, $l \in \mathcal{L}$ et $k \in \mathcal{K}$. Comme mentionné précédemment, estimer les θ_{lk} est important pour résoudre le problème de *label-switching* lors de l'étape de classification. Les matrices $\mathbb{B}_{kk'}$ estimées lors de l'apprentissage peuvent être, ou non, ré-estimées lors de la phase de classification. Signalons de plus qu'il n'est pas toujours possible d'apprendre toutes les matrices $\mathbb{B}_{kk'}$. En particulier, lorsque la structure de voisinage est telle qu'il n'y a pas de voisins dans les classes k et k', les termes en $\mathbb{B}_{kk'}$ n'apparaîtront pas dans les formules du modèle et cette matrice ne pourra être estimée. Plus généralement, lorsque le nombre de paires de voisins dans les classes k et k' est trop faible, l'estimation de $\mathbb{B}_{kk'}$ sera probablement erronée et

il est alors préférable de l'ignorer.

Enfin, remarquons que, pour toute matrice $(\varepsilon_{kk'})_{k,k'\in\mathcal{K}}$:

$$\frac{\exp(\sum_{i\sim j}\mathbf{y}_i'(\mathbb{B}_{z_iz_j}+\varepsilon_{z_iz_j})\mathbf{y}_j)}{\sum_{\mathbf{y}\in\mathcal{L}^n}\exp(\sum_{i\sim j}\mathbf{y}_i'(\mathbb{B}_{z_iz_j}+\varepsilon_{z_iz_j})\mathbf{y}_j)}=\frac{\exp(\sum_{i\sim j}\mathbf{y}_i'\mathbb{B}_{z_iz_j}\mathbf{y}_j)}{\sum_{\mathbf{y}\in\mathcal{L}^n}\exp(\sum_{i\sim j}\mathbf{y}_i'\mathbb{B}_{z_iz_j}\mathbf{y}_j)}=P_G(\mathbf{y}|\mathbf{z})$$

si bien que, lors de l'apprentissage sur $(\mathbf{X},\mathbf{Y})$ conditionnellement à $\mathbf{Z}=\mathbf{z}$, les matrices $\mathbb{B}_{kk'}$ sont estimées à un coefficient additif près $\varepsilon_{kk'}$. Or, d'après (V.7),

$$\frac{\exp(\sum_{i\sim j}\mathbf{y}_i'(\mathbb{B}_{z_iz_j}+\varepsilon_{z_iz_j})\mathbf{y}_j+\mathbf{z}_i'\mathbb{C}\mathbf{z}_j)}{\sum_{\mathbf{z}\in\mathcal{L}^n}\sum_{\mathbf{y}\in\mathcal{L}^n}\exp(\sum_{i\sim j}\mathbf{y}_i'(\mathbb{B}_{z_iz_j}+\varepsilon_{z_iz_j})\mathbf{y}_j+\mathbf{z}_i'\mathbb{C}\mathbf{z}_j)}\neq P(\mathbf{y},\mathbf{z})$$

Plusieurs estimateurs de $\mathbb{B}_{kk'}$, aussi problables les uns que les autres, conduisent donc à des modèles pour $(\mathbf{Y},\mathbf{Z})$ différents. Néanmoins, ce problème d'identifiabilité est résolu lorsque l'estimation de $\mathbb{C}$ est effectuée lors de l'étape de classification. En d'autre terme, si les matrices $\mathbb{B}_{kk'}+\varepsilon_{kk'}$ plutôt que $\mathbb{B}_{kk'}$ ont été apprises lors de la phase d'apprentissage, le modèle conduira à estimer lors de la phase de classification la matrice $(c_{kk'}-\varepsilon_{kk'})_{k,k'\in\mathcal{K}}$ plutôt que $\mathbb{C}=(c_{kk'})_{k,k'\in\mathcal{K}}$ mais les lois correspondantes, et donc les classifications obtenues, restent les mêmes.

Remarque. *Supposons que l'apprentissage soit effectué classe par classe, indépendamment les unes des autres (c'est par exemple le cas avec les images uni-textures d'apprentissage de l'application détaillée au chapitre VI, section 2). D'après l'équation* (V.10)*, on a :*

$$P_G(\mathbf{y}|\forall i\in\mathcal{I}^1,\ z_i=k)\propto\exp(\sum_{i\sim j}\mathbf{y}_i'\mathbb{B}_{kk}\mathbf{y}_j) \tag{V.14}$$

qui est l'équation d'un modèle de Potts étendu (voir chapitre I, section 1.3.2) de matrice de compatibilité $\mathbb{B}_{kk}$. L'apprentissage de la classe k revient alors à estimer un modèle de Potts étendu, c'est à dire à appliquer exactement l'algorithme NREM décrit au chapitre I, section 4.4.

L'algorithme NREM en pratique. Nous détaillons la procédure d'estimation lors de la phase d'apprentissage par l'algorithme NREM (voir chapitre I, section 4.4.2), bien que d'autres algorithmes (ICE...) auraient pu être utilisés. Notons $\mathbb{B}=(\mathbb{B}_{kk'})_{k,k'\in\mathcal{K}}$ les paramètres de la distribution markovienne $P_G(\mathbf{Y}|\mathbf{z})$ (équations V.10) et $\boldsymbol{\theta}=(\theta_{lk})_{l\in\mathcal{L},k\in\mathcal{K}}$. Soit encore $\boldsymbol{\psi}=(\boldsymbol{\theta},\mathbb{B})$ l'ensemble des paramètres du champ de Markov caché à bruit indépendant $(\mathbf{X},\mathbf{Y})$ conditionnellement à $\mathbf{Z}=\mathbf{z}$ (équation V.11). L'approximation en champ moyen de la distribution markovienne $P_G(\mathbf{y}|\mathbf{z})$ obtenue en fixant le champ des voisins à $\tilde{\mathbf{y}}$ (éventuellement dépendant des observations $\mathbf{x}^1$ et de leurs classes $\mathbf{z}^1$) s'écrit :

$$P_G(\mathbf{y}|\mathbf{z})\approx\prod_{i\in\mathcal{I}^1}P_G(y_i|\tilde{\mathbf{y}}_{N_i},\mathbf{z})\propto\prod_{i\in\mathcal{I}^1}\exp(\sum_{j\in N_i}\mathbf{y}_i'\mathbb{B}_{z_iz_j}\tilde{\mathbf{y}}_j) \tag{V.15}$$

Le principe de l'algorithme consiste à alterner une étape (NR) de choix du voisinage pour mettre à jour le champ des voisins $\tilde{\mathbf{y}}$ utilisé dans l'approximation de type champ moyen et

une étape (EM) d'estimation des paramètres sur le mélange indépendant défini par (V.3) et l'approximation (V.15). L'espérance à maximiser à l'itération $(q+1)$ s'écrit :

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)}) = \langle \log P_G(\mathbf{x}^1, \mathbf{Y}|\mathbf{z}^1, \boldsymbol{\psi})|\mathbf{x}^1, \boldsymbol{\psi}^{(q)}\rangle \tag{V.16}$$

Pour alléger les notations, nous noterons x_i (resp. z_i) plutôt que x_i^1 (resp. z_i^1) l'observation (resp. la classe) d'apprentissage au site $i \in \mathcal{I}^1$. Partant des valeurs initiales $\tilde{\mathbf{y}}$ du champ des voisins et $\boldsymbol{\psi}^{(0)}$ des paramètres, l'itération $(q+1)$ de l'algorithme est la suivante :

(E) Calcul des probabilités a posteriori pour tout $i \in \mathcal{I}^1$ et $l \in \mathcal{L}$:

$$\tilde{t}_{il}^{(q)} = P_G(Y_i = l|\mathbf{x}^1, \mathbf{z}^1, \boldsymbol{\psi}^{(q)}) \approx \frac{\tilde{\pi}_{il}^{(q)} f(x_i|\theta_{lz_i}^{(q)})}{\sum_{l' \in \mathcal{L}} \tilde{\pi}_{il'}^{(q)} f(x_i|\theta_{l'z_i}^{(q)})}$$

$$\text{où } \tilde{\pi}_{iy_i}^{(q)} = P_G(y_i|\tilde{\mathbf{y}}_{N_i}, \mathbf{z}, \mathbb{B}^{(q)}).$$

(M) Mise à jour des paramètres $\mathbb{B} = (\mathbb{B}_{kk'})_{k,k' \in \mathcal{K}}$ et $\boldsymbol{\theta} = (\theta_{lk})_{l \in \mathcal{L}, k \in \mathcal{K}}$:

$$\mathbb{B}^{(q+1)} = \arg\max_{\mathbb{B}} \sum_{i \in \mathcal{I}^1} \sum_{l \in \mathcal{L}} \tilde{t}_{il}^{(q)} \log \tilde{\pi}_{il} \tag{V.17}$$

$$\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}} \sum_{i \in \mathcal{I}^1} \sum_{l \in \mathcal{L}} \tilde{t}_{il}^{(q)} \log f(x_i|\theta_{lz_i}) \tag{V.18}$$

$$\text{où } \tilde{\pi}_{il} = P_G(Y_i = l|\tilde{\mathbf{y}}_{N_i}, \mathbf{z}, \mathbb{B}).$$

(NR) Choix des voisins : créer, à partir des observations $\mathbf{x}^1$, de leurs étiquettes $\mathbf{z}^1$ et de l'estimation courante $\boldsymbol{\psi}^{(q+1)}$ des paramètres, un nouveau champ des voisins $\tilde{\mathbf{y}}$:

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

En particulier, dans le cas de densités $f(.|\theta_{lk}) \sim \mathcal{N}(\mu_{lk}, \Sigma_{lk})$, la mise à jour des paramètres est explicite et donnée par les équations :

$$\mu_{lk}^{(q+1)} = \frac{\sum_{i/z_i=k} t_{il}^{(q)} x_i}{\sum_{i/z_i=k} t_{il}^{(q)}} \tag{V.19}$$

$$\Sigma_{lk}^{(q+1)} = \frac{\sum_{i/z_i=k} t_{il}^{(q)} (x_i - \mu_{lk}^{(q+1)})(x_i - \mu_{lk}^{(q+1)})'}{\sum_{i/z_i=k} t_{il}^{(q)}} \tag{V.20}$$

$\mu_{lk}^{(q+1)}$ et $\Sigma_{lk}^{(q+1)}$ s'interprètent alors comme les moyenne et variance empiriques calculées sur les observations d'apprentissage $(x_i)_{i \in \mathcal{I}^1}$ de la classe k, affectées des poids $(t_{il}^{(q)})_{i \in \mathcal{I}^1}$. Il n'y a pas de formule explicite pour la mise à jour des paramètres $\mathbb{B}$, mais ils peuvent être obtenus par descente de gradient (voir Annexe 5).

3.2 Etape de classification

Lors de cette phase, les observations $\mathbf{x}^2$ aux sites $i \in \mathcal{I}^2$ sont non étiquetées, si bien que les champs $\mathbf{Y}$ et $\mathbf{Z}$ sont manquants. Avec $\mathbf{U} = (\mathbf{Y}, \mathbf{Z})$, le couple $(\mathbf{X}, \mathbf{U})$ est un champ de Markov caché à bruit indépendant sur lequel on peut appliquer les traitements bayésiens décrits au chapitre I, section 4. Les paramètres du modèle (équation V.1) sont alors les K^2 matrices $\mathbb{B}_{kk'}$ de dimension $L \times L$, les LK paramètres θ_{lk}, ainsi que la matrice additionnelle $\mathbb{C}$ de dimension $K \times K$. Lors de cette étape, les θ_{lk} sont supposés fixés aux valeurs estimées lors de la phase d'apprentissage. Concernant les $\mathbb{B}_{kk'}$, plusieurs stratégies sont envisageables : les fixer entièrement aux valeurs estimées, en partie, ou les ré-estimer totalement. La meilleure stratégie à adopter dépend bien entendu de la base d'apprentissage et du type d'interaction que l'on souhaite considérer. La matrice $\mathbb{C}$, elle, devra être estimée.

L'algorithme NREM en pratique. L'algorithme est appliqué au champ de Markov caché à bruit indépendant $(\mathbf{X}, \mathbf{U})$ (équation V.4 et V.5). Les paramètres du modèles sont alors $\boldsymbol{\psi}' = (\boldsymbol{\psi}, \mathbb{C})$. L'approximation de type champ moyen de $P_G(\mathbf{U})$ en fixant les voisins au champ $\tilde{\mathbf{u}}$ (éventuellement dépendant des observations $\mathbf{x}^2$) s'écrit :

$$P_G(\mathbf{u}) \approx \prod_{i \in \mathcal{I}^2} P_G(u_i | \tilde{\mathbf{u}}_{N_i}) \propto \prod_{i \in \mathcal{I}^2} \exp\Big(\sum_{j \in N_i} \mathbf{u}_i' \mathbb{V} \tilde{\mathbf{u}}_j\Big) \tag{V.21}$$

Le principe de l'algorithme consiste à alterner une étape (NR) de choix du voisinage pour mettre à jour le champ des voisins $\tilde{\mathbf{u}}$ utilisé dans l'approximation de type champ moyen et une étape (EM) d'estimation des paramètres sur le mélange indépendant défini par l'approximation (V.21). L'espérance à maximiser à l'itération $(q+1)$ s'écrit :

$$Q(\boldsymbol{\psi}' | \boldsymbol{\psi}'^{(q)}) = \langle \log P_G(\mathbf{x}^2, \mathbf{U} | \boldsymbol{\psi}') | \mathbf{x}^2, \boldsymbol{\psi}'^{(q)} \rangle \tag{V.22}$$

Partant des valeurs initiales $\tilde{\mathbf{u}}$ du champ des voisins et $\boldsymbol{\psi}^{(0)}$ des paramètres, l'itération $(q+1)$ de l'algorithme est la suivante :

(E) Calcul des probabilités a posteriori pour tout $i \in \mathcal{I}^2$, $l \in \mathcal{L}$ et $k \in \mathcal{K}$:

$$\tilde{t}_{ilk}^{(q)} = P_G(y_i = l, z_i = k | \mathbf{x}^2, \boldsymbol{\psi}'^{(q)}) \approx \frac{\tilde{\pi}_{ilk}^{(q)} f(x_i | \theta_{lk}^{(q)})}{\sum_{l' \in \mathcal{L}} \sum_{k' \in \mathcal{K}} \tilde{\pi}_{il'k'}^{(q)} f(x_i | \theta_{l'k'}^{(q)})}$$

où $\tilde{\pi}_{ilk}^{(q)} = P_G(U_i = e''_{lk} | \tilde{\mathbf{u}}_{N_i}, \mathbb{V}^{(q)})$.

(M) Mise à jour des paramètres $\mathbb{B} = (\mathbb{B}_{kk'})_{k,k' \in \mathcal{K}}$ (éventuellement) et $\mathbb{C}$:

$$(\mathbb{B}, \mathbb{C})^{(q+1)} = \arg\max_{(\mathbb{B}, \mathbb{C})} \sum_{i \in \mathcal{I}^1} \sum_{l \in \mathcal{L}} \sum_{k \in \mathcal{K}} \tilde{t}_{ilk}^{(q)} \log \tilde{\pi}_{ilk}$$

(NR) Choix des voisins : créer, à partir des observations $\mathbf{x}^2$ et de l'estimation courante $\boldsymbol{\psi}'^{(q+1)}$ des paramètres, un nouveau champ des voisins $\tilde{\mathbf{u}}$:
- *Algorithme en champ moyen* : fixer $\tilde{\mathbf{u}}$ à l'estimation en champ moyen de l'espérance de la distribution conditionnelle $P_G(\mathbf{u} | \mathbf{x}^2, \boldsymbol{\psi}'^{(q+1)})$

– *Algorithme en champ modal* : fixer $\tilde{\mathbf{u}}$ à l'estimation en champ modal du mode de la distribution conditionnelle $P_G(\mathbf{u}|\mathbf{x}^2, \boldsymbol{\psi}'^{(q+1)})$
– *Algorithme en champ simulé* : simuler $\tilde{\mathbf{u}}$ selon la loi conditionnelle $P_G(\mathbf{u}|\mathbf{x}^2, \boldsymbol{\psi}'^{(q+1)})$, via l'échantillonneur de Gibbs.

Il n'y a pas de formule explicite pour la mise à jour des paramètres $\mathbb{V} = (\mathbb{B}, \mathbb{C})$, mais ils peuvent être obtenus par descente de gradient (voir Annexe 5).

3.3 Nécessité du cadre supervisé

Nous mettons ici en évidence le problème d'identifiabilité de nos modèles dans un cadre non supervisé. Nous nous plaçons dans un cadre d'approximation en champ moyen pour illustration. Considérons la vraisemblance des observations $\mathbf{x}$ donnée par :

$$P(\mathbf{x}|\boldsymbol{\psi}') = \sum_{\mathbf{y},\mathbf{z}} P(\mathbf{x}|\mathbf{y},\mathbf{z},\boldsymbol{\theta}) P(\mathbf{y},\mathbf{z}|\mathbb{B},\mathbb{C})$$
$$\text{avec } P(\mathbf{x}|\mathbf{y},\mathbf{z},\boldsymbol{\theta}) = \prod_i f(x_i|\theta_{y_i z_i}) \tag{V.23}$$

Dans une approche de type champ moyen-EM, à l'étape EM de l'algorithme NREM, on aurait :

$$P(\mathbf{y},\mathbf{z}|\mathbb{B},\mathbb{C}) \approx \prod_i P(y_i, z_i|\tilde{y}_{N_i}, \tilde{z}_{N_i}, \mathbb{B}, \mathbb{C})$$

et donc

$$\begin{aligned} P(\mathbf{x}|\boldsymbol{\psi}') &\approx \sum_{\mathbf{y},\mathbf{z}} \prod_i f(x_i|\theta_{y_i z_i}) P(y_i, z_i|\tilde{y}_{N_i}, \tilde{z}_{N_i}, \mathbb{B}, \mathbb{C}) \\ &= \prod_i \sum_{y_i, z_i} f(x_i|\theta_{y_i z_i}) P(y_i, z_i|\tilde{y}_{N_i}, \tilde{z}_{N_i}, \mathbb{B}, \mathbb{C}) \end{aligned}$$

Posons alors $\tilde{\pi}_{ilk} = P(Y_i = l, Z_i = k|\tilde{y}_{N_i}, \tilde{z}_{N_i}, \mathbb{B}, \mathbb{C})$ et $\tilde{q}_{ik} = \sum_{l=1}^{L} \tilde{\pi}_{ilk}$. Il vient :

$$P(\mathbf{x}|\boldsymbol{\psi}') \approx \prod_i \sum_{k=1}^{K} \tilde{q}_{ik} \sum_{l=1}^{L} f(x_i|\theta_{lk}) \frac{\pi_{ilk}}{\tilde{q}_{ik}} \tag{V.24}$$

de sorte qu'en entrée de l'étape EM de l'algorithme NREM, $P(\mathbf{x}|\boldsymbol{\psi})$ est approximée par un mélange de K mélanges à L composantes. Ceci pose un problème dans le cas non supervisé car de tels mélanges ne sont pas identifiables (voir Annexe 1). Néanmoins, notons que si, à la place de (V.23), on fait l'hypothèse de [7], à savoir

$$P(\mathbf{x}|\mathbf{y},\mathbf{z},\boldsymbol{\theta}) = \prod_i P(x_i|z_i),$$

cela revient à supposer que $L = 1$ dans l'expression (V.24) et on obtient un mélange classique identifiable à permutation des classes près. Notons que ce problème n'est pas propre au champ moyen mais lié au modèle. Pour y remédier, une solution est l'ajout d'une étape d'apprentissage, ce qui revient à imposer un ordre sur les paramètres et lève le problème d'identifiabilité (voir Annexe 1).

4 Sélection du modèle de Markov triplet

Choisir le modèle probabiliste le plus approprié aux données observées est une étape importante pour pouvoir classer "au mieux" les observations. Le critère BIC (voir chapitre II, section 5.2) [114] est le plus généralement utilisé pour sélectionner entre des modèles de paramétrisation différentes, celui qui semble le mieux adapté. Concernant le modèle de champ de Markov triplet (V.1), plusieurs caractéristiques peuvent être sélectionnées : le nombre de sous-classes, la forme de chacune des matrices $\mathbb{B}_{kk'}$, les distributions $f(.|\theta_{lk})$ et la forme de la matrice $\mathbb{C}$. Pour que le modèle soit consistant, ces 3 premières caractéristiques doivent être sélectionnées lors de l'apprentissage, alors que la forme de la matrice $\mathbb{C}$ ne peut être sélectionnée que lors de la phase de test.

4.1 Sélection du nombre de sous-classes

On peut choisir de déterminer le nombre de sous-classes de chacune des classes (c'est-à-dire la valeur L_k de chaque classe $k \in \mathcal{K}$). Il n'est néanmoins pas possible de faire la sélection de tous les nombres de sous-classes en même temps, car si, pour chaque k, L_k peut prendre M valeurs possibles, $(L_1, \cdots, L_K)$ peut prendre M^K valeurs, qui est exponentiel. A titre d'exemple, si on dispose de $K = 4$ classes et si on souhaite tester un nombre de sous-classes égal à $2, 3, 4, 5$ ou 6, il faudra calculer $5^4 = 625$ critères BIC, et donc effectuer 625 procédures d'estimation des paramètres correspondant à chacun des modèles. Une alternative est de sélectionner le nombre de sous-classes L_k de chaque classe k indépendamment les unes des autres, c'est-à-dire en n'utilisant, pour le calcul du critère BIC, que les données d'apprentissage étiquetées dans la classe k. Notons que, en séparant de cette sorte les données de chaque classe, nous brisons les arcs reliant deux sites voisins situés dans des classes k et k' distinctes. Pour sélectionner le nombre de sous-classes de la classe k, il faut alors calculer la valeur du critère BIC (ou son approximation en champ moyen, voir chapitre II, section 5.2) sur les données de cette classe, pour différentes valeurs de L_k. Pour détecter le L_k le "meilleur", il faut alors tracer la courbe du BIC en fonction du nombre de sous-classes et y détecter un coude ou un maximum (voir figure VIII.17 pour illustration). Il existe des critères permettant de détecter automatiquement ce coude comme le critère EL (*Elbow Likelihood*) de [40] ou le critère de [12] basé sur la notion de précision des données.

4.2 Sélection de la forme des matrices $\mathbb{B}_{\mathbf{kk'}}$

Pour simplifier les notations, nous nous plaçons dans le cas où toutes les classes ont le même nombre L de sous-classes. Sous la forme la plus générale, chacune des K^2 matrices $\mathbb{B}_{kk'}$ est alors composée de $\frac{1}{2}L(L+1)$ termes distincts, soit un total de $\nu = \frac{1}{2}K^2L(L+1)$ paramètres à estimer. A titre d'exemple, lorsque $K = 10$ et $L = 5$, $\nu = 1500$. Rappelons que les termes $B^{ll'}_{kk'}$ modélisent la compatibilité entre les sous classes l et l' des classes k et k'. Pour que ces termes soient bien estimés, il faut donc que les données d'apprentissage contiennent un grand nombre de sites k et k' voisins les uns des autres. Aussi est-il naturel de supposer que, pour $k \neq k'$, $\mathbb{B}_{kk'} = 0_L$ (la matrice nulle de dimension $L \times L$). Notons que, du point de vue de la modélisation de $P_G(\mathbf{y}, \mathbf{z})$, supposer $\mathbb{B}_{kk'} = b_{kk'}\mathbb{1}_L$ (où $\mathbb{1}_L$ désigne la matrice $L \times L$ composée uniquement de 1) revient à considérer $\mathbb{B}_{kk'} = 0_L$ puisque cette interaction générale entre les classes k et k' est déjà prise en compte à travers la

composante $c_{kk'}$ de $\mathbb{C}$. Pour $\mathbb{B}_{kk'} = 0_L$ lorsque $k \neq k'$, le nombre de paramètres à estimer est alors $\nu = \frac{1}{2}KL(L+1)$ ($\nu = 150$ lorsque $K = 10$ et $L = 5$). Concernant les K matrices $\mathbb{B}_{kk}$ (les blocs diagonaux de $\mathbb{V}$), plusieurs modèles peuvent être envisagés, selon le type de corrélations considérées. Soit la décomposition :

$$\mathbb{B}_{kk} = \Delta_k + \Gamma_k$$

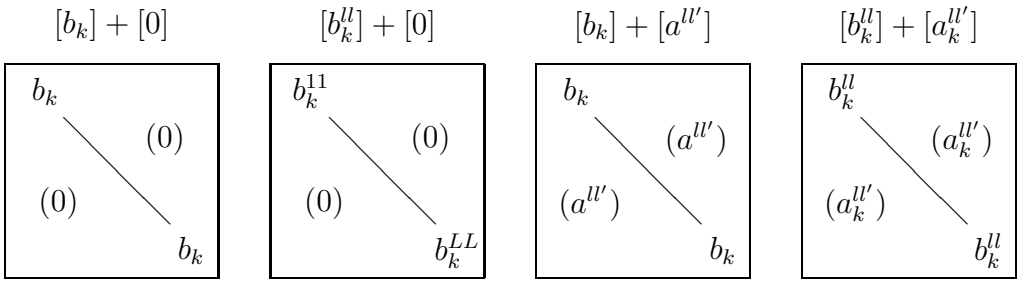
où Δ_k est une matrice diagonale de dimension $L \times L$ et Γ_k une matrice de dimension $L \times L$ de diagonale nulle. Cinq hypothèses peuvent être faites sur Δ_k :

$\Delta_k = [0]$	tous les termes diagonaux sont nuls.
$\Delta_k = [b]$	tous les termes diagonaux sont égaux à b, indépendamment de k.
$\Delta_k = [b_k]$	tous les termes diagonaux sont égaux à b_k et dépendent de k.
$\Delta_k = [b^{ll}]$	les termes diagonaux varient, indépendamment de k.
$\Delta_k = [b_k^{ll}]$	les termes diagonaux varient et dépendent de k.

De même trois hypothèses peuvent être faites sur Γ_k :

$\Gamma_k = [0]$	tous les termes sont nuls.
$\Gamma_k = [a^{ll'}]$	les termes non diagonaux varient, indépendamment de k.
$\Gamma_k = [a_k^{ll'}]$	les termes non diagonaux varient et dépendent de k.

Notons que les modèles $\Gamma_k = [a]$ et $\Gamma_k = [a_k]$ (tous les termes hors diagonaux sont égaux à a ou a_k) sont équivalents, du point de vue de la loi $P_G(\mathbf{y}, \mathbf{z})$, au modèle $\Gamma_k = [0]$ puisque cette corrélation est déjà modélisée via le terme c_{kk} de $\mathbb{C}$. Par combinaisons des différents modèles possibles pour Δ_k et Γ_k, nous obtenons alors 15 modèles pour $\mathbb{B}_{kk}$. A titre d'exemple, voici quatre combinaisons possibles des Δ_k et Γ_k, ainsi que le nombre de paramètres ν_k correspondant :



$\nu_k = 1$ $\quad$ $\nu_k = L$ $\quad$ $\nu_k = 1 + \frac{L(L-1)}{2}$ $\quad$ $\nu_k = \frac{L(L+1)}{2}$

Remarquons encore que le modèle $[0] + [0]$ revient à supposer (voir équation V.12) que :

$$P_G(\mathbf{y}) = \frac{1}{L^n}$$

et donc que les sous-classes $\mathbf{Y}$ sont indépendantes et identiquement distribuées. De plus, d'après l'équation (V.13), on a alors :

$$P_G(\mathbf{z}) \propto \exp(\sum_{i \sim j} \mathbf{z}_i' \mathbb{C} \mathbf{z}_j)$$

si bien que le champ des classes $\mathbf{Z}$ est markovien.

Parmi les modèles précédemment décrits, le modèle sélectionné correspondra à celui ayant le critère BIC le plus élevé.

Remarque. *Si l'apprentissage se fait classe par classe, indépendamment les unes des autres, les modèles* $\Delta_k = [b]$, $\Delta_k = [b^{ll}]$ *et* $\Gamma_k = [a^{ll'}]$ *ne peuvent être considérés puisqu'ils sont indépendants de* k. *Or, l'apprentissage de la classe* k *revient à apprendre un modèle de Potts étendu de corrélation spatiale* $\mathbb{B}_{kk}$ *(voir la remarque de la section 3.1). Les modèles* $[0] + [a_k^{ll'}]$ *et* $[b_k] + [a_k^{ll'}]$ *sont alors équivalents puisque, dans l'équation (V.14), la matrice* $\mathbb{B}_{kk}$ *est définie à un coefficient additif près. En éliminant le cas indépendant* $[0] + [0]$, *il ne reste donc, pour chaque matrice* $\mathbb{B}_{kk}$, *que quatre modèles possibles, à savoir* $[b_k] + [0]$, $[b_k^{ll}] + [0]$, $[b_k] + [a_k^{ll'}]$ *et* $[b_k^{ll}] + [a_k^{ll'}]$.

4.3 Sélection de la forme de la matrice $\mathbb{C}$

En décomposant la matrice $\mathbb{C}$ en termes diagonaux et hors diagonaux, on peut dégager trois modèles distincts :

$\mathbb{C} = [c]$	les termes sont égaux à c sur la diagonale, 0 hors diagonale.
$\mathbb{C} = [c_k]$	les termes sont égaux à c_k sur la diagonale, 0 hors diagonale.
$\mathbb{C} = [c_{kk'}]$	tous les termes varient.

Notons que l'hypothèse $\mathbb{C} = [0]$ ne peut être prise en compte car, à l'apprentissage, les matrice $\mathbb{B}_{kk'}$ sont apprises à un coefficient additif près et c'est cette matrice $\mathbb{C}$ qui permet de recaler les paramètres. Dans la pratique, l'hypothèse $\mathbb{C} = [c_k]$ ou $[c]$ semble être suffisante. Bien que $\mathbf{Z}$ ne soit en général pas markovien, le terme $\mathbf{z}_i' \mathbb{C} \mathbf{z}_j = c \sum_{i \sim j} \mathbb{1}_{z_i = z_j}$ agit alors comme un terme de régularisation (ou de lissage) sur la classification.

4.4 Sélection des lois des classes

Concernant les distributions $f(.|\theta_{lk})$ des LK sous-classes, le modèle le plus répandu est le modèle gaussien, pour lequel θ_{lk} se décompose en un vecteur moyenne μ_{lk} de dimension D (la dimension des observations x_i) et en une matrice de covariance Σ_{lk} symétrique de dimension $D \times D$. Plusieurs paramétrisation de Σ_{lk} sont envisageables, allant du plus parcimonieux ($\Sigma_{lk} = \sigma_{lk}^2 I_{LK}$), au modèle complet $\Sigma_{lk} = (\Sigma_{lk}^{dd'})_{d,d' \in [\![1,D]\!]}$. Dans le cas de données de grande dimension, il peut être intéressant d'utiliser la paramétrisation décrite au chapitre III, section 2. Parmi ces modèles, le modèle sélectionné correspondra à celui ayant le critère BIC le plus élevé.

Chapitre VI

Expériences

Dans ce chapitre, nous nous intéressons tout d'abord (section 1) à un exemple très simple de champ de Markov $(\mathbf{Y}, \mathbf{Z})$ et mettons en évidence un phénomène de transition de phase pour ces modèles avec potentiellement un effet sur l'estimation en pratique du modèle de Markov triplet correspondant. Nous montrons ensuite (section 2) l'efficacité des modèles de Markov triplets plus généraux sur une application à la classification de textures réelles.

1 Illustration d'un phénomène de transition de phase

Soit $\mathcal{I}$ un ensemble de n sites muni d'une structure de voisinage, $\mathbf{X} = (X_i)_{i\in\mathcal{I}}$ un champ aléatoire réel sur $\mathcal{I}$ et $\mathbf{Y} = (Y_i)_{i\in\mathcal{I}}$, $\mathbf{Z} = (Z_i)_{i\in\mathcal{I}}$ deux champs aléatoires discrets, à valeurs respectivement dans $\mathcal{L}^n = [\![1, L]\!]^n$ et $\mathcal{K}^n = [\![1, K]\!]^n$. Nous nous intéressons au modèle de Markov triplet $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ défini par :

$$P(\mathbf{x}, \mathbf{y}, \mathbf{z}) \propto \exp(b \sum_{i\sim j} 1_{y_i=y_j} \mathbb{1}_{z_i=z_j} + c \sum_{i\sim j} \mathbb{1}_{z_i=z_j} + \sum_{i\in\mathcal{I}} \log f(x_i|\theta_{y_i z_i})) \tag{VI.1}$$

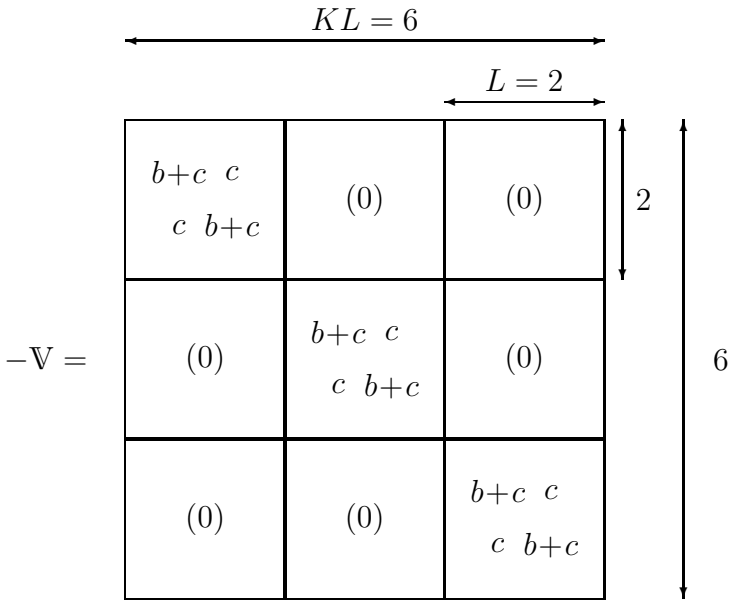
paramétré par les réels b et c, ainsi que les paramètres $\theta_{lk} = (\mu_{lk}, \Sigma_{lk})$ des distributions gaussiennes $f(.|\theta_{lk})$, pour $l \in \mathcal{L}$ et $k \in \mathcal{K}$. Le couple $(\mathbf{Y}, \mathbf{Z})$ est alors markovien, de distribution :

$$P(\mathbf{y}, \mathbf{z}) \propto \exp(b \sum_{i\sim j} 1_{y_i=y_j} 1_{z_i=z_j} + c \sum_{i\sim j} 1_{z_i=z_j}) \tag{VI.2}$$

En comparaison avec l'équation (V.7) cela revient à supposer que :

- pour tout $k \in \mathcal{K}$, $\mathbb{B}_{kk} = b\mathbb{I}_L$ et pour tout $k' \neq k$, $\mathbb{B}_{kk'} = 0_L$ (la matrice nulle)
- la matrice $\mathbb{C}$ est diagonale, ses termes diagonaux sont égaux à c

Pour $K = 3$ et $L = 2$, la matrice $\mathbb{V}$ de (V.9) est alors de la forme :



Plusieurs cas particuliers sont à souligner :

- Pour $L = 1$, le modèle (VI.2) est un modèle de Potts à K classes et coefficient de régularité égal à $b + c$.
- Pour $K = 1$, il s'agit d'un modèle de Potts à L classes et coefficient de régularité b.
- Pour $c = 0$, il s'agit d'un modèle de Potts à LK classes et coefficient de régularité b.

Transition de phase Comme décrit au chapitre I, section 1.4, il est possible d'observer sur les champs de Markov un phénomène de transition de phase correspondant aux valeurs $\boldsymbol{\phi}_c$ des paramètres pour lesquelles la fonction de partition $W(\boldsymbol{\phi})$ cesse d'être analytique lorsque le nombre de sites $n \to \infty$. Pour le modèle markovien défini par l'équation (VI.2), $\boldsymbol{\phi} = (b, c)$ et on a :

$$\begin{aligned}
\log W(\boldsymbol{\phi}) &= \log[\sum_{\mathbf{y},\mathbf{z}} \exp(b \sum_{i\sim j} \mathbb{1}_{y_i=y_j}\mathbb{1}_{z_i=z_j} + c \sum_{i\sim j} \mathbb{1}_{z_i=z_j})] \\
\frac{\partial \log W(\boldsymbol{\phi})}{\partial b} &= \langle \sum_{i\sim j} \mathbb{1}_{Y_i=Y_j}\mathbb{1}_{Z_i=Z_j} \rangle = \langle N(\mathbf{Y},\mathbf{Z}) \rangle \\
\frac{\partial \log W(\boldsymbol{\phi})}{\partial c} &= \langle \mathbb{1}_{Z_i=Z_j} \rangle = \langle N(\mathbf{Z}) \rangle \\
\frac{\partial^2 \log W(\boldsymbol{\phi})}{\partial b^2} &= Var(N(\mathbf{Y},\mathbf{Z})) \\
\frac{\partial^2 \log W(\boldsymbol{\phi})}{\partial c^2} &= Var(N(\mathbf{Z})) \\
\frac{\partial^2 \log W(\boldsymbol{\phi})}{\partial b \partial c} &= Cov(N(\mathbf{Y},\mathbf{Z}), N(\mathbf{Z}))
\end{aligned}$$

où $N(\mathbf{y},\mathbf{z})$ désigne le nombre de couples i et j voisins pour lesquels $y_i = y_j$ et $z_i = z_j$ et $N(\mathbf{z})$ le nombre de couples i et j voisins pour lesquels $z_i = z_j$.

En Figure VI.1 on a représenté la superposition des composantes de la hessienne pour b (en ordonnée) et c (en abscisse) variant entre -2 et 2. Les zones non bleues correspondent

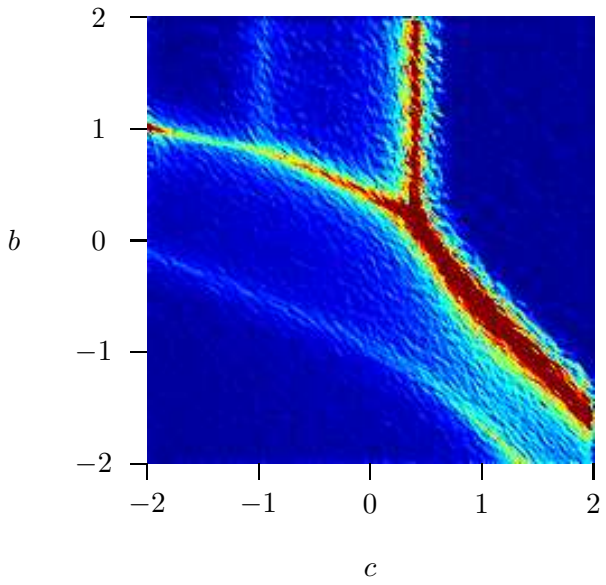


FIG. VI.1 – Image représentant les valeurs de b (en ordonnée) et c (en abscisse) de transition de phase.

aux valeurs de b et c correspondant à une transition de phase. On peut y observer une courbe en "Y" principale le long de laquelle $\nabla^2_\phi W(\boldsymbol{\phi})$ est très piqué. On y voit également une autre branche le long de $c \approx -1$, entre $b \approx 0.8$ et $b \approx 2$, et une courbe reliant les points $(c, b) \approx (1.3, -2)$ et $(c, b) \approx (-2, -0.1)$. L'ensemble de ces courbes définissent dont 5 régions de l'espace.

En Figure VI.2, on peut voir des simulations du champ de Markov $(\mathbf{Y}, \mathbf{Z})$ défini par (VI.2) sur une image 128×128 et un voisinage d'ordre 2, pour différentes valeurs de b (en ordonnée) et c (en abscisse). Chacune des $LK = 4$ valeurs possible du couple (y_i, z_i) est associée à un niveau de gris. La Figure VI.3 donne les réalisations du champ $\mathbf{Z}$ correspondant. Enfin, on montre en Figure VI.4 des réalisations des données $\mathbf{X}$ associées aux images de la Figure VI.2 lorsque, dans l'équation (VI.1), les distributions $f(.|\theta_{lk})$ sont des gaussiennes $\mathcal{N}(\mu_{lk}, \sigma^2)$ en dimension 1, de moyenne $\mu_{lk} = (k-1)L + l$ et d'écart-type $\sigma = 0.3$.

2 Application à des données réelles : reconnaissance de textures

Une étude préliminaire à la reconnaissance de textures peut être trouvée dans [15]. Les résulats présentés dans cette section font l'objet d'un rapport de recherche, soumis pour publication [16].

2.1 Les données

L'objectif est la classification supervisée (ou reconnaissance) d'images de textures réelles. La base de données dont nous disposons est constituée de 140 images unitextures, et 63 images multitextures. Les images ont été prises selon des angles de vue différents, ainsi que sous des luminosités et à des échelles variables. La base de données comporte

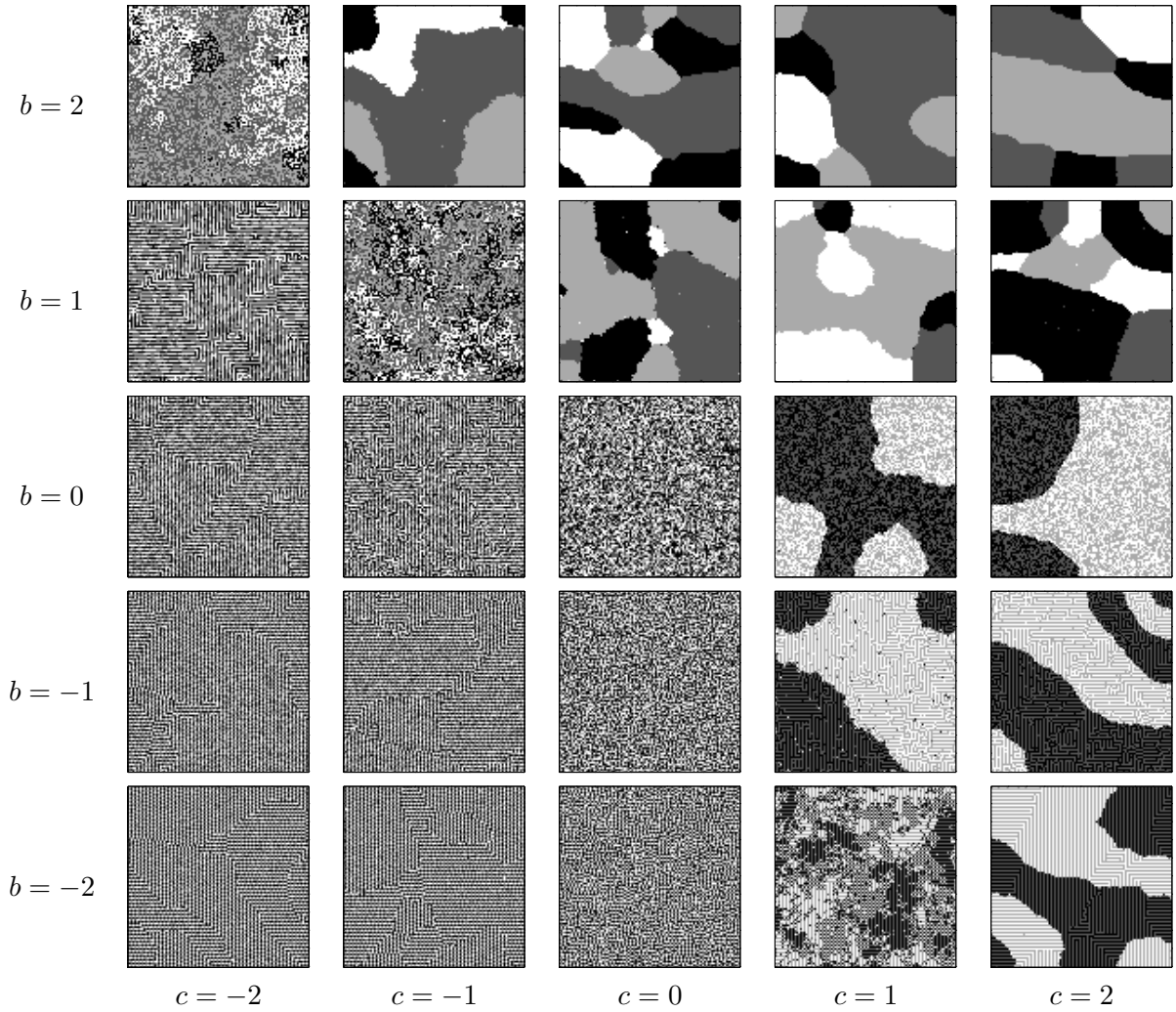


FIG. VI.2 – Simulation du champ de Markov $(\mathbf{Y}, \mathbf{Z})$ défini par (VI.2) pour différentes valeurs de b et c, sur une image 128×128 et un voisinage d'ordre 2, avec $K = 2$ et $L = 2$.

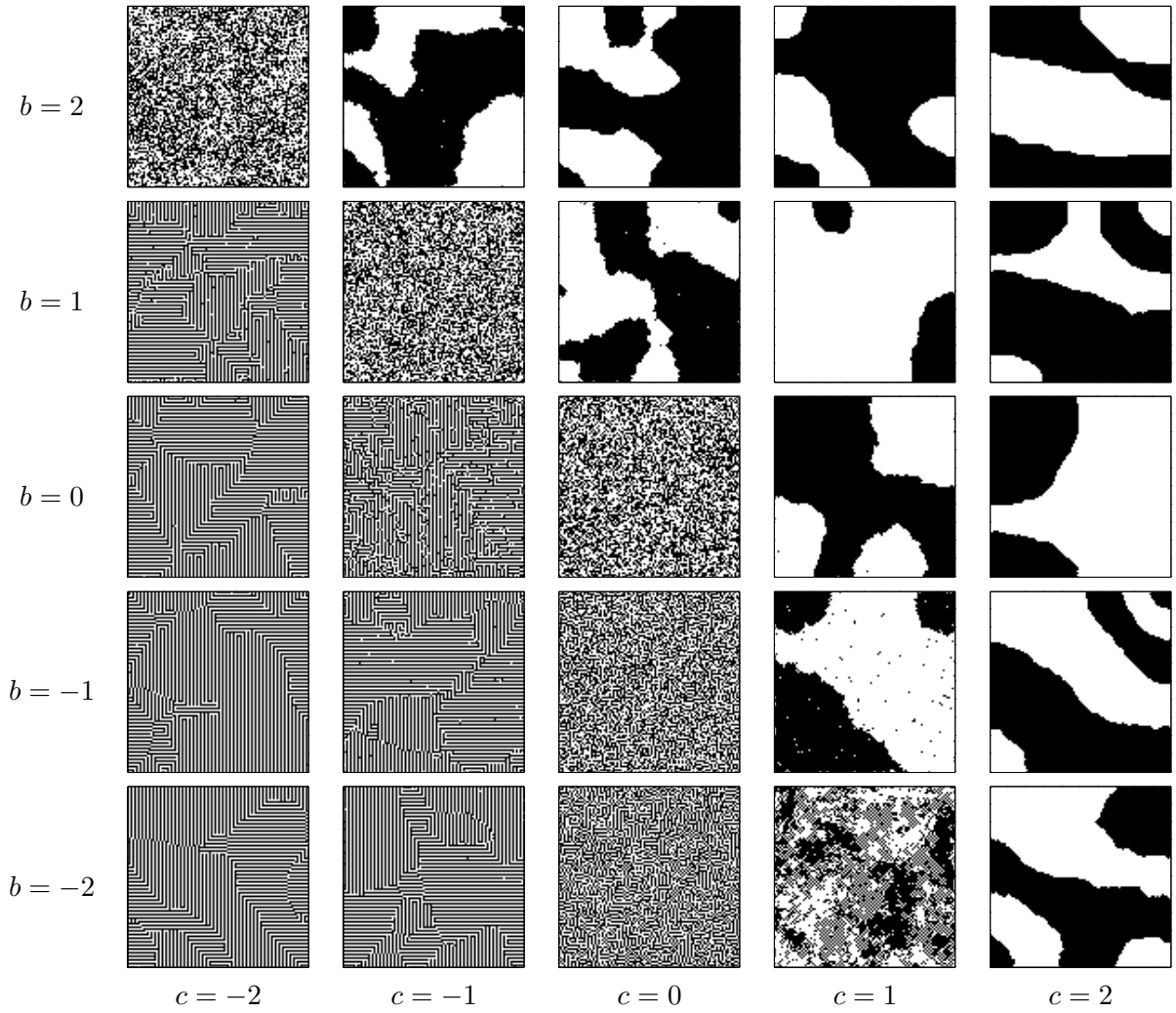


FIG. VI.3 – Réalisations du champ $\mathbf{Z}$ correspondant à la Figure VI.2 pour différentes valeurs de b et c, sur une image 128×128 et un voisinage d'ordre 2, avec $K = 2$ et $L = 2$.

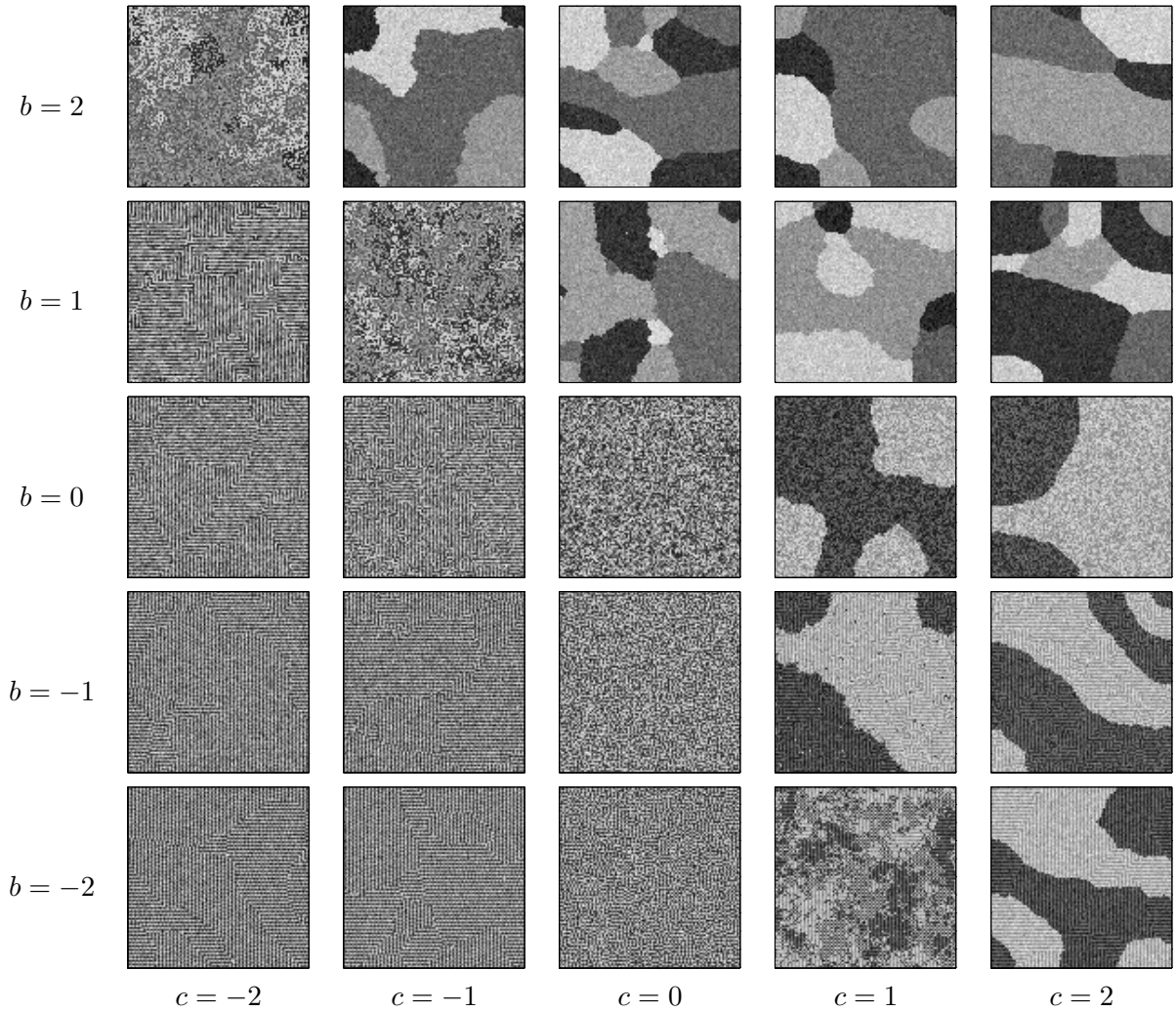


FIG. VI.4 – Réalisations des données $\mathbf{X}$ définies par (VI.1) pour les réalisations $(\mathbf{y}, \mathbf{z})$ de la Figure VI.2, lorsque les distributions $f(.|\theta_{lk})$ sont des gaussiennes $\mathcal{N}(\mu_{lk}, \sigma^2)$ en dimension 1, de moyenne $\mu_{lk} = (k-1)L + l$ et d'écart-type $\sigma = 0.3$. Les réalisations sont obtenues pour différentes valeurs de b et c, sur une image 128×128 et un voisinage d'ordre 2, avec $K = 2$ et $L = 2$.

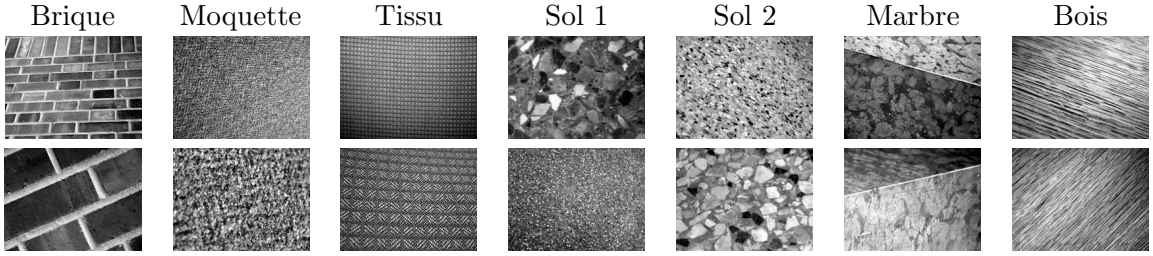


FIG. VI.5 – Echantillon des 7 textures de la base d'aprrentissage

$K = 7$ textures réelles différentes dont un échantillon est visible en figure VI.5. Pour chacune des 7 textures, il y a 20 images unitextures dont 10 sont gardées pour constituer la base d'apprentissage. Les autres images, à savoir les 10×7 autres images unitextures ainsi que les 63 images multitextures consituent la base de test.

Les niveaux de gris d'une image peuvent difficilement tenir compte des changements d'angles, d'échelle ou de luminosité. Aussi nos images sont-elles décrites par des *descripteurs locaux* invariants par transformation affine, ainsi que par leurs relations spatiales.

Descripteurs d'image. L'extraction de ces descripteurs d'image comporte deux phases : la *détection* de points d'intérêt et la *description* de la zone de l'image située autour de chacun de ces points d'intérêt. La détection des points d'intérêt est réalisée grâce à un opérateur de détection qui parcourt l'image et identifie les zones de l'image ayant des caractéristiques particulières. Par exemple, le détecteur de Harris-Laplace [93] recherche les zones de fort gradient dans toutes les directions et détecte ainsi principalment les angles et bords de l'image. Citons encore, parmi les plus utilisés, le détecteur DoG (Difference of Gaussian) [87] qui extrait des zones homogènes (*blobs*) de l'image et le détecteur de Laplace [85] détectant les régions les plus saillantes. A chacun des points d'intérêt détectés est également associée une échelle caractéristique qui servira à déterminer la zone autour du point sur laquelle calculer le descripteur, permettant de ce fait l'invariance par transformation affine. Le descripteur SIFT [87] par exemple divise chaque zone autour du point d'intérêt courant en 4×4 zones et calcule le gradient dans les 8 directions de l'image à l'intérieur de ces 16 zones. Ainsi, chaque zone autour d'un point d'intérêt est décrite par un vecteur de taille $128 = 4 \times 4 \times 8$. Au terme de ce procédé d'extraction, à une image sont associés n descripteurs locaux irrégulièrement espacés, en dimension D ($D = 128$ pour SIFT). Pour l'étape d'extraction et description de nos images, nous suivons la méthode décrite dans [82] pour ses avantages sur les autres méthodes. Le détecteur utilisé est le détecteur de Laplace [85] et le descripteur SIFT [87].

Structure de voisinage. Concernant la structure de voisinage associée à ces descripteurs, nous avons choisi d'utiliser le graphe de Delaunay. D'autres graphes auraient pu être construits, notamment par exemple en se basant sur les zones autour des points d'intérêt comme dans [82], ou en construisant les graphes décrits au chapitre I, section 1.1. Néanmoins, le graphe de Delaunay semble permettre une meilleure description de nos images (voir Figure VI.6).

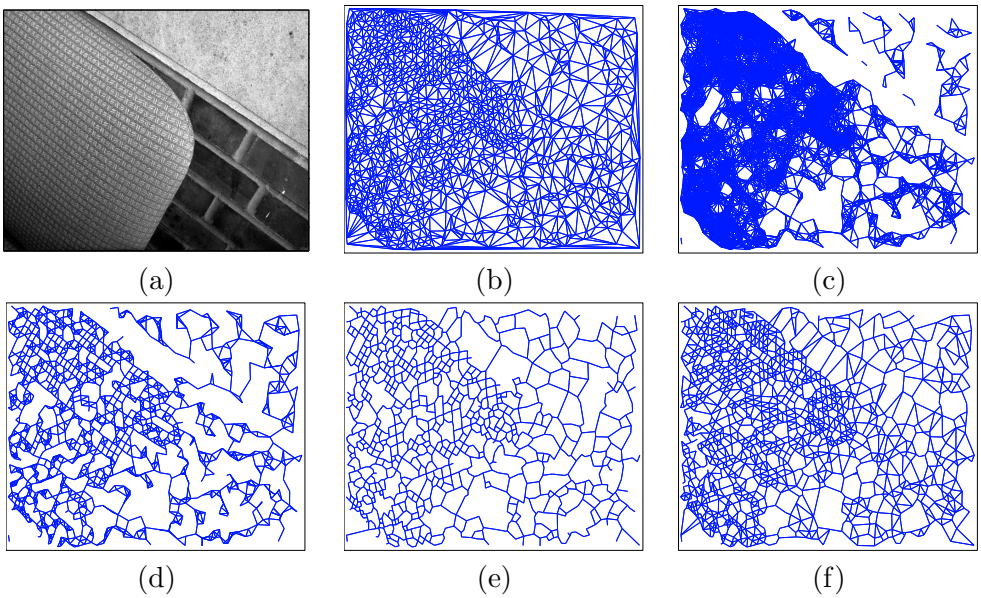


FIG. VI.6 – (a) Image multitexture. (b) Graphe de Delaunay associé aux points d'intérêt détectés. (c) ϵ-graphe ($\varepsilon = 30$). (d) Graphe des k-voisins réciproques ($k = 5$). (e) Graphe de voisinage relatif. (f) Graphe de Gabriel.

2.2 Sélection de modèle

Notre modèle suppose que les descripteurs sont des variables aléatoires $(\mathbf{X}_i)_{i\in\mathcal{I}}$ situés aux points d'intérêt $i \in \mathcal{I}$. Il est clair que les textures sont des classes complexes, ne pouvant donc être modélisées par des distributions unimodales (une seule gaussienne par texture par exemple). De plus, étant par nature composées de motifs qui se répètent, elles présentent, même visuellement, une certaine structure. Dès lors, l'hypothèse de bruit indépendant apparaît trop restrictive puisqu'elle revient à supposer qu'au sein d'une texture, les observations sont indépendantes. Nous proposons donc d'utiliser le modèle de champ de Markov triplet décrit au chapitre V. Le nombre de sous-classes L_k de chacune des $K = 7$ textures est supposé être fixe, égal à $L = 10$. Sélectionner L_k pour chaque texture k individuellement est possible, quoique laborieux (voir chapitre V, section 4.1), mais nous n'avons pas observé d'amélioration significative sur le modèle à matrices $\mathbb{B}_{kk}$ pleines. De même, nous avons choisi de supposer $\mathbb{C} = c\mathbb{I}_K$ (où $\mathbb{I}_K$ désigne la matrice identité de taille $K \times K$), qui correspond au terme de l'énergie d'un modèle de Potts et favorise les régions homogènes (c'est-à-dire les régions de même texture), ce qui est le cas dans les images réelles. Les lois $f(.|\theta_{lk})$ pour $k = 1, \cdots, 7$ et $l = 1, \cdots, 10$ sont supposées gaussiennes, paramétrées par un vecteur moyenne μ_{lk} de dimension $D = 128$ et par une matrice de covariance Σ_{lk} de dimension $D \times D$. Reste donc à sélectionner la forme des matrices Σ_{lk} des gaussiennes et des matrices $\mathbb{B}_{kk'}$ des lois $P_G(\mathbf{y}, \mathbf{z})$, pour $l = 1, \cdots, 10$ et $k, k' = 1, \cdots, 7$.

Sélection des Σ_{lk}. Les descripteurs étant en dimension $D = 128$, l'estimation d'une matrice de covariance Σ_{lk} pleine nécessiterait l'estimation de $\frac{1}{2}D(D+1) = 8256$ paramètres. Une paramétrisation de ces matrices de covariance est donc nécessaire. Nous avons choisi

$\mathbb{B}_{kk}$	Brique	Moquette	Tissu	Sol 1	Sol 2	Marbre	Bois
$[b_k]+[0]$	1739730	2403890	3141300	2556280	3147610	2967630	2092090
$[b_k^{ll}]+[0]$	1739840	2403970	3141440	2556410	**3147900**	2967770	2092030
$[b_k]+[a_k^{ll'}]$	1740010	2404450	3141930	2556630	3145570	**2968290**	1946380
$[b_k^{ll}]+[a_k^{ll'}]$	**1740080**	**2404600**	**3142170**	**2556730**	3145600	2968280	**2092760**

TAB. VI.1 – Valeurs du critère BIC^w pour chaque texture lorsque les matrices de covariance Σ_{lk} sont diagonales. En gras, le modèle $\mathbb{B}_{kk}$ sélectionné pour chaque texture par notre critère BIC^w.

de comparer deux possibilités : le cas d'une matrice de covariance diagonale et celui d'une matrice paramétrée pour la grande dimension comme décrit au chapitre III, section 2 [23].

Sélection des $\mathbb{B}_{kk'}$. Concernant les matrices $\mathbb{B}_{kk'}$, $k \neq k'$, la nature même des données d'apprentissage (des images unitextures) fait qu'il est impossible d'estimer ces matrices. En effet, il ne peut y avoir de sites voisins dans les classes k et k' (les descripteurs d'apprentissage de la classes k proviennent d'image unitextures, donc ne contenant que des descripteur de la classe k). Nous avons donc fixé les matrices $\mathbb{B}_{kk'}$ à 0, ce qui est consistant avec le fait que nous cherchons à retrouver des régions homogènes de la même texture. Une autre alternative serait de retarder leur estimation à l'étape de classification mais les images test considérées ne sont composées que de quelques unes des 7 textures si bien que la plupart des $\mathbb{B}_{kk'}$ ne pourraient être estimées. Concernant les matrices $\mathbb{B}_{kk}$, nous avons considéré, pour chaque k, les 4 formes possibles décrites au chapitre V, section 4.2, à savoir les modèles diagonaux $[b_k]+[0]$ et $[b_k^{ll}]+[0]$, ou pleins $[b_k]+[a_k^{ll'}]$ et $[b_k^{ll}]+[a_k^{ll'}]$. Encore une fois, l'estimation de ces matrices $\mathbb{B}_{kk}$ aurait pu être réeffectuée à l'étape de classification (seules les distributions $f(.|\theta_{lk})$ sont nécessaires pour résoudre le problème d'identifiabilité) mais, pour chaque image test, il faudrait alors estimer K matrices de dimension $L \times L$. Etant donné le nombre de descripteurs de chaque image (de quelques centaines à quelques milliers), l'estimation ne pourrait vraisemblablement être effectuée que pour un des modèles très simples, comme le modèle $[b_k]+[0]$ ou à la rigueur $[b_k^{ll}]+[0]$ et réduirait ainsi grandement la flexibilité de notre modèle.

Résultats. Nous avons donc calculé les approximations en champ moyen BIC^p (chapitre II, section 5.2.1) et BIC^w (chapitre II, section 5.2.2) du critère BIC pour chacun des modèles présentés précédemment. La Table VI.1 rapporte les valeurs de BIC^w pour chaque texture dans le cas où les matrices Σ_{lk} sont diagonales et la Table VI.2 dans le cas où elles sont paramétrées pour la grande dimension. Nous ne reportons pas les valeurs BIC^p qui sont presque équivalentes aux BIC^w.

Il apparaît que, en terme de BIC, les modèles pour données de grande dimension sont à préférer aux modèles diagonaux, et ce quel que soit le modèle pour les matrices $\mathbb{B}_{kk}$. Notons de plus que le modèle sélectionné pour les matrices $\mathbb{B}_{kk}$ dépend de la texture. Pour la texture "Bois", le modèle sélectionné est le plus simple ($[b_k]+[0]$) alors que pour les textures "Sol 2" et "Marbre", le modèle le plus compliqué est sélectionné ($[b_k^{ll}]+[a_k^{ll'}]$).

$\mathbb{B}_{kk}$	Brique	Moquette	Tissu	Sol 1	Sol 2	Marbre	Bois
$[b_k]$	1902700	2515860	3516630	2696700	3290280	3172210	**2263160**
$[b_k^{ll}]$	1882040	**2524590**	**3529860**	**2697180**	3286890	3172450	2260870
$[b_k] + [a_k^{ll'}]$	**1905800**	2518320	3521420	2692730	3292960	3177140	2260650
$[b_k^{ll}] + [a_k^{ll'}]$	1876510	2518430	3495890	2691310	**3293230**	**3178150**	2262300

TAB. VI.2 – Valeurs du critère BIC^w pour chaque texture lorsque les matrices de covariance Σ_{lk} sont paramétrées pour la grande dimension. En gras, le modèle $\mathbb{B}_{kk}$ sélectionné pour chaque texture par notre critère BIC^w.

2.3 Résultats de classification

La Table VI.3 donne les taux de classification correcte τ_k de chaque texture k :

$$\tau_k = \frac{\text{nombre de descripteurs de la texture } k \text{ classés dans la texture } k}{\text{nombre de descripteurs de la texture } k}$$

Notons que ces résultats ont été obtenus sur des images unitextures uniquement. Nous y comparons les performance de plusieurs modèles :

– Les lignes "Mélange de mélange" correspondent à une hypothèse d'indépendance des descripteurs, et à une modélisation par mélange de mélange indépendant gaussien, comme utilisé dans [68]. Sous un tel modèle, la distribution du triplet $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ est de la forme :

$$P(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \prod_{i \in \mathcal{I}} \pi_{z_i} \lambda_{y_i z_i} f(x_i | \theta_{y_i z_i}) \tag{VI.3}$$

avec $\sum_{k \in \mathcal{K}} \pi_k = 1$ et, pour tout $k \in \mathcal{K}$, $\sum_{l \in \mathcal{L}} \lambda_{lk} = 1$. L'algorithme EM est utilisé pour l'estimation des paramètres. L'apprentissage consiste à estimer, pour chaque texture $k \in \mathcal{K}$, les proportions $(\lambda_{lk})_{l \in \mathcal{L}}$ et les paramètres $(\theta_{lk})_{l \in \mathcal{L}}$ des gaussiennes $f(.|\theta_{lk})$ sur les données d'apprentissage de la texture k. Lors de la phase de test, on estime les proportions $(\pi_k)_{k \in \mathcal{K}}$ des $K = 7$ textures.

– Les lignes "TMF" se réfèrent à notre méthode par champ de Markov triplet (équation V.1) lorsque le modèle le plus général pour $\mathbb{B}_{kk}$ ($[b_k^{ll}] + [a_k^{ll'}]$) est utilisé pour chaque texture :

$$P(\mathbf{x}, \mathbf{y}, \mathbf{z}) \propto \exp\Big(\sum_{i \sim j} \mathbb{B}^{y_i y_j}_{z_i z_j} 1_{z_i = z_j} + c \sum_{i \sim j} 1_{z_i = z_j} + \sum_i \log f(x_i | \theta_{y_i z_i})\Big) \tag{VI.4}$$

Ce modèle est le pendant markovien du modèle de mélange de mélange indépendant (équation VI.3).

– Les lignes "Mélange" se réfèrent au modèle de mélange gaussien classique, sous lequel le couple $(\mathbf{X}, \mathbf{Z})$ a pour distribution :

$$P(\mathbf{x}, \mathbf{z}) = \prod_{i \in \mathcal{I}} \pi_{z_i} f(x_i | \theta_{z_i})$$

où $\sum_{k \in \mathcal{K}} \pi_k = 1$ et $f(.|\theta_k)$ sont K distributions gaussiennes. Lors de l'étape d'apprentissage pour la texture k, on estime les paramètres de la gaussienne $f(.|\theta_k)$ par

maximum de vraisemblance. Lors de la phase de test, les K gaussiennes ont été apprises et on estime les K proportions π_k des différentes textures. Notons que ce modèle est équivalent au modèle de mélange de mélange (VI.3) avec $L = 1$.

– Les lignes "HMF" correspondent au modèle de champ de Markov caché (modèle de Potts) classique, sous lequel le couple $(\mathbf{X}, \mathbf{Z})$ a pour distribution :

$$P(\mathbf{x}, \mathbf{z}) \quad \propto \quad \exp(c \sum_{i \sim j} 1_{z_i = z_j} + \sum_i \log f(x_i | \theta_{z_i}))$$

L'étape d'apprentissage de la texture k est alors exactement identique au modèle de mélange indépendant. Lors de la phase de test, seul le paramètre de régularité c reste à estimer. Notons que ce modèle est équivalent au modèle par champ de Markov triplet (VI.4) avec $L = 1$.

Pour chacun de ces quatre modèles, on considère deux alternatives pour la forme de la matrice de covariance des distributions gaussiennes : diagonale ou paramétrée pour la grande dimension (voir chapitre III, section 2). Enfin, la ligne "TMF+BIC" se réfère au cas où la forme de ces matrices $\mathbb{B}_{kk}$ ainsi que celle des covariances Σ_{lk} ont été sélectionnées par le critère BIC (Tables VI.1 et VI.2).

Les résultats de la Table VI.3 montrent que le taux de classification correcte est légèrement amélioré entre le modèle de mélange indépendant et le modèle de champ de Markov caché, et ce quelle que soit la forme des matrices de covariance. Néanmoins les taux de reconnaissance restent très faibles et les améliorations, à modèle commun, peu significatives (de l'ordre de 3%). Rappelons que, pour les modèles de mélange indépendant et de champ de Markov caché, l'apprentissage est identique et seule diffère l'étape de test. Le modèle markovien permet alors d'obtenir des classifications plus lisses mais, le modèle étant très mal adapté aux données, la classification reste très "bruitée". De manière analogue, les performances du modèle de mélange de mélange indépendant sont inférieures à celle de notre modèle par champ de Markov triplet pour lequel les taux de reconnaissance sont tout à fait satisfaisants. Mentionnons encore que, de manière générale, utiliser un modèle triplet (équations VI.3 et VI.4) permet d'améliorer significativement les taux de reconnaissance. En effet, par l'ajoût d'un champ auxillaire $\mathbf{Y}$ jouant le rôle de sous-classes, une texture n'est plus modélisée par une unique distribution gaussienne, donc unimodale, mais par un mélange de 10 distributions gaussiennes. Une telle loi s'avère être mieux adaptée au caractère probablement multimodal de la vraie distribution. Enfin, remarquons que pour chacun des modèles étudiés, utiliser une paramétrisation des matrices de covariance adaptée aux données de grande dimension améliore les résultats. Un tel modèle permet en effet de tenir compte des dépendances entre les variables d'un descripteur, tout en ne nécessitant l'estimation que d'un nombre limité de paramètres. Les dépendances entre les 128 variables de nos descripteurs proviennent de la façon même dont ils ont été calculés. En effet, avec le descripteur SIFT, chaque variable correspond à un gradient dans une direction. Par exemple si le gradient dans la direction "Nord" est très forte, celui dans la direction "Sud" sera vraisemblablement faible. Notons que les résultats avec le modèle par champ de Markov triplet et Σ_{lk} pour la grande dimension donne de très bons résultats (98% et plus) sur toutes les textures excepté sur la texture "Marbre". En effet pour cette texture, les images sont très hétérogènes en terme de luminosité. Sur

Modèle	Brique	Moquette	Tissu	Sol 1	Sol 2	Marbre	Bois
Mélange Diagonal	34.08	27.63	43.70	27.41	33.80	26.27	29.78
Mélange Grande Dimension	42.12	35.11	52.05	29.37	46.42	28.44	31.06
HMF Diagonal	36.03	29.96	43.80	31.14	39.58	29.15	32.48
HMF Grande Dimension	42.46	35.65	52.65	33.06	48.34	29.83	34.91
Mélange de mélange Diagonal	77.58	31.60	58.26	28.26	58.79	33.87	58.56
Mélange de mélange Grande Dimension	81.18	56.94	62.48	35.64	67.43	37.05	65.02
TMF Diagonal	96.59	80.70	83.60	82.69	83.90	46.05	95.18
TMF Grande Dimension	99.33	98.61	99.28	97.36	99.57	56.24	99.28
TMF-BIC	**99.37**	**98.71**	**99.30**	**98.16**	**99.62**	**56.77**	**99.52**

TAB. VI.3 – Pourcentage de descripteurs correctement classés sur les images test unitextures. Les lignes correspondent à différents modèles. Les chiffres en gras correspondent aux meilleurs taux pour chaque texture. Les termes "Diagonal" et "Grande Dimension" se réfèrent à la forme de la matrice de covariance Σ_{lk} considérée, respectivement diagonale et paramétrée pour la grande dimension.

une même image, une partie peut être presque totalement noire car dans l'ombre et une autre, presque totalement blanche de lumière. La mauvaise qualité des descripteurs qui ne peuvent faire face à de telles variations, empêche de bien apprendre le modèle de la texture "Marbre". Enfin, remarquons qu'utiliser le modèle de champ de Markov triplet sur le modèle sélectionné améliore légèrement les résultats (de l'ordre de 0.3%).

Sur les images multitextures, des améliorations sont également observées sur toutes les images entre les modèles indépendants et markoviens. De plus, les modèles triplets donnent de bien meilleurs résultats que les modèles cachés classiques. Les taux de reconnaissance augmentent d'environ 53% en moyenne entre le modèle de mélange indépendant à matrice de covariance diagonale et le modèle par champ de Markov triplet sélectionné par BIC. Une illustration est donnée en Figure VI.7. Parmi les modèles non triplets, nous ne reportons que la classification correspondant au modèle de champ de Markov caché avec matrice de covariance paramétrée pour la grande dimension. Les autres modèles non triplets ont des taux de reconnaissance légèrement plus faibles mais l'allure générale de la classification reste très proche.

Les taux relatifs à notre TMF-BIC sont tous supérieurs à 90%. Il arrive que, pour très peu d'images, le modèle TMF avec le modèle le plus complet pour $\mathbb{B}_{kk}$ ($[b_k^{ll}]$) donne de meilleurs taux de classification (de 1 à 2%). La Figure VI.8 illustre un tel cas, alors que la Figure VI.9 illustre le cas le plus général où la sélection de modèle donne des résultats sensiblement meilleurs.

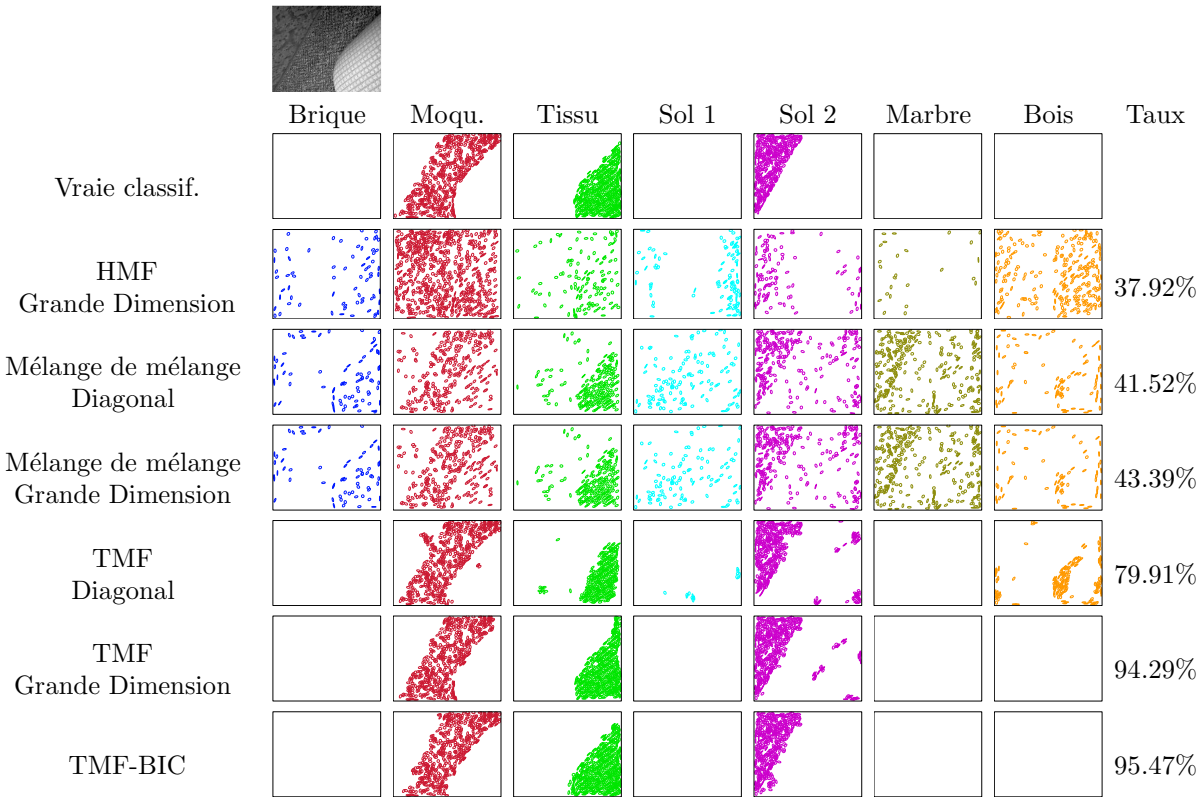


FIG. VI.7 – Image composée de 3 textures (Moquette, Tissu, Sol2) : la première ligne est la vraie classification, les suivantes correspondent aux différents modèles. Les colonnes montrent les points d'intérêt classés dans chacune des 7 textures. La dernière colonne donne le taux de classification correcte.

Comme mentionné précédemment (Table VI.3), la texture "Marbre" souffre de faibles taux de reconnaissance du fait de la nature des images de cette texture (forts changements de luminosité) qui rendent l'apprentissage d'un modèle pour cette texture très difficile. La Figure VI.10 montre le comportement du modèle sur une image composée de 3 textures, dont le marbre. Le taux de reconnaissance global est supérieur à 90% mais la plus grande partie des erreurs proviennent des points de la texture "Marbre" (seuls 50% des descripteurs du "Marbre" sont bien classés). Notons de plus que, du fait du caractère spatial de notre modèle, l'algorithme a tendance à classer les descripteurs de la texture "Marbre" dans les textures avoisinantes, à savoir "Moquette" et "Sol 2". Les autres algorithmes, dont nous ne reportons pas les résultats ici, ont un comportement similaire, mais le taux de classification correcte est bien plus faible.

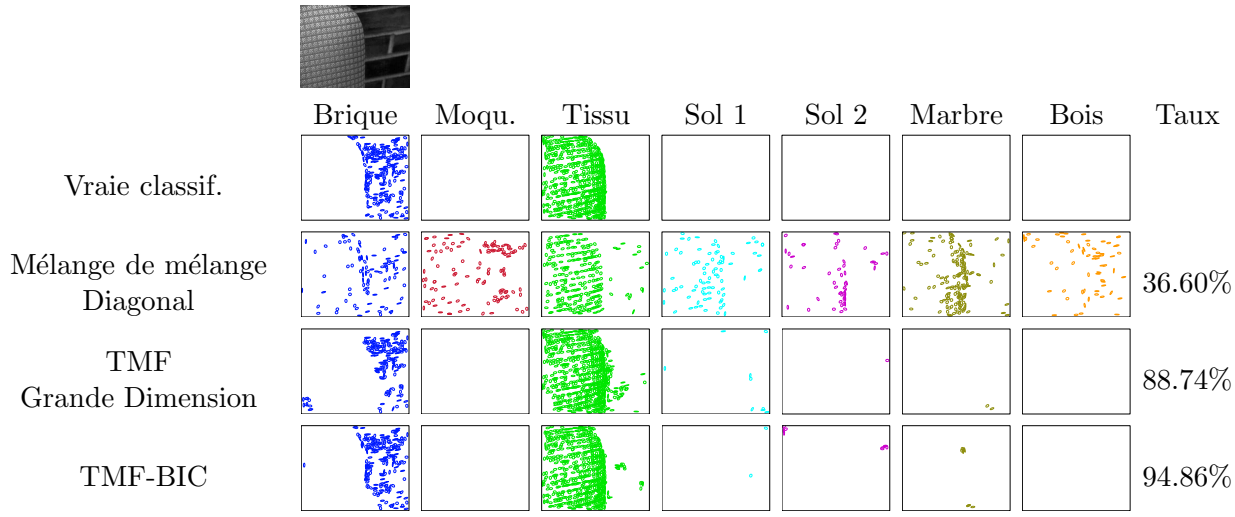


FIG. VI.8 – Image composée de deux textures (Tissu et Brique) : la première ligne est la vraie classification, les suivantes correspondent respectivement au modèle de mélange indépendant à matrice de covariance diagonale, au champ de Markov triplet à matrice de covariance pour la grande dimension et $\mathbb{B}_{kk}$ non contraints et au champ de Markov triplet avec le modèle sélectionné par BIC. Les colonnes montrent les points d'intérêt classés dans chacune des 7 textures. La dernière colonne donne le taux de classification correcte.

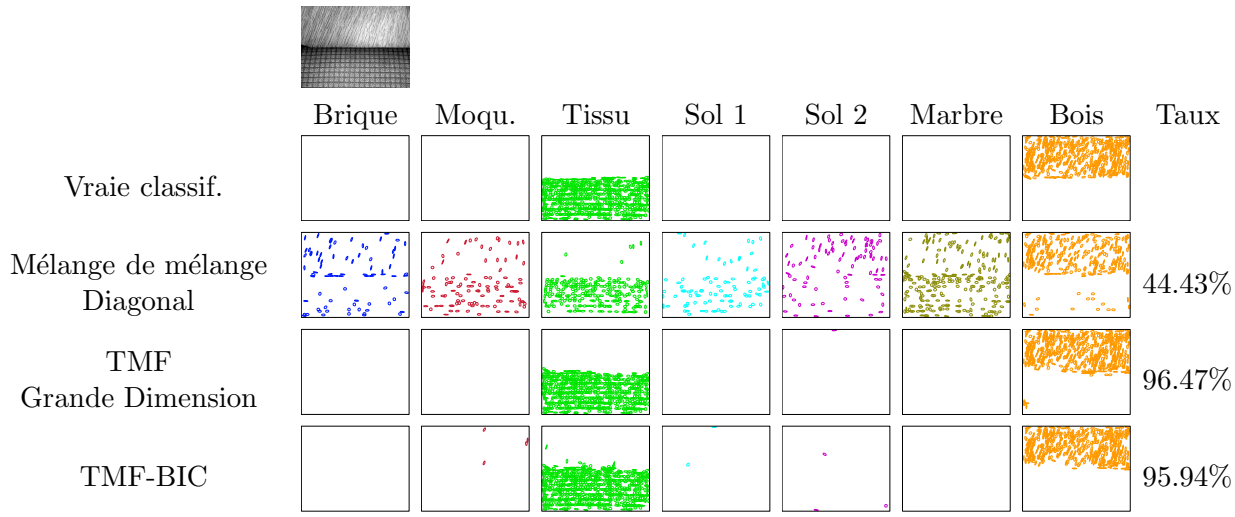


FIG. VI.9 – Image composée de deux textures (Tissu et Bois) : la première ligne est la vraie classification, les suivantes correspondent respectivement au modèle de mélange indépendant à matrice de covariance diagonale, au champ de Markov triplet à matrice de covariance pour la grande dimension et $\mathbb{B}_{kk}$ non contraints et au champ de Markov triplet avec le modèle sélectionné par BIC. Les colonnes montrent les points d'intérêt classés dans chacune des 7 textures. La dernière colonne donne le taux de classification correcte.

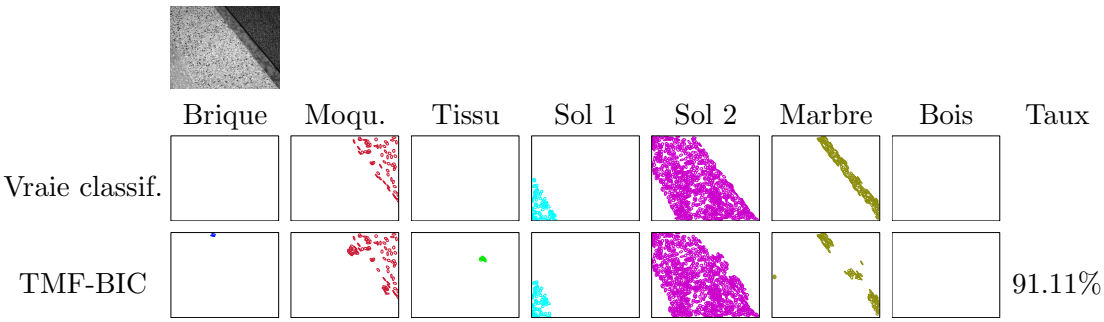


FIG. VI.10 – Image composée de 3 textures, dont du marbre : la première ligne est la vraie classification, la seconde correspond au champ de Markov triplet avec le modèle sélectionné par BIC. Les colonnes montrent les points d'intérêt classés dans chacune des 7 textures. La dernière colonne donne le taux de classification correcte.

Partie C

Classification d'individus avec observations incomplètes

Chapitre VII

Modèles et méthodes pour données incomplètes

1 Observations incomplètes

Lorsqu'on analyse un problème réel, on est souvent confronté à des données manquantes, c'est-à-dire à une population d'individus qui ne sont pas tous observés ou pas tous observés de la même manière. Les raisons pour lesquelles ces valeurs manquent peuvent être très diverses. Dans le cas des sondages, certains sondés peuvent choisir de ne pas répondre à certaines questions (voir l'exemple de l'introduction de [108]). Le vide peut également être accidentel en cas de défaillance d'un appareil de mesure. Dans d'autres cas les mesures ne seront pas effectuées volontairement. Elles pourront être jugées peu révélatrices ou trop coûteuses par l'expérimentateur au vue de résultats antérieurs.

Un premier exemple est celui des images satellitaires, comme la très coûteuse mission *Mars Express* de la NASA. Les images transmises par le satellite sont des images hyperspectrales, consituées de spectre en dimension 256. Les conditions atmosphériques très rudes ou un dysfonctionnement de l'appareil de mesure font que, sur certaines images, certains spectres sont manquants.

Un second exemple est celui des données d'expression de gènes issues de puces ADN. Plusieurs raisons conduisent les expériences à ne pas fournir des données complètes : des poussières ou éraflures sur la lame ([123, 18]), une erreur systématique du robot qui dépose les sondes, une image à analyser corrompue, une résolution non adaptée qui peut engendrer des problèmes pour identifier les gènes sur la lame [90],...

1.1 Notations

Comme précédemment, on note $\mathbf{x} = (x_1, \cdots, x_n)$ les données en chacun des n sites ($\forall i \in \mathcal{I}$, $x_i \in \mathbb{R}^D$). On note encore $o_i \subset [\![1, D]\!]$ l'ensemble des indices des composantes observées x_{id} pour le site i et m_i celui des indices des composantes manquantes (on a donc $o_i \cup m_i = [\![1, D]\!]$). Nous écrirons encore :

$$x_i^{o_i} = \{x_{id}, d \in o_i\} \quad x_i^{m_i} = \{x_{id}, d \in m_i\}$$
$$\mathbf{x^o} = \{x_i^{o_i}, i \in \mathcal{I}\} \quad \mathbf{x^m} = \{x_i^{m_i}, i \in \mathcal{I}\}$$

Chaque observation x_i peut n'avoir aucune valeur manquante ($o_i = [\![1, D]\!]$ et $m_i = \varnothing$),

comme aucune valeur observée ($o_i = \varnothing$ et $m_i = [\![1, D]\!]$), ou toute autre possibilité entre ces deux cas extrêmes.

1.2 Traitements heuristiques des données manquantes

Suppression La technique la plus souvent employée en présence de données manquantes consiste à supprimer les vecteurs d'observations incomplètes (*case deletion*), de façon à ne travailler qu'avec des données complètes. Cette technique simple peut se justifier lorsque peu de données sont manquantes (typiquement moins de 5% [111]). Cependant, il est courant qu'une proportion notable d'observations soient incomplètes. A titre d'exemple, les données manquantes affectent en général de l'ordre de 90% des gènes présents dans une expérience [96]. De plus, dans le cadre de données dépendantes, cette technique conduit à briser certains arcs dans le graphe de dépendance, et donc à déconnecter artificiellement des individus interconnectés.

Imputation simple Une autre technique heuristique permettant de ne pas supprimer d'information consiste à remplacer les données manquantes par une valeur jugée "raisonnable" (*single imputation*) et à travailler ensuite sur ces données complétées comme si aucune valeur n'était manquante. La méthode *cold-deck* remplace les données manquantes par une valeur issue d'une source externe, la réponse à un précédent questionnaire par exemple dans le cas des sondages. La méthode *hot-deck* utilise comme valeur de remplacement celle observée en un site "similaire". Une extension de cette méthode consiste à remplacer une donnée manquante x_{id} par la moyenne de la d-ième composante des k-plus proches observations(*k-nearest neighbors imputation*). Une autre technique très répandue est de remplacer chaque valeur manquante x_{id} par la moyenne $\langle \mathbf{x}_{.d} \rangle$ des données observées selon cette variable d (*mean imputation*). En classification, la procédure équivalente est de remplacer la valeur manquante x_{id} par l'estimation courante de la moyenne $\langle \mathbf{x}_{C_k d} \rangle$ de la variable d à l'intérieur de la classe C_k à laquelle est affecté le site i. Ces types de remplacement permettent de préserver la moyenne empirique de l'échantillon complété, mais ont pour inconvénient de biaiser la variance empirique vers 0. Une autre méthode consiste à prédire les données manquantes par régression linéaire sur les données observées (*regression imputation*) mais une telle approche augmente la corrélation entre les données. De manière générale, toutes ces méthodes d'imputation simple ont le désavantage de ne pas tenir compte de l'incertitude quant à la prédiction des données manquantes, ni de tenir compte du fait que l'absence de l'observation peut elle-même être informative (comme c'est le cas des données censurées par exemple).

Imputation multiple Plutôt que de remplacer chaque valeur manquante par une unique valeur, on peut la remplacer par un ensemble de valeurs (*multiple imputation*) [107], représentant l'incertitude quant à la vraie valeur à utiliser. La procédure se fait alors en trois étapes :

1. Remplacer chaque donnée manquante par q valeurs possibles, de manière à obtenir q jeux de données complets
2. Analyser ces q jeux de données par une procédure standard
3. Combiner les résultats obtenus sur les q jeux complets

Plus le nombre q sera grand, meilleur sera le résultat. Néanmoins en pratique, on observe qu'un petit nombre d'imputations est suffisant (entre 3 et 10). En effet (voir [107], p117), l'efficacité (*efficiency*) d'un estimateur basé sur q imputations est de l'ordre de $(1 + \frac{\gamma}{q})^{-1}$ où γ est le taux d'information manquante pour la quantité à estimer. A moins que ce taux soit extrêmement élevé ($\gamma > 0.9$), on ne tire donc que peu d'avantage à produire et analyser plus de 10 imputations. Ces q imputations peuvent être obtenues par simulation en utilisant des techniques de type MCMC [111]. Pour combiner les résultats obtenus sur les q jeux de données complets, on peut faire la moyenne empirique des q estimateurs obtenus [107].

1.3 Modélisation probabiliste intégrant l'absence de données

Dans ce travail, on s'intéresse aux méthodes de classification qui reposent sur une modélisation probabiliste explicite de la distribution des données. Le formalisme que nous présentons est celui utilisé dans l'ouvrage de Little et Rubin [86], auquel le lecteur pourra se référer pour de plus amples informations.
Dans le cas d'observations incomplètes, on peut considérer que les données observées sont constituées de deux parties :

- les valeurs observées $\mathbf{x^o}$
- les indices $\mathbf{m}$ des valeurs manquantes.

Dans un problème de classification, les données manquantes sont alors de deux natures :

- les observations manquantes $\mathbf{x^m}$
- la classification $\mathbf{z}$

Il s'agit donc de définir la distribution des champs aléatoires $\mathbf{X}$, $\mathbf{M}$ et $\mathbf{Z}$ dont $\mathbf{x} = (\mathbf{x^o}, \mathbf{x^m})$, $\mathbf{m}$, et $\mathbf{z}$ sont des réalisations. Le but de la modélisation reste naturellement la recherche d'une classification $\mathbf{z}$ optimale au sens d'un critère statistique en fonction des observations $\mathbf{x^o}$ et $\mathbf{m}$.

1.3.1 Mécanismes d'absence de données

Pour un jeu de données complètes $(\mathbf{x}, \mathbf{z})$, la distribution de l'absence de données $\mathbf{M}$ peut s'écrire de manière générale $P(\mathbf{M}|\mathbf{x}, \mathbf{z})$. On peut alors distinguer trois types de dépendances entre $\mathbf{M}$ et $(\mathbf{x}, \mathbf{z})$, de la plus simple à la plus complexe :

- Absence complètement aléatoire (*Missing Completely At Random* ou MCAR) : l'absence est indépendante des données complètes. En d'autres termes, le fait qu'une observation pour l'individu i soit manquante est indépendant des valeurs $\mathbf{x}$ et des classes $\mathbf{z}$:

$$P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}) = P(\mathbf{m})$$

  Exemple : Pour 50% des individus, la 2ème composante est manquante, les autres composantes ne manquent jamais.

$$P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}) = \prod_{i \in \mathcal{I}} (0.5\ (\mathbb{1}_{m_i=2} + \mathbb{1}_{m_i=\varnothing})).$$

- Absence aléatoire (*Missing At Random* ou MAR) : l'absence dépend seulement des données observées $\mathbf{x^o}$:

$$P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}) = P(\mathbf{m}|\mathbf{x^o})$$

Exemple : la valeur x_{i2} manque si et seulement si $x_{i1} < 0$, les autres composantes ne manquent jamais.

$$P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}) = \prod_{i \in \mathcal{I}} (\mathbb{1}_{x_{i1}<0} \mathbb{1}_{m_i=2} + \mathbb{1}_{x_{i1}>0} \mathbb{1}_{m_i=\varnothing}).$$

– Absence non aléatoire (*Not Missing At Random* ou NMAR) : lorsque l'absence n'est ni MAR, ni MCAR. En particulier, elle peut dépendre de la valeur manquante elle-même :

$$P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}) = P(\mathbf{m}|\mathbf{x^m})$$

ou de la partition non observée :

$$P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}) = P(\mathbf{m}|\mathbf{z})$$

Exemple 1 : La valeur x_{i2} manque si et seulement si elle est négative ou nulle, les autres composantes ne manquent jamais.

$$P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}) = \prod_{i \in \mathcal{I}} (\mathbb{1}_{x_{i2}\leq 0} \mathbb{1}_{m_i=2} + \mathbb{1}_{x_{i2}>0} \mathbb{1}_{m_i=\varnothing})$$

On dit alors que les données sont *censurées*.

Exemple 2 : Dans la classe 1, 70% des données sont manquantes selon la 2ème composante, les autres composantes ne manquent jamais.

$$P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}) = \prod_{i \in \mathcal{I}} [(0.7\ \mathbb{1}_{m_i=2} + 0.3\ \mathbb{1}_{m_i=\varnothing}) \mathbb{1}_{z_i=1} + \mathbb{1}_{m_i=\varnothing} \mathbb{1}_{z_i \neq 1}]$$

Une illustration de ces différents processus d'absence est visible en Figure VII.1.

Remarque. *Notons que, de manière générale, on a les inclusions :*

$$MCAR \subset MAR \subset NMAR$$

1.3.2 Propriétés de l'hypothèse MAR en classification

La grande majorité des méthodes traitant de données incomplètes se limitent aux modèles MAR ou MCAR car ceux-ci permettent de concevoir des méthodes générales d'estimation.

Factorisation de la vraisemblance En général, la distribution $P(\mathbf{m}, \mathbf{x})$ dépend de paramètres. Supposons que les paramètres d'absence $\boldsymbol{\rho}$ et de génération des observations $\boldsymbol{\psi}$ soient séparables, c'est-à-dire :

$$P(\mathbf{m}, \mathbf{x}|\boldsymbol{\rho}, \boldsymbol{\psi}) = P(\mathbf{m}|\mathbf{x}, \boldsymbol{\rho})\ P(\mathbf{x}|\boldsymbol{\psi})$$

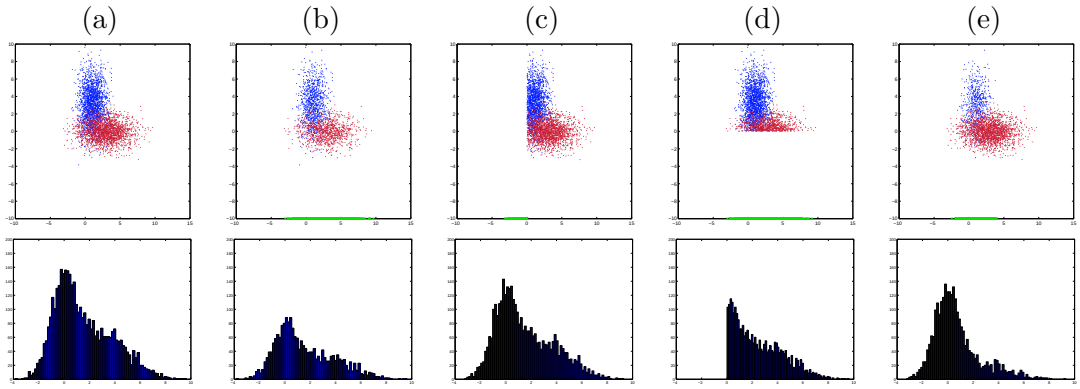


FIG. VII.1 – Illustration de différents mécanismes d'absence en dimension 2. La première ligne représente le nuage de point, c'est à dire l'ensemble des couples (x_{i1}, x_{i2}) (respectivement en abscisse et ordonnée). La deuxième ligne correspond à l'histogramme des valeurs $\{x_{i2}\}$. (a) Nuage de point initial, (b) Pour 50% des points, la deuxième composante x_{i2} est manquante (MCAR), (c) la valeur x_{i2} manque si $x_{i1} < 0$, (d) la valeur x_{i2} manque si $x_{i2} < 0$, (e) dans la classe 1, 70% des données sont manquantes selon la 2ème composante

Alors, sous l'hypothèse MAR, la vraisemblance des valeurs observées $\mathbf{x^o}$ et $\mathbf{m}$ s'écrit [86] :

$$\begin{aligned} P(\mathbf{x^o}, \mathbf{m}|\boldsymbol{\psi}, \boldsymbol{\rho}) &= \sum_{\mathbf{z}} \int P(\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}|\boldsymbol{\psi}, \boldsymbol{\rho})\ P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}, \boldsymbol{\psi}, \boldsymbol{\rho}) d\mathbf{x^m} \\ &= \sum_{\mathbf{z}} \int P(\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}|\boldsymbol{\psi})\ P(\mathbf{m}|\mathbf{x^o}, \boldsymbol{\rho}) d\mathbf{x^m} \\ &= P(\mathbf{m}|\mathbf{x^o}, \boldsymbol{\rho})\ P(\mathbf{x^o}|\boldsymbol{\psi}) \end{aligned}$$

Notons que cette équation n'est pas la simple application de la règle de Bayes puisque le paramètre $\boldsymbol{\psi}$ est absent de la première distribution et $\boldsymbol{\rho}$ de la deuxième. Cette équation nous assure que le calcul des paramètres de maximum de vraisemblance $\boldsymbol{\psi}$ peut donc se faire indépendamment de celui de $\boldsymbol{\rho}$, par maximisation de la vraisemblance $P(\mathbf{x^o}|\boldsymbol{\psi})$ des données observées $\mathbf{x^o}$.

Factorisation de la vraisemblance complète Sous l'hypothèse MAR, la vraisemblance des données complètes $(\mathbf{x^o}, \mathbf{x^m}, \mathbf{m}, \mathbf{z})$ est donnée par :

$$\begin{aligned} P(\mathbf{x^o}, \mathbf{x^m}, \mathbf{m}, \mathbf{z}|\boldsymbol{\psi}, \boldsymbol{\rho}) &= P(\mathbf{m}|\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}, \boldsymbol{\psi}, \boldsymbol{\rho})\ P(\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}|\boldsymbol{\psi}, \boldsymbol{\rho}) \\ &= P(\mathbf{m}|\mathbf{x^o}, \boldsymbol{\rho})\ P(\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}|\boldsymbol{\psi}) \end{aligned} \quad \text{(VII.1)}$$

Chercher $\boldsymbol{\psi}$ maximisant la vraisemblance complète revient donc à chercher $\boldsymbol{\psi}$ maximisant la vraisemblance $P(\mathbf{x^o}, \mathbf{x^m}, \mathbf{z}|\boldsymbol{\psi})$, indépendemment du mécanisme d'absence.

1.3.3 Applicabilité de l'hypothèse MAR

Si le modèle MAR paraît attrayant en raison des propriétés détaillées en section 1.3.2, on peut se demander dans quelle mesure il se justifie dans une situation réelle. On est certain que l'hypothèse MAR est valide lorsque l'expérimentateur décide volontairement

de ne pas mesurer certaines variables ([111], p20-22). On parle alors d'*absence planifiée* (*planned missingness*). C'est par exemple le cas lors de l'utilisation de questionnaires multiples, c'est-à-dire composés de questions qui ne sont pas toutes posées [65]. L'absence planifiée correspond en général à un modèle MCAR, voire MAR si, par exemple, les questions posées aux participants dépendent de leurs réponses antérieures.
Or, comme relevé dans [112], dans de nombreux cas réels, le mécanisme d'absence n'est pas MAR. A titre d'exemple, l'enquête clinique de [69] auprès de patients dépressifs et concernant l'efficacité d'un traitement, révèle que le taux de non-réponse est plus élevé chez les patients allant mieux et chez ceux allant encore plus mal. La non-réponse au questionnaire dépend de l'état de dépression du patient, donc de l'efficacité du traitement, c'est-à-dire de la variable que l'on cherche à mesurer. Sous l'hypothèse NMAR, l'estimation des paramètres par maximum de vraisemblance requiert alors la définition d'un modèle pour le mécanisme d'absence et la maximisation de la vraisemblance des données complètes $(\mathbf{x}, \mathbf{m}, \mathbf{z})$.
Néanmoins, les méthodes fondées sur un modèle MAR donnent souvent des résultats satisfaisants, même lorsque le phénomène d'absence semble lié aux valeurs manquantes [37]. En effet, la partie observée $\mathbf{x^o}$ est généralement suffisamment informative pour prédire les valeurs non observées $(\mathbf{x^m}, \mathbf{z})$ et la dépendance de $\mathbf{M}$ à $(\mathbf{x^m}, \mathbf{z})$ peut être négligée.

Dans la suite de ce chapitre, on suppose être dans le cas d'une absence des données selon le modèle MAR.

Les prochaines sections s'intéressent à la classification automatique de données incomplètes. Nous nous intéressons dans un premier temps à la classification d'individus indépendants puis généraliserons la méthode en section 3 au cadre d'une dépendance markovienne.

2 Mélange indépendant avec données incomplètes

Cette section présente l'algorithme EM [44] pour l'estimation et la classification automatique de données incomplètes dans un cadre de mélange indépendant, donc sans prise en compte des éventuelles dépendances entre les individus. Le lecteur intéressé pourra également se reporter à [61] ou [86].

2.1 Modèle

Comme au chapitre II, section 3, on suppose que les classes $\mathbf{Z} = \{Z_i, i \in \mathcal{I}\}$ sont indépendantes et identiquement distribuées (i.i.d), de loi paramétrée par $(\pi_k)_{k\in\mathcal{K}}$:

$$P(\mathbf{z}) = \prod_{i\in\mathcal{I}} \pi_{z_i} \tag{VII.2}$$

On suppose de plus être dans le cadre d'un bruit indépendant :

$$P(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \prod_{i\in\mathcal{I}} P(x_i^{o_i}, x_i^{m_i}|z_i) = \prod_{i\in\mathcal{I}} f(x_i|\theta_{z_i}), \tag{VII.3}$$

ce qui implique l'indépendance conditionnelle des données observées $\mathbf{X^o}$:

$$P(\mathbf{x^o}|\mathbf{z}) = \sum_{\mathbf{x^m}} \prod_{i\in\mathcal{I}} P(x_i^{o_i}, x_i^{m_i}|z_i) = \prod_{i\in\mathcal{I}} \sum_{x_i^{m_i}} P(x_i^{o_i}, x_i^{m_i}|z_i) = \prod_{i\in\mathcal{I}} f(x_i^{o_i}|\theta_{z_i}) \qquad \text{(VII.4)}$$

où on a noté $P(x_i|z_i) = f(x_i|\theta_{z_i})$ et $P(x_i^{o_i}|z_i) = f(x_i^{o_i}|\theta_{z_i})$. Notons que par abus d'écriture, nous adoptons la même notation "$f(.|\theta_{z_i})$" pour la distribution de x_i et de la marginale $x_i^{o_i}$ bien que ces deux distributions n'appartiennent pas nécessairement à la même famille de lois, ni au même espace.

2.2 Classification à paramètres connus

On suppose dans cette section que les paramètres $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\phi})$ du modèle sont connus. Leur estimation fait l'objet de la section 2.3.

Cas général Dans le cadre d'observations incomplètes, la règle du MAP (voir chapitre II, section 2.2.2) revient à choisir la configuration la plus probable conditionnellement aux données observées $\mathbf{x^o}$:

$$\mathbf{z}^{map} = \arg\max_{\mathbf{z}} P(\mathbf{z}|\mathbf{x^o}) \qquad \text{(VII.5)}$$

La règle du MPM (voir voir chapitre II, section 2.2.3) choisit, pour chaque site $i \in \mathcal{I}$, la classe la plus probable conditionnellement aux données observées $\mathbf{x^o}$:

$$z_i^{mpm} = \arg\max_{z_i} P(z_i|\mathbf{x^o}) \qquad \text{(VII.6)}$$

Cas du mélange indépendant Pour un modèle de mélange indépendant, du fait des factorisations (VII.2) et (VII.4), les règles du MAP et du MPM sont équivalentes et reviennent à classer chaque site $i \in \mathcal{I}$ dans la classe la plus probable connaissant $x_i^{o_i}$:

$$z_i^{map} = \arg\max_{z_i} P(z_i|x_i^{o_i}) = \arg\max_{z_i} \begin{cases} \pi_{z_i} f(x_i^{o_i}|\theta_{z_i}) & \text{si } o_i \neq \varnothing \\ \pi_{z_i} & \text{sinon} \end{cases}$$

Notons que tout site i pour lequel aucune observation n'est disponible ($o_i = \varnothing$) est classé dans la classe la plus probable (celle ayant la plus forte probabilité π_k).

2.3 Estimation des paramètres par l'algorithme EM

2.3.1 Principe général

Le principe de l'algorithme EM [44] s'applique très naturellement au cadre de la classification d'observations incomplètes. Le principe de l'algorithme est en effet de maximiser, à chaque itération, l'espérance de la vraisemblance des données complétées, connaissant les observations. Dans le cas de la classification de données incomplètes, les informations manquantes sont la partition cachée $\mathbf{Z}$, ainsi que les données non observées $\mathbf{X^m}$, alors que les observations sont les données observées $\mathbf{x^o}$, ainsi que les indices des données manquantes $\mathbf{m}$. L'espérance à maximiser à l'itération $(q+1)$ s'écrit donc :

$$Q(\boldsymbol{\psi}, \boldsymbol{\rho}|\boldsymbol{\psi}^{(q)}, \boldsymbol{\rho}^{(q)}) = \langle \log P(\mathbf{x^o}, \mathbf{X^m}, \mathbf{Z}, \mathbf{m}|\boldsymbol{\psi}, \boldsymbol{\rho}) | \mathbf{x^o}, \mathbf{m}, \boldsymbol{\psi}^{(q)}, \boldsymbol{\rho}^{(q)} \rangle \qquad \text{(VII.7)}$$

Sous l'hypothèse MAR, d'après (VII.1), cette espérance s'écrit encore :

$$\begin{aligned} Q(\boldsymbol{\psi}, \boldsymbol{\rho}|\boldsymbol{\psi}^{(q)}, \boldsymbol{\rho}^{(q)}) &= \langle \log P(\mathbf{x^o}, \mathbf{X^m}, \mathbf{Z}|\boldsymbol{\psi}) + \log P(\mathbf{m}|\mathbf{x^o}, \boldsymbol{\rho}) | \mathbf{x^o}, \mathbf{m}, \boldsymbol{\psi}^{(q)}, \boldsymbol{\rho}^{(q)} \rangle \\ &= \langle \log P(\mathbf{x^o}, \mathbf{X^m}, \mathbf{Z}|\boldsymbol{\psi}) | \mathbf{x^o}, \mathbf{m}, \boldsymbol{\psi}^{(q)}, \boldsymbol{\rho}^{(q)} \rangle + \log P(\mathbf{m}|\mathbf{x^o}, \boldsymbol{\rho}) \\ &= \langle \log P(\mathbf{x^o}, \mathbf{X^m}, \mathbf{Z}|\boldsymbol{\psi}) | \mathbf{x^o}, \boldsymbol{\psi}^{(q)} \rangle + \log P(\mathbf{m}|\mathbf{x^o}, \boldsymbol{\rho}) \end{aligned}$$

Sous l'hypothèse MAR, mettre à jour les paramètres $\boldsymbol{\psi}^{(q+1)}$ à l'itération $(q+1)$ revient donc à maximiser l'espérance $Q''(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)})$ indépendamment du modèle d'absence :

$$\begin{aligned} Q''(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)}) &= \langle \log P(\mathbf{x^o}, \mathbf{X^m}, \mathbf{Z}|\boldsymbol{\psi}) | \mathbf{x^o}, \boldsymbol{\psi}^{(q)} \rangle \\ &= \langle \log P(\mathbf{x^o}, \mathbf{X^m}|\mathbf{Z}, \boldsymbol{\theta}) | \mathbf{x^o}, \boldsymbol{\psi}^{(q)} \rangle + \langle \log P(\mathbf{Z}|\boldsymbol{\phi}) | \mathbf{x^o}, \boldsymbol{\psi}^{(q)} \rangle \end{aligned} \quad \text{(VII.8)}$$

L'espérance $Q''(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)})$ se décompose donc en deux termes, $Q''_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\psi}^{(q)})$ ne dépendant que des paramètres $\boldsymbol{\theta}$ des classes et $Q''_{\boldsymbol{\phi}}(\boldsymbol{\phi}, \boldsymbol{\psi}^{(q)})$ ne dépendant que des paramètres $\boldsymbol{\phi}$ de la loi a priori :

$$\begin{aligned} Q''_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\psi}^{(q)}) &= \langle \log P(\mathbf{x^o}, \mathbf{X^m}|\mathbf{Z}, \boldsymbol{\theta}) | \mathbf{x^o}, \boldsymbol{\psi}^{(q)} \rangle \\ Q''_{\boldsymbol{\phi}}(\boldsymbol{\phi}, \boldsymbol{\psi}^{(q)}) &= \langle \log P(\mathbf{Z}|\boldsymbol{\phi}) | \mathbf{x^o}, \boldsymbol{\psi}^{(q)} \rangle \end{aligned}$$

L'itération $(q+1)$ de l'algorithme EM se décompose alors en deux étapes :
(E) Calcul des termes de $Q''(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)})$ ne faisant pas intervenir $\boldsymbol{\psi}$
(M) Mise à jour des paramètres par :

$$\begin{aligned} \boldsymbol{\theta}^{(q+1)} &= \arg\max Q''_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\psi}^{(q)}) \\ \boldsymbol{\phi}^{(q+1)} &= \arg\max Q''_{\boldsymbol{\phi}}(\boldsymbol{\phi}, \boldsymbol{\psi}^{(q)}) \end{aligned}$$

2.3.2 Cas du mélange indépendant

Sous l'hypothèse de bruit indépendant (VII.3), la fonction $Q''_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\psi}^{(q)})$ à maximiser à l'itération $(q+1)$ s'écrit :

$$\begin{aligned} Q''_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\psi}^{(q)}) &= \sum_{i \in \mathcal{I}} \langle \log P(x_i^{o_i}, X_i^{m_i}|Z_i, \boldsymbol{\theta}) | x_i^{o_i}, \boldsymbol{\psi}^{(q)} \rangle \\ &= \sum_{i \in \mathcal{I}} \sum_{z_i \in \mathcal{K}} \int P(x_i^{m_i}, z_i | x_i^{o_i}, \boldsymbol{\psi}^{(q)}) \log f(x_i^{o_i}, x_i^{m_i}|\theta_{z_i}) dx_i^{m_i} \\ &= \sum_{i \in \mathcal{I}} \sum_{z_i \in \mathcal{K}} P(z_i|x_i^{o_i}, \boldsymbol{\psi}^{(q)}) \int f(x_i^{m_i}|x_i^{o_i}, \theta_{z_i}^{(q)}) \log f(x_i^{o_i}, x_i^{m_i}|\theta_{z_i}) dx_i^{m_i} \\ &= \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} t_{ik}^{(q)} \langle \log f(x_i^{o_i}, X_i^{m_i}|\theta_k) | x_i^{o_i}, \theta_k^{(q)} \rangle \end{aligned}$$

où le coefficient $t_{ik}^{(q)}$ désigne la probabilité a posteriori sachant les données observées $x_i^{o_i}$:

$$t_{ik}^{(q)} = \begin{cases} P(Z_i = k|x_i^{o_i}, \boldsymbol{\psi}^{(q)}) & \text{si } o_i \neq \varnothing \\ P(Z_i = k|\boldsymbol{\psi}^{(q)}) & \text{sinon} \end{cases}$$

Sous l'hypothèse d'indépendance des classes (VII.2), la fonction $Q''_{\boldsymbol{\phi}}(\boldsymbol{\phi}, \boldsymbol{\psi}^{(q)})$ à maximiser à l'itération $(q+1)$ s'écrit :

$$Q''_{\boldsymbol{\phi}}(\boldsymbol{\phi}, \boldsymbol{\psi}^{(q)}) = \sum_{i \in \mathcal{I}} \langle \log P(Z_i|\boldsymbol{\phi}) | x_i^{o_i}, \boldsymbol{\psi}^{(q)} \rangle = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} t_{ik}^{(q)} \log \pi_k$$

Les deux étapes (E) et (M) de cette itération $(q+1)$ sont alors :

(E) Calcul des probabilités a posteriori pour tout $i \in \mathcal{I}$ et $k \in \mathcal{K}$:

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} f(x_i^{o_i}|\theta_k^{(q)})}{\sum_{k' \in \mathcal{K}} \pi_{k'}^{(q)} f(x_i^{o_i}|\theta_{k'}^{(q)})} \text{ si } o_i \neq \varnothing \qquad \text{ou} \qquad \pi_k^{(q)} \text{ si } o_i = \varnothing$$

(M) Mise à jour des proportions $(\pi_k)_{k \in \mathcal{K}}$ et des paramètres $\boldsymbol{\theta} = (\theta_k)_{k \in \mathcal{K}}$ des densités des classes : pour tout $k \in \mathcal{K}$

$$\pi_k^{(q+1)} = \frac{\sum_{i \in \mathcal{I}} t_{ik}^{(q)}}{n} \tag{VII.9}$$

$$\theta_k^{(q+1)} = \arg\max_{\theta_k} \sum_{i \in \mathcal{I}} t_{ik}^{(q)} \langle \log f(x_i^{o_i}, X_i^{m_i}|\theta_k) | x_i^{o_i}, \theta_k^{(q)} \rangle \tag{VII.10}$$

2.3.3 Cas gaussien

Supposons que la loi de la classe $k \in \mathcal{K}$ soit une distribution gaussienne, c'est-à-dire que, pour tout $x_i \in \mathbb{R}^D$, $f(x_i|\theta_k) = \mathcal{N}(x_i|\mu_k, \Sigma_k)$. Cette section décrit l'estimation des paramètres $\theta_k = (\mu_k, \Sigma_k)$ de cette distribution gaussienne lors de l'étape (M) de l'algorithme EM.

Les équations classiques donnant les moments de vecteurs gaussiens nous assurent que :

$$\begin{aligned} f(x_i^{o_i}|\theta_k) &= \mathcal{N}(x_i^{o_i}|\mu_k^{o_i}, \Sigma_k^{o_i o_i}) \\ f(x_i^{m_i}|x_i^{o_i}, \theta_k) &= \mathcal{N}(x_i^{m_i}|\eta_{ik}, \Gamma_{ik}) \end{aligned}$$

où on a noté :

$$\begin{aligned} \Sigma_k^{o_i o_i} &= \{(\Sigma_k)_{st}, s \in o_i, t \in o_i\} \\ \Sigma_k^{m_i o_i} &= \{(\Sigma_k)_{st}, s \in m_i, t \in o_i\} = (\Sigma_k^{o_i m_i})' \\ \Sigma_k^{m_i m_i} &= \{(\Sigma_k)_{st}, s \in m_i, t \in m_i\} \\ \mu_k^{m_i} &= \{(\mu_k)_s, s \in m_i\} \\ \mu_k^{o_i} &= \{(\mu_k)_s, s \in o_i\} \\ \eta_{ik} &= \mu_k^{m_i} + \Sigma_k^{m_i o_i} (\Sigma_k^{o_i o_i})^{-1} (x_i^{o_i} - \mu_k^{o_i}) \\ \Gamma_{ik} &= \Sigma_k^{m_i m_i} - \Sigma_k^{m_i o_i} (\Sigma_k^{o_i o_i})^{-1} \Sigma_k^{o_i m_i} \end{aligned}$$

Mise à jour de la moyenne Puisque $f(x_i|\theta_k) = \mathcal{N}(x_i|\mu_k, \Sigma_k)$, on a :

$$2 \log f(x_i^{o_i}, x_i^{m_i}|\theta_k) = 2 \log f(x_i|\theta_k) = cste - \log|\Sigma_k| - (x_i - \mu_k)\Sigma_k^{-1}(x_i - \mu_k)'$$

Quitte à réordonner les indices sous la forme :

$$(x_i - \mu_k) \;=\; \left(\begin{array}{c} x_i^{o_i} - \mu_k^{o_i} \\ x_i^{m_i} - \mu_k^{m_i} \end{array} \right),$$

la dérivée de $\log f(x_i^{o_i}, x_i^{m_i}|\theta_k)$ par rapport à μ_k donne :

$$\frac{\partial \log f(x_i^{o_i}, x_i^{m_i}|\theta_k)}{\partial \mu_k} = \Sigma_k^{-1}(x_i - \mu_k) = \Sigma_k^{-1} \left(\begin{array}{c} x_i^{o_i} - \mu_k^{o_i} \\ x_i^{m_i} - \mu_k^{m_i} \end{array} \right)$$

Donc,

$$\begin{aligned} \frac{\partial}{\partial \mu_k} \langle \log f(x_i^{o_i}, X_i^{m_i}|\theta_k)|x_i^{o_i}, \theta_k^{(q)} \rangle &= \langle \frac{\partial}{\partial \mu_k} \log f(x_i^{o_i}, X_i^{m_i}|\theta_k)|x_i^{o_i}, \theta_k^{(q)} \rangle \\ &= \Sigma_k^{-1} \langle \left(\begin{array}{c} x_i^{o_i} - \mu_k^{o_i} \\ X_i^{m_i} - \mu_k^{m_i} \end{array} \right) |x_i^{o_i}, \theta_k^{(q)} \rangle \\ &= \Sigma_k^{-1} \left(\begin{array}{c} x_i^{o_i} - \mu_k^{o_i} \\ \langle X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)} \rangle - \mu_k^{m_i} \end{array} \right) \\ &= \Sigma_k^{-1} \left(\begin{array}{c} x_i^{o_i} - \mu_k^{o_i} \\ \eta_{ik}^{(q)} - \mu_k^{m_i} \end{array} \right) \end{aligned}$$

On déduit alors de l'équation (VII.10) que

$$\mu_k^{(q+1)} = \arg\max_{\mu_k} \sum_{i \in \mathcal{I}} t_{ik}^{(q)} \left(\begin{array}{c} x_i^{o_i} - \mu_k^{o_i} \\ \eta_{ik}^{(q)} - \mu_k^{m_i} \end{array} \right)$$

La mise à jour de la composante $d \in [\![1, D]\!]$ de la moyenne μ_k à l'itération $(q+1)$ s'écrit donc :

$$(\mu_k^d)^{(q+1)} = \frac{\sum_i t_{ik}^{(q)} l_{ik}^d}{\sum_i t_{ik}^{(q)}} \tag{VII.11}$$

où, pour tout $i \in \mathcal{I}$, $k \in \mathcal{K}$ et $d \in [\![1, D]\!]$

$$l_{ik}^d = \left\{ \begin{array}{ll} x_i^d & \text{si } d \in o_i \\ (\eta_{ik}^d)^{(q)} & \text{sinon} \end{array} \right.$$

La mise à jour (VII.11) consiste donc à remplacer les variable manquantes $x_i^{m_i}$ par la moyenne $\eta_{ik}^{(q)}$ de la loi conditionnelle $f(X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)})$.

Mise à jour de la matrice de covariance En notant que :

$$\begin{aligned} \frac{\partial \log |\Sigma_k|}{\partial \Sigma_k^{-1}} = -\frac{\partial \log |\Sigma_k^{-1}|}{\partial \Sigma_k^{-1}} &= -(\Sigma_k)' = -\Sigma_k \\ \frac{\partial (x_i - \mu_k)' \Sigma_k^{-1} (x_i - \mu_k)}{\partial \Sigma_k^{-1}} &= (x_i - \mu_k)(x_i - \mu_k)', \end{aligned}$$

la dérivée de $\log f(x_i^{o_i}, x_i^{m_i}|\theta_k)$ par rapport à Σ_k^{-1} donne :

$$2\frac{\partial \log f(x_i^{o_i}, x_i^{m_i}|\theta_k)}{\partial \Sigma_k^{-1}} = \Sigma_k - (x_i - \mu_k)(x_i - \mu_k)'$$

Or, quitte à réordonner les indices,

$$(x_i - \mu_k)(x_i - \mu_k)' = \begin{pmatrix} (x_i^{o_i} - \mu_k^{o_i})(x_i^{o_i} - \mu_k^{o_i})' & (x_i^{o_i} - \mu_k^{o_i})(x_i^{m_i} - \mu_k^{m_i})' \\ (x_i^{m_i} - \mu_k^{m_i})(x_i^{o_i} - \mu_k^{o_i})' & (x_i^{m_i} - \mu_k^{m_i})(x_i^{m_i} - \mu_k^{m_i})' \end{pmatrix}$$

En notant que $\eta_{ik}^{(q)} = \langle X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)}\rangle$ et que

$$\begin{aligned} & \langle (X_i^{m_i} - \mu_k^{m_i})(X_i^{m_i} - \mu_k^{m_i})'|x_i^{o_i}, \theta_k^{(q)}\rangle \\ = & (\langle X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)}\rangle - \mu_k^{m_i})(\langle X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)}\rangle - \mu_k^{m_i})' + Var(X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)}) \\ = & (\eta_{ik}^{(q)} - \mu_k^{m_i})(\eta_{ik}^{(q)} - \mu_k^{m_i})' + \Gamma_{ik}^{(q)}, \end{aligned}$$

on a donc

$$\begin{aligned} & 2\frac{\partial}{\partial \Sigma_k^{-1}} \langle \log f(x_i^{o_i}, X_i^{m_i}|\theta_k)|x_i^{o_i}, \theta_k^{(q)}\rangle \\ = & \Sigma_k - \begin{pmatrix} (x_i^{o_i} - \mu_k^{o_i})(x_i^{o_i} - \mu_k^{o_i})' & (x_i^{o_i} - \mu_k^{o_i})(\eta_{ik}^{(q)} - \mu_k^{m_i})' \\ (\eta_{ik}^{(q)} - \mu_k^{m_i})(x_i^{o_i} - \mu_k^{o_i})' & (\eta_{ik}^{(q)} - \mu_k^{m_i})(\eta_{ik}^{(q)} - \mu_k^{m_i})' + \Gamma_{ik}^{(q)} \end{pmatrix} \end{aligned}$$

D'après l'équation (VII.10),

$$\Sigma_k^{(q+1)} = \arg\max_{\Sigma_k} \sum_i t_{ik}^{(q)} \left[\Sigma_k - \begin{pmatrix} (x_i^{o_i} - \mu_k^{o_i})(x_i^{o_i} - \mu_k^{o_i})' & (x_i^{o_i} - \mu_k^{o_i})(\eta_{ik}^{(q)} - \mu_k^{m_i})' \\ (\eta_{ik}^{(q)} - \mu_k^{m_i})(x_i^{o_i} - \mu_k^{o_i})' & (\eta_{ik}^{(q)} - \mu_k^{m_i})(\eta_{ik}^{(q)} - \mu_k^{m_i})' + \Gamma_{ik}^{(q)} \end{pmatrix} \right]$$

La mise à jour de la composante $d, d' \in [\![1, D]\!]$ de la matrice de covariance Σ_k à l'itération $(q+1)$ s'écrit donc :

$$\Sigma_k^{dd'(q+1)} = \frac{\sum_i t_{ik}^{(q)} S_{ik}^{dd'(q)}}{\sum_i t_{ik}^{(q)}} \tag{VII.12}$$

en notant pour tout $i \in \mathcal{I}$, $k \in \mathcal{K}$, $d, d' \in [\![1, D]\!]$,

$$(S_{ik}^{dd'})^{(q)} = \begin{cases} (x_i^d - \mu_k^{d(q)})(x_i^{d'} - \mu_k^{d'(q)}) & \text{si } d \in o_i \text{ et } d' \in o_i \\ (x_i^d - \mu_k^{d(q)})(\eta_{ik}^{d'(q)} - \mu_k^{d'(q)}) & \text{si } d \in o_i \text{ et } d' \in m_i \\ (\eta_{ik}^{d\ (q)} - \mu_k^{d(q)})(x_i^{d'} - \mu_k^{d'(q)}) & \text{si } d \in m_i \text{ et } d' \in o_i \\ (\eta_{ik}^{d\ (q)} - \mu_k^{d(q)})(\eta_{ik}^{d'(q)} - \mu_k^{d'(q)}) + \Gamma_{ik}^{dd'(q)} & \text{si } d \in m_i \text{ et } d' \in m_i \end{cases}$$

Notons que, du fait du terme $\Gamma_{ik}^{dd'(q)}$ de la dernière ligne, cette mise à jour n'est pas équivalente à remplacer les variables manquantes $x_i^{m_i}$ par leur moyenne $\eta_{ik}^{(q)}$ conditionnellement aux observations $x_i^{o_i}$. La technique d'imputation par la moyenne (voir section 1.2) tend donc à biaiser la variance estimée [61] vers 0.

Remarque. *Il existe également des formules explicites de mise à jour des paramètres θ_k des lois $f(x_i|\theta_k)$ lorsque f n'est pas gaussienne, en particulier pour des lois de Laplace diagonales (voir [41], p.196). Les calculs montrent que la mise à jour de la composante $d \in [\![1, D]\!]$ de la dispersion λ_k n'est pas équivalente à remplacer les valeurs manquantes par la moyenne et que cette technique d'imputation tend à biaiser la dispersion estimée vers* 0.

2.4 Classer les données suite à l'algorithme EM

Une fois les paramètres estimés, la classification $\mathbf{z}$ recherchée peut être obtenue par application des règles du MAP (VII.5) ou MPM (VII.6) sans aucun calcul supplémentaire. Sous l'hypothèses de mélange indépendant, ces deux règles sont équivalentes et s'écrivent :

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{z_i \in \mathcal{K}} P(z_i|x_i^{o_i}, \psi_{z_i}) = \arg\max_{k \in \mathcal{K}} t_{ik}$$

Les valeurs des t_{ik} en sortie de l'algorithme EM permettent donc directement de restaurer les classes $\mathbf{z}^{map}$. Notons que si au site i, aucune observation n'est disponible ($o_i = \varnothing$), $t_{ik} = \pi_k$ et le site est naïvement classé dans la classe k la plus probable, faisant de ce fait grossir artificiellement la classe la plus représentée.

Cette section a présenté le cadre général de l'estimation et la classification d'individus indépendants les uns des autres. Nous proposons maintenant de prendre en compte les dépendances éventuelles entre individus à l'aide d'un graphe et une modélisation markovienne.

3 Champ de Markov caché avec données incomplètes

Cette section présente une méthode de classification automatique de données incomplètes sous une modélisation par champ de Markov caché à bruit indépendant. Les individus, ou sites, sont supposés être en interaction à travers un graphe définissant une structure de voisinage.

3.1 Modèle

Comme au chapitre II, section 4, on suppose que les classes $\mathbf{Z} = \{Z_i, i \in \mathcal{I}\}$ définissent un champ de Markov, de distribution de Gibbs P_G paramétrée par $\boldsymbol{\phi} = (\phi_k)_{k \in \mathcal{K}}$:

$$P_G(\mathbf{z}|\boldsymbol{\phi}) = W(\boldsymbol{\phi})^{-1} \exp(-H(\mathbf{z}; \boldsymbol{\phi}))$$

On suppose de plus l'indépendance des données conditionnellement aux classes (équation VII.3 et VII.4). Sous cette hypothèse, les champs $\mathbf{Z}|\mathbf{x}$ et $\mathbf{Z}|\mathbf{x^o}$ sont tous deux des champs de Markov paramétrés par $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\phi})$, d'énergie respectives :

$$\begin{aligned} H(\mathbf{z}|\mathbf{x}; \boldsymbol{\psi}) &= H(\mathbf{z}; \boldsymbol{\phi}) - \sum_{i \in \mathcal{I}} \log f(x_i|\theta_{z_i}) \\ H(\mathbf{z}|\mathbf{x^o}; \boldsymbol{\psi}) &= H(\mathbf{z}; \boldsymbol{\phi}) - \sum_{i \in \mathcal{I}} \log f(x_i^{o_i}|\theta_{z_i}) \end{aligned}$$

et de fonction de partition $W(\boldsymbol{\phi})$.

3.2 Approximation en champ moyen

Notons $\tilde{\mathbf{z}}^{\mathbf{x}^o}$ le champ des voisins, déterminé conditionnellement aux observations $\mathbf{x^o}$, dans l'approximation de type champ moyen de la loi conditionnelle $P_G(\mathbf{z}|\mathbf{x^o}, \boldsymbol{\psi})$:

$$P_G(\mathbf{z}|\mathbf{x^o}, \boldsymbol{\psi}) \approx \prod_{i \in \mathcal{I}} P_{\tilde{\mathbf{z}}^{\mathbf{x}^o}}(z_i|x_i^{o_i}, \boldsymbol{\psi}) = \prod_{i \in \mathcal{I}} P_G(z_i|x_i^{o_i}, \tilde{\mathbf{z}}_{N_i}^{\mathbf{x}^o}, \boldsymbol{\psi}) \qquad \text{(VII.13)}$$

La loi du couple $(\mathbf{X}, \mathbf{Z})$ est alors approximée par un modèle indépendant :

$$P_G(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}) \approx \prod_{i \in \mathcal{I}} P_{\tilde{\mathbf{z}}^{\mathbf{x}^o}}(z_i|\boldsymbol{\phi}) f(x_i|\theta_{z_i}) \tag{VII.14}$$

3.3 Estimation des paramètres par l'algorithme NREM

Si les paramètres $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\phi})$ du modèle sont connus, dans le cadre d'observations incomplètes, les règle du MAP (voir chapitre II, section 2.2.2) et du MPM (voir chapitre II, section 2.2.3) sont données par les équations (VII.5) et (VII.6). Néanmoins, la distribution $P_G(\mathbf{z}|\mathbf{x^o})$ étant markovienne, ces règles ne peuvent être appliquées sans approximation. Pour cela, les algorithmes de recuit simulé ou l'algorithme ICM (chapitre II, section 4.3) peuvent être utilisés sur le champ de Markov $\mathbf{Z}|\mathbf{x^o}$.

Lorsque ces paramètres sont à estimer, dans le cadre d'une modélisation par champ de Markov caché, l'algorithme EM ne peut être appliqué sans approximation. En effet, l'espérance $Q''(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)})$ à maximiser à l'itération $(q+1)$ (équation VII.8) nécessite le calcul de la distribution markovienne $P(\mathbf{z}|\mathbf{x})$. Le principe de l'algorithme NREM [29] (voir chapitre II, section 4.4.2), initialement proposé dans le cadre de la classification d'observations complètes, peut néanmoins être utilisé pour l'estimation des paramètres $\boldsymbol{\psi}$ dans le cadre d'observations incomplètes. Une itération de cet algorithme consiste alors à effectuer une étape (NR) de restauration du champ des voisins $\tilde{\mathbf{z}}^{\mathbf{x}^o}$ en fonction des données observées $\mathbf{x^o}$ et une étape (EM) de l'algorithme EM décrit en section 2.3 sur le mélange indépendant ainsi obtenu (équation VII.14) :

(EM) **Estimation :** Mettre à jour les estimateurs $\boldsymbol{\psi}^{(q+1)}$ des paramètres en appliquant l'algorithme EM sur le modèle de mélange indépendant (section 2.3.2) défini par la loi jointe

$$P_{\tilde{\mathbf{z}}^{\mathbf{x}^o}}(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}) = \prod_{i \in \mathcal{I}} \tilde{\pi}_{iz_i} f(x_i|\theta_{z_i}) \qquad \text{où } \tilde{\pi}_{iz_i} = P_G(z_i|\tilde{\mathbf{z}}^{\mathbf{x}^o}_{N_i}, \boldsymbol{\phi}) \tag{VII.15}$$

avec $\boldsymbol{\psi}^{(q)}$ comme valeur initiale des paramètres.

(NR) **Choix des voisins :** créer, à partir des données observées $\mathbf{x^o}$ et de l'estimation courante $\boldsymbol{\psi}^{(q+1)}$ des paramètres, un nouveau champ des voisins $\tilde{\mathbf{z}}^{\mathbf{x}^o}$.

L'étape (EM) en pratique Sous l'approximation en champ moyen (VII.13), l'espérance $Q''(\boldsymbol{\psi}|\boldsymbol{\psi}^{(q)})$ à maximiser à l'itération $(q+1)$ se décompose en somme de deux termes, $Q''_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\psi}^{(q)})$ ne dépendant que des paramètres $\boldsymbol{\theta}$ des classes et $Q''_{\boldsymbol{\phi}}(\boldsymbol{\phi}, \boldsymbol{\psi}^{(q)})$ ne dépendant que des paramètres $\boldsymbol{\phi}$ de la loi a priori :

$$\begin{aligned} Q''_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\psi}^{(q)}) &= \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \tilde{t}^{(q)}_{ik} \langle \log f(x_i^{o_i}, X_i^{m_i}|\theta_k)|x_i^{o_i}, \theta_k^{(q)} \rangle \\ Q''_{\boldsymbol{\phi}}(\boldsymbol{\phi}, \boldsymbol{\psi}^{(q)}) &= \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \tilde{t}^{(q)}_{ik} \log \tilde{\pi}_{ik} \end{aligned}$$

où $\tilde{\pi}_{ik} = P_G(Z_i = k|\tilde{\mathbf{z}}^{\mathbf{x}^o}_{N_i}, \boldsymbol{\phi})$ et $\tilde{t}^{(q)}_{ik}$ désigne la probabilité a posteriori sachant les données observées $x_i^{o_i}$:

$$\tilde{t}^{(q)}_{ik} = \begin{cases} P_{\tilde{\mathbf{z}}^{\mathbf{x}^o}}(Z_i = k|x_i^{o_i}, \boldsymbol{\psi}^{(q)}) & \text{si } o_i \neq \varnothing \\ P_{\tilde{\mathbf{z}}^{\mathbf{x}^o}}(Z_i = k|\boldsymbol{\psi}^{(q)}) & \text{sinon} \end{cases}$$

Les deux étapes de l'itération $(q+1)$ de (EM) sont alors :

(E) Calcul des probabilités a posteriori pour tout $i \in \mathcal{I}$ et $k \in \mathcal{K}$:

$$\tilde{t}^{(q)}_{ik} = \frac{\tilde{\pi}^{(q)}_{ik} f(x_i^{o_i}|\theta_k^{(q)})}{\sum_{k' \in \mathcal{K}} \tilde{\pi}^{(q)}_{ik'} f(x_i^{o_i}|\theta_{k'}^{(q)})} \text{ si } o_i \neq \varnothing \qquad \text{ou} \qquad \tilde{\pi}^{(q)}_{ik} \text{ si } o_i = \varnothing$$

(M) Mise à jour des paramètres $\boldsymbol{\psi} = (\psi_k)_{k\in\mathcal{K}}$ de la distribution $P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{z}|\psi)$ et des paramètres $\boldsymbol{\theta} = (\theta_k)_{k\in\mathcal{K}}$ des densités $f(.|\theta_k)$:

$$\boldsymbol{\phi}^{(q+1)} = \arg\max_{\boldsymbol{\phi}} \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} \tilde{t}^{(q)}_{ik} \log \tilde{\pi}_{ik} \tag{VII.16}$$

$$\theta_k^{(q+1)} = \arg\max_{\theta_k} \sum_{i\in\mathcal{I}} \tilde{t}^{(q)}_{ik} \langle \log f(x_i^{o_i}, X_i^{m_i}|\theta_k)|x_i^{o_i}, \theta_k^{(q)}\rangle \tag{VII.17}$$

Notons que l'équation (VII.17) est la même que dans le cas du mélange indépendant (voir section 2.3.2), en remplaçant t_{ik} par $\tilde{t}_{ik}$. En particulier, dans le cas de densités $f(.|\theta_k) \sim \mathcal{N}(\mu_k, \Sigma_k)$, la mise à jour des paramètres est explicite et donnée par les équations (VII.11) et (VII.12). De plus, l'équation (VII.16) est la même que pour la mise à jour des paramètres $\boldsymbol{\phi}$ sans observation incomplète (équation II.33).

L'étape (NR) en pratique A l'itération $(q+1)$, la mise à jour du champ $\tilde{\mathbf{z}}^{\mathbf{x}^o}$ est la suivante :

- *Algorithme en champ moyen* : fixer $\tilde{\mathbf{z}}^{\mathbf{x}^o}$ à l'estimation en champ moyen de l'espérance de la distribution conditionnelle $P_G(\mathbf{z}|\mathbf{x^o}, \boldsymbol{\psi}^{(q+1)})$
- *Algorithme en champ modal* : fixer $\tilde{\mathbf{z}}^{\mathbf{x}^o}$ à l'estimation en champ modal du mode de la distribution conditionnelle $P_G(\mathbf{z}|\mathbf{x^o}, \boldsymbol{\psi}^{(q+1)})$
- *Algorithme en champ simulé* : simuler $\tilde{\mathbf{z}}^{\mathbf{x}^o}$ selon la loi conditionnelle $P_G(\mathbf{z}|\mathbf{x^o}, \boldsymbol{\psi}^{(q+1)})$, via l'échantillonneur de Gibbs.

3.4 Classer les données suite à l'algorithme NREM.

Sous l'approximation en champ moyen (VII.15), on est ramené à un modèle de mélange indépendant et le calcul du MAP, comme du MPM, conduisent à choisir en chaque site i la classe la plus probable connaissant l'observation $x_i^{o_i}$:

$$\forall i \in \mathcal{I},\ z_i^{map} = \arg\max_{z_i\in\mathcal{K}} P_{\tilde{\mathbf{z}}^{\mathbf{x}^o}}(z_i|x_i^{o_i}) \approx \arg\max_{z_i\in\mathcal{K}} \tilde{\pi}_{iz_i} f(x_i^{o_i}|\theta_{z_i}) = \arg\max_{k\in\mathcal{K}} \tilde{t}_{ik} \tag{VII.18}$$

Remarquons que, si au site i, aucune observation n'est disponible ($o_i = \varnothing$), $\tilde{t}_{ik} = \tilde{\pi}_{ik}$ et le site est classé dans la classe k la plus probable en fonction du champ de ses voisins $\tilde{\mathbf{z}}^{\mathbf{x^o}}_{N_i}$. Cette caractéristique est un avantage certain en comparaison au choix naïf du modèle indépendant (voir section 2.2) classant i dans la classe la plus représentée.

Chapitre VIII

Expériences

Les expériences présentées dans ce chapitre, et en particulier l'application à la classification de données d'expression de gènes (voir section 2) sont le fruit d'une collaboration avec Matthieu Vignes de l'équipe *BioSS* du *Scottish Crop Research Institute* (`http://www.bioss.ac.uk`). On pourra encore se reporter à l'article [17] pour description et analyse de ces expériences.

1 Sur données simulées

Nous avons effectué un grand nombre d'expériences sur données simulées par simulation d'un modèle de champ de Markov caché (section 1.1) ou par bruitage d'une image synthétique (section 1.2). Pour chacune de ces simulations, nous avons considéré qu'une certaine proportion ρ d'observations étaient manquantes, et ce, selon divers mécanismes :

– *mécanisme d'absence aléatoire* : on choisit au hasard les valeurs manquantes. Pour chaque observation $x_i \in \mathbb{R}^D$, la composante x_{id}, $d \in [\![1, D]\!]$, est manquante avec la probabilité ρ. Avec les notations du chapitre VII, section 1.1,

$$P(\mathbf{m}|\mathbf{x}, \mathbf{z}) = P(\mathbf{m}) = \prod_{i \in \mathcal{I}} \rho^{|m_i|}(1-\rho)^{D-|m_i|}$$

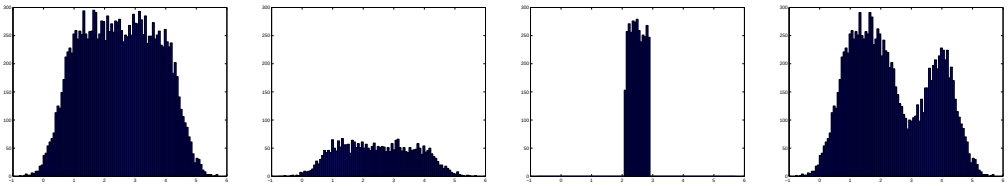
Il s'agit d'un mécanisme d'absence complètement aléatoire (MCAR).

– *mécanisme d'absence aléatoire totale* : on choisit aléatoirement les observations manquantes dans leur intégralité. En chaque site $i \in \mathcal{I}$, l'observation $x_i \in \mathbb{R}^D$ est totalement manquante (tous les x_{id} sont manquants) avec la probabilité ρ.

$$P(\mathbf{m}|\mathbf{x}, \mathbf{z}) = P(\mathbf{m}) = \prod_{i \in \mathcal{I}} (\rho \, \mathbb{1}_{|m_i|=D} + (1-\rho) \, \mathbb{1}_{|m_i|=0})$$

Il s'agit d'un mécanisme d'absence complètement aléatoire (MCAR). Notons que en dimension 1, les mécanisme d'absence aléatoire et aléatoire totale sont équivalents.

– *mécanisme d'absence par censure* : les plus grandes et plus petites observations sont manquantes. Si on note $q_{\frac{\rho}{2}}$ le quantile empirique d'ordre $\frac{\rho}{2}$ (c'est à dire tel qu'une proportion $\frac{\rho}{2}$ des observations soient $\leq q_{\frac{\rho}{2}}$) et $q_{1-\frac{\rho}{2}}$ le quantile empirique d'ordre



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

FIG. VIII.1 – (a) Histogramme correspondant à un mélange de 4 classes équiprobables et un bruit gaussien $\mathcal{N}(0, \sigma^2 = 0.25)$ (b) 80% des observations sont manquantes aléatoirement (mécanisme d'absence aléatoire) (c) 80% des observations sont censurées à gauche et à droite (mécanisme d'absence par censure) (d) 80% des observations sont manquantes dans la classe 3 (mécanisme d'absence dans une classe)

$1 - \frac{\rho}{2}$ (c'est à dire tel qu'une proportion $\frac{\rho}{2}$ des observations soient $\geq q_{1-\frac{\rho}{2}}$),

$$P(\mathbf{m}|\mathbf{x}, \mathbf{z}) = P(\mathbf{m}|\mathbf{x}) = \prod_{i \in S} \prod_{d=1}^{D} [(\mathbb{1}_{x_{id} > q_{1-\frac{\rho}{2}}} + \mathbb{1}_{x_{id} < q_{\frac{\rho}{2}}}) \mathbb{1}_{d \in m_i} + \mathbb{1}_{q_{\frac{\rho}{2}} \leq x_{id} \leq q_{1-\frac{\rho}{2}}} \mathbb{1}_{d \notin m_i}]$$

On parle alors de données censurées à gauche et à droite. Il s'agit d'un mécanisme d'absence non aléatoire (NMAR).

– *mécanisme d'absence dans une classe* : les données sont manquantes aléatoirement pour une classe k déterminée. Pour chaque obervation $x_i \in \mathbb{R}^D$ telle que $z_i = k$, la composante $d \in [\![1, D]\!]$ est manquante avec la probabilité ρ.

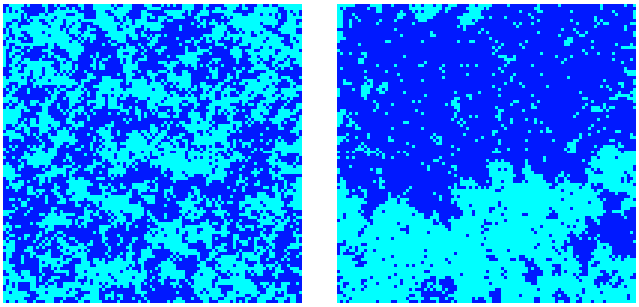
$$P(\mathbf{m}|\mathbf{x}, \mathbf{z}) = P(\mathbf{m}|\mathbf{z}) = \prod_{i \in \mathcal{I}} (\rho^{|m_i|} (1 - \rho)^{D - |m_i|} \mathbb{1}_{z_i = k} + \mathbb{1}_{|m_i| = 0} \mathbb{1}_{z_i \neq k})$$

Il s'agit d'un mécanisme d'absence non aléatoire (NMAR).

Une illustration de ces différents mécanismes d'absence est visible en Figure VIII.1.

Pour chaque simulation, les paramètres du modèle sont estimés selon :
– EMmiss : l'algorithme EM avec observations incomplètes, sous l'hypothèse d'un modèle de mélange gaussien indépendant (voir chapitre VII, section 2.3).
– SFmiss : l'algorithme en champ simulé avec observations incomplètes, sous l'hypothèse d'un champ de Markov caché à bruit indépendant (voir chapitre VII, section 3.3).
– SF : l'algorithme en champ simulé (décrit au chapitre II, section 4.4.2) sur les observations complétées par imputation .

Concernant l'imputation des données, plusieurs alternatives sont comparées : trois méthodes effectuées en pré-traitement, avant l'estimation du modèle (*off-line*) :
– KNN : la méthode des k-plus proches voisins ($k = 15$)
– ZERO : les observations manquantes sont remplacées par des zéro
– MEAN : une observation x_{id} manquante est remplacée par la moyenne des observations x_{jd} non manquantes ($d \in o_j$).



$\beta = 0.3$ $\beta = 0.4$

FIG. VIII.2 – Simulation d'un modèle de Potts à 2 couleurs, pour différentes valeurs du paramètre spatial β.

ainsi que deux méthodes d'imputation effectuées au cours de l'algorithme d'estimation (*on-line*) :

- UMEAN : à l'itération $(q+1)$ de l'algorithme en champ simulé, lors du calcul des paramètres de la classe k, $x_i^{m_i}$ est remplacé par la moyenne $\mu_k^{(q)}$ de la loi $f(X_i^{m_i}|\theta_k^{(q)})$.
- CMEAN : à l'itération $(q+1)$ de l'algorithme en champ simulé, lors du calcul des paramètres de la classe k, $x_i^{m_i}$ est remplacé par la moyenne $\eta_{ik}^{(q)}$ de la loi conditionnelle $f(X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)})$.

La classification finale est obtenue par application de la règle du MAP ou de son approximation en champ moyen (chapitre II, sections 4.4.3 et 2.4 ou chapitre VII, section 3.4).

1.1 Simulation d'un champ de Markov caché bruité

On simule un modèle de Potts à 2 couleurs ($K = 2$) sur une grille régulière de taille 100×100 pour différentes valeurs du paramètre spatial β (0.3 et 0.4) et un voisinage d'ordre 2, en utilisant l'échantillonneur de Gibbs (1000 itérations). Les images correspondantes sont celles de la figure VIII.2. Chacun des pixels est dupliqué D fois pour constituer une image en dimension D. Cette image est ensuite bruitée par ajout d'un bruit blanc gaussien D-dimensionnel. Nous présentons les résultats obtenus en dimension 1 avec $\sigma^2 = 0.25$ (Figure VIII.3), en dimension 10 avec $\Sigma = 2\mathbb{I}_{10}$ (Figure VIII.4) et en dimension 4 avec Σ non diagonale, de coefficient $\Sigma_{dd'} = 0.2$ si $d \neq d'$, $\Sigma_{dd} = 0.5$ (Figure VIII.5). Notons que la méthode d'imputation par les k-plus proches voisins ne peut être utilisée aux sites i pour lesquels $o_i = \varnothing$. C'est en particulier le cas en dimension 1, ainsi qu'avec le mécanisme d'absence totale. Remarquons encore que lorsque la matrice de covariance Σ est diagonale, $\eta_{ik} = \mu_k$ puisque, pour tout x_i, $f(X_i^{m_i}|x_i^{o_i}) = f(X_i^{m_i})$ et donc les imputations par la moyenne conditionnelle (CMEAN) et inconditionnelle (UMEAN) sont équivalentes.

Il apparaît sur ces simulations que, parmi les méthodes d'imputation *off-line* (en prétraitement), l'imputation par des 0 (ZERO+SF) est la moins performante. L'imputation par les k-plus proches voisins (lorsqu'elle est possible), donne des résultats équivalents

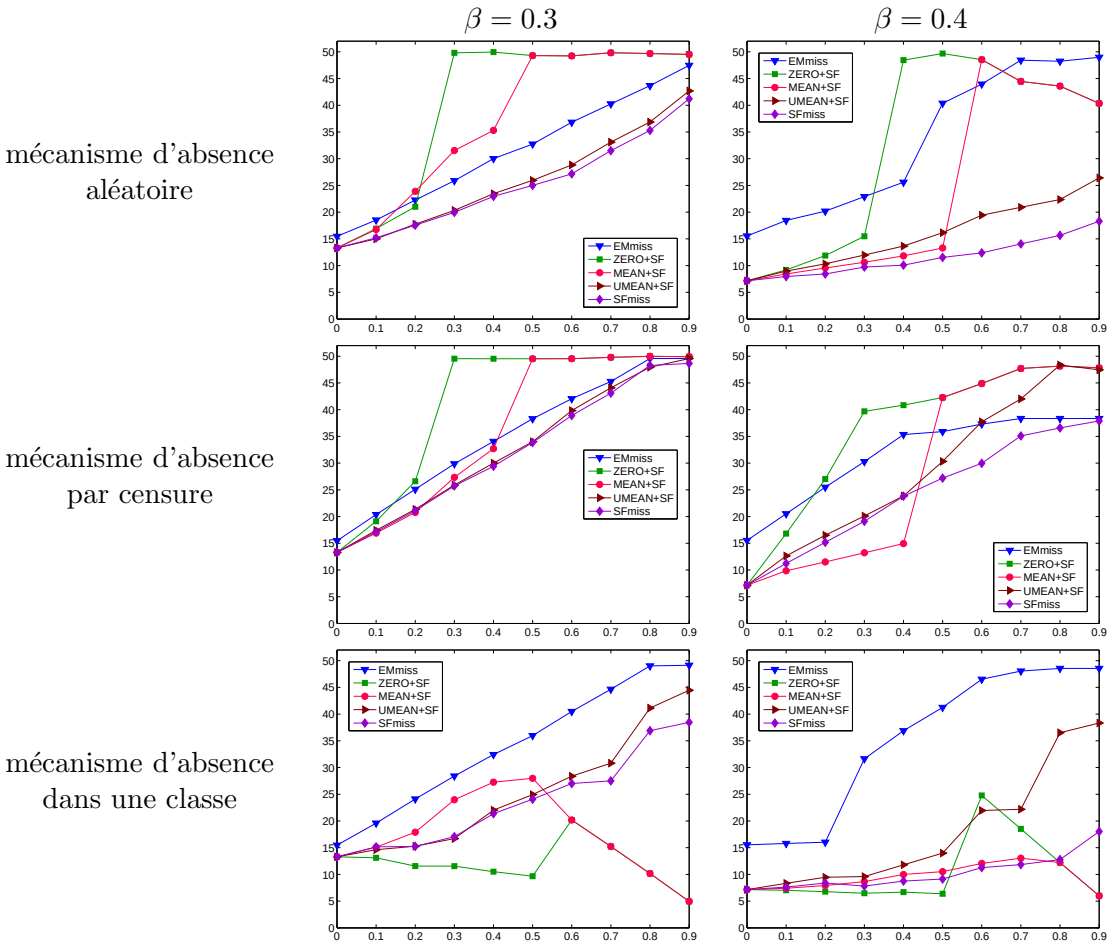


FIG. VIII.3 – Taux d'erreur de classification en fonction du paramètre ρ sur simulation d'un champ de Markov caché (modèle de Potts de paramètres $\beta = 0.3$ et 0.4 bruité par un bruit additif gaussien en dimension $D = 1$ de variance $\sigma^2 = 0.25$).

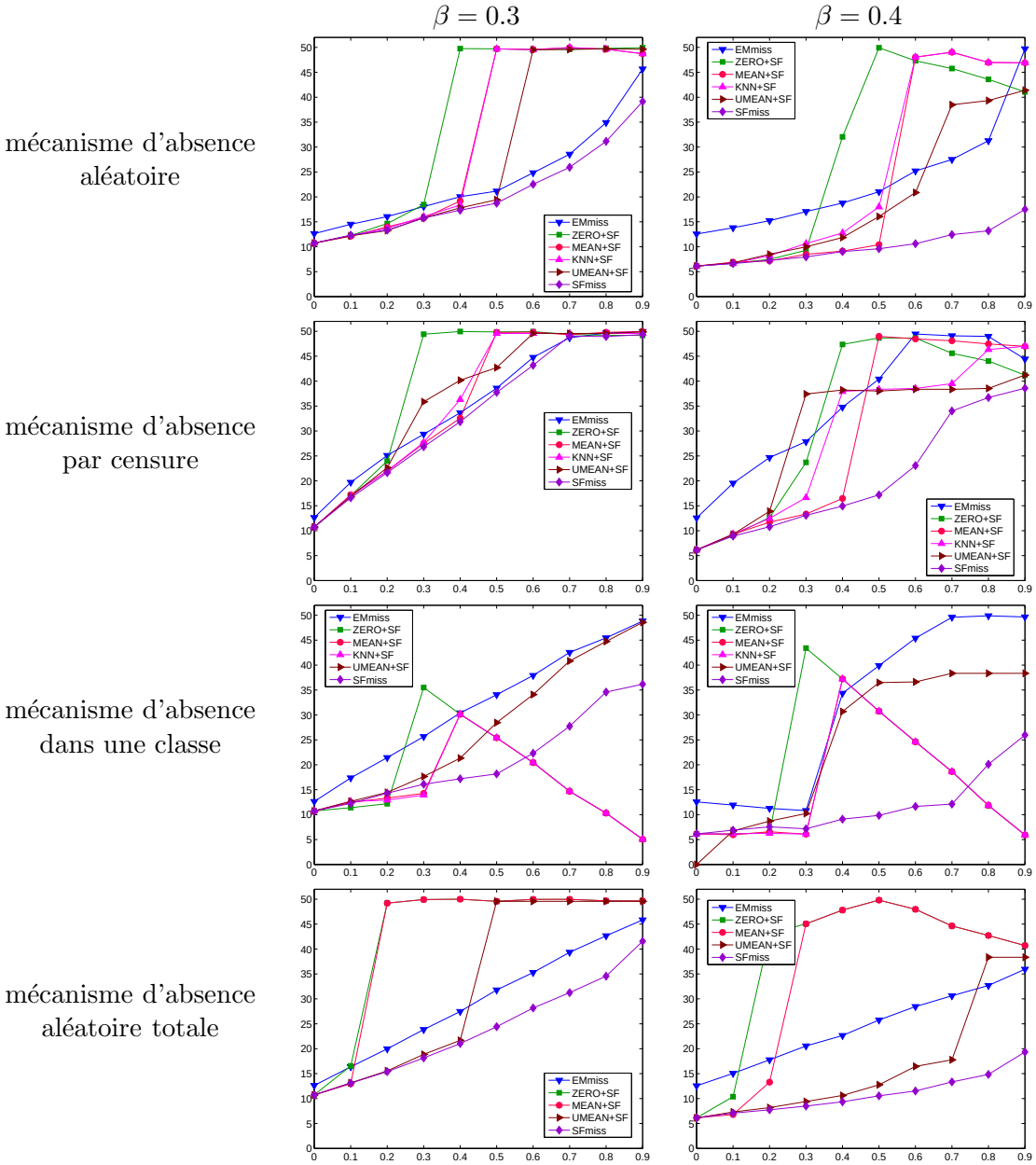


FIG. VIII.4 – Taux d'erreur de classification en fonction du paramètre ρ sur simulation d'un champ de Markov caché (modèle de Potts de paramètres $\beta = 0.3$ et 0.4 bruité par un bruit additif gaussien en dimension $D = 10$ de matrice de covariance $\Sigma = 2\mathbb{I}_{10}$).

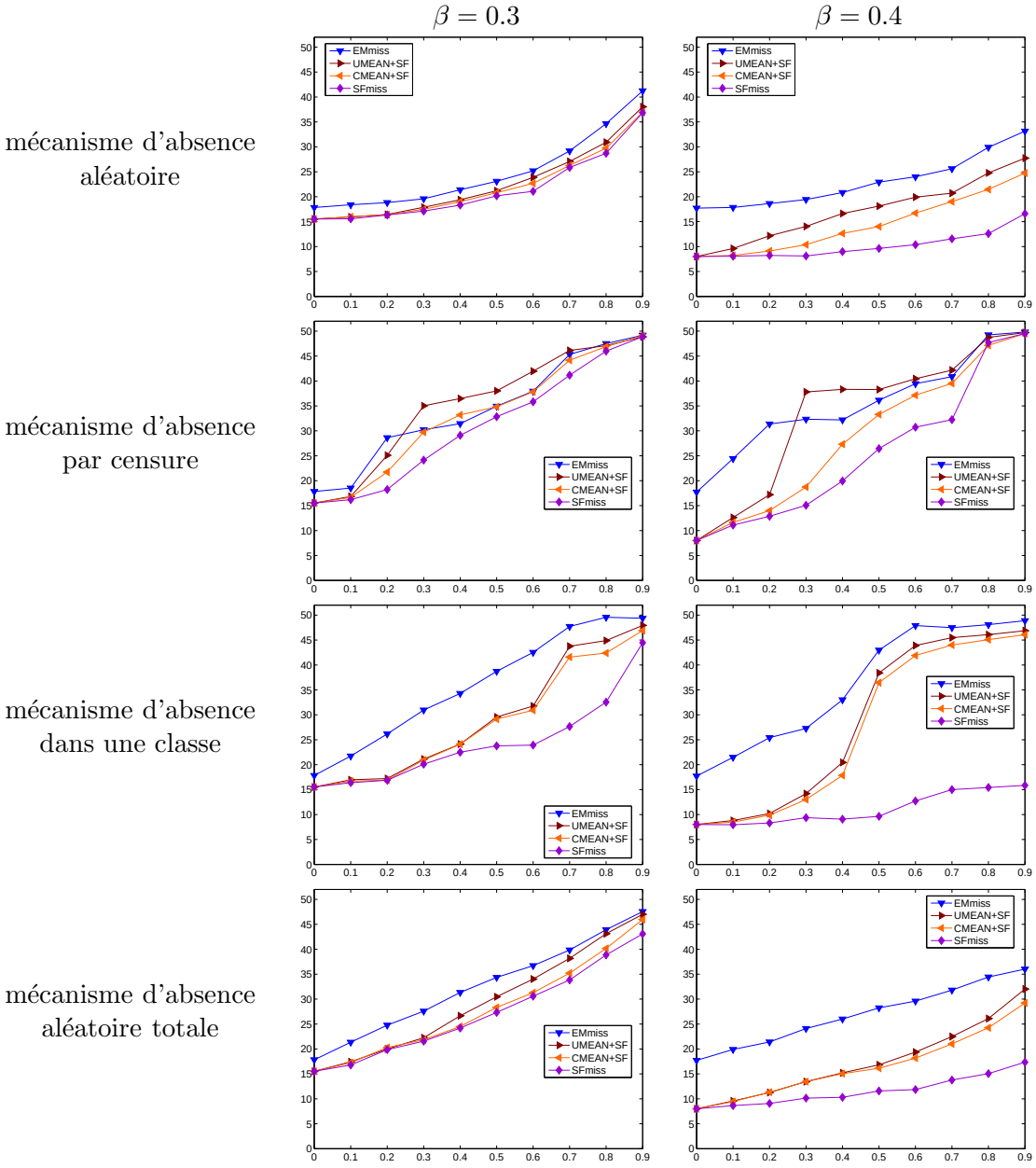


FIG. VIII.5 – Taux d'erreur de classification en fonction du paramètre ρ sur simulation d'un champ de Markov caché (modèle de Potts de paramètres $\beta = 0.3$ et 0.4 bruité par un bruit additif gaussien en dimension $D = 4$ de matrice de covariance Σ telle que $\Sigma_{dd} = 0.5$ et $\Sigma_{dd'} = 0.2$ pour $d \neq d'$).

à l'imputation par la moyenne (MEAN+SF), voire un peu moins bon. Notons qu'à partir d'un certain pourcentage de données manquantes (typiquement autour de 40%), le taux d'erreur avec ces deux types d'imputation explose. La méthode d'imputation *on-line* (au cours de l'algorithme d'estimation) par la moyenne inconditionnelle (UMEAN+SF) semble être plus stable, sauf sur données censurées. En effet, à cause de la censure, la variance empirique des données observées est plus faible que si toutes les données étaient observées (les $\mathbf{x^o}$ appartiennent nécessairement à l'intervalle $[q_{\frac{\rho}{2}}, q_{1-\frac{\rho}{2}}]$). Dès lors, une telle imputation ne peut "redresser" la variance des classes (absence du terme $\Gamma_{ik}^{dd'(q)}$ dans la mise à jour de $\Sigma_k^{dd'(q+1)}$, voir équation VII.12) et les valeurs imputées dans chaque classe seront plus proches de la moyenne que les vraies valeurs, accentuant de ce fait encore plus le biais de la variance des classes vers 0 et ainsi de suite. Notons encore que, dans le cas d'une matrice de covariance non diagonale, l'imputation par la moyenne conditionnelle (CMEAN+SF) est toujours meilleure. En effet, les valeurs imputées pour le calcul des paramètres de chaque classe tiennent alors compte des valeurs observées selon les autres dimensions, valeurs qui sont informatives. Mais de manière générale, l'algorithme sans imputation sous modélisation markovienne (SFmiss) surpasse les autres algorithmes et apparaît être le plus stable. Même pour 90% de données manquantes, le taux d'erreur n'explose pas.

Remarquons néanmoins que, lorsque les données manquantes proviennent toutes d'une même classe et à partir d'un certain pourcentage de données manquantes, les méthodes d'imputation par 0 ou la moyenne donnent de meilleurs résultats que toutes les autres méthodes (et notamment que SFmiss). On observe en effet, autour de 40% ou 50% de données manquantes, une forte décroissance des courbes d'erreur MEAN+SF et ZERO+SF. Cette décroissance s'explique par le fait qu'il y a alors assez de données imputées par la même valeur (0 ou la moyenne) pour les regrouper dans une même classe. Mais ce comportement est particulier à ce type de mécanisme d'absence.

Notons que le fait que l'algorithme SFmiss sans imputation soit plus performant qu'avec imputation par la moyenne inconditionnelle (UMEAN+SF) ou conditionnelle (CMEAN+SF) s'accorde avec la théorie : ces deux imputations reviennent à ne pas considérer le terme Γ_{ik} dans l'estimation de Σ_k (équation VII.12) et pour cette raison biaisent la variance vers zéro [61].

Enfin, sous modélisation par mélange indépendant, le comportement est relativement linéaire la plupart du temps. Les résultats sont toujours moins bons que sous modélisation markovienne (courbe SFmiss), et ce même lorsqu'aucune observation n'est manquante.

1.2 Image synthétique bruitée

Les résultats présentés sont ceux obtenus sur une image synthétique de damier de taille 128×128 et composée de 4 classes. Chacun des pixels est dupliqué D fois pour constituer une image en dimension D. Cette image est ensuite bruitée par ajout d'un bruit blanc gaussien D-dimensionnel. Nous présentons les résultats obtenus en dimension 1 avec $\sigma^2 = 0.25$ (Figure VIII.6), en dimension 10 avec $\Sigma = 2\mathbb{I}_{10}$ (Figure VIII.7) et en dimension 4 avec Σ non diagonale, de coefficient $\Sigma_{dd'} = 0.2$ si $d \neq d'$, $\Sigma_{dd} = 0.5$ (Figure

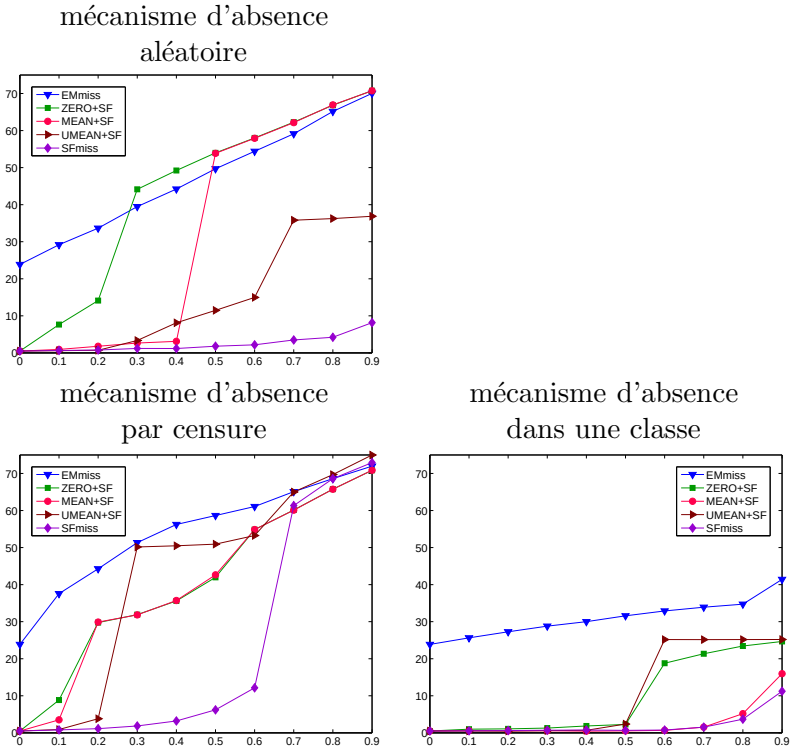


FIG. VIII.6 – Taux d'erreur de classification en fonction du paramètre ρ sur une image synthétique bruitée par un bruit additif gaussien en dimension $D = 1$ de variance $\sigma^2 = 0.25$.

VIII.8). Lors de la classification par modèle de champ de Markov caché, les observations sont supposées suivre un modèle de Potts de paramètre β à estimer, et défini sur un voisinage d'ordre 2. Pour illustration, les Figures VIII.9 à VIII.12 présentent les classifications obtenues avec l'algorithme SFmiss pour les 4 mécanismes d'absence considérés, dans le cas de la dimension 4.

On observe de manière générale que, sous modélisation markovienne, le problème de classification en présence de données manquantes est mieux résolu sur cette image synthétique que sur les données simulées de la section 1.1. En effet, la composante spatiale y est plus importante (β est estimé à 2.14), si bien que, sous hypothèse markovienne, l'estimation des paramètres et la classification sont plus aisées. C'est également pour cette raison que l'hypothèse de mélange indépendant (EMmiss) donne de bien pauvres résultats, même sans donnée manquante (plus de 20% d'erreur, contre moins de 1% sous hypothèse markovienne). Le comportement général des différentes méthodes (avec ou sans imputation) est semblable au cas des données simulées de la section 1.1. L'imputation par les k-plus proches voisins et la moyenne donnent des résultats relativement similaires et présentent la caractéristique d'exploser à partir d'un certain taux d'observations manquantes. L'imputation par la moyenne inconditionnelle est meilleure et plus stable sauf sur données censurées. Lorsque les différentes variables des observations sont corrélées (matrice Σ non diagonale), l'imputation par la moyenne conditionnelle est préférable à l'imputation par la moyenne inconditionnelle. Elle permet en effet, lors de l'estimation des paramètres, de prendre en compte les valeurs observées selon les autres dimensions,

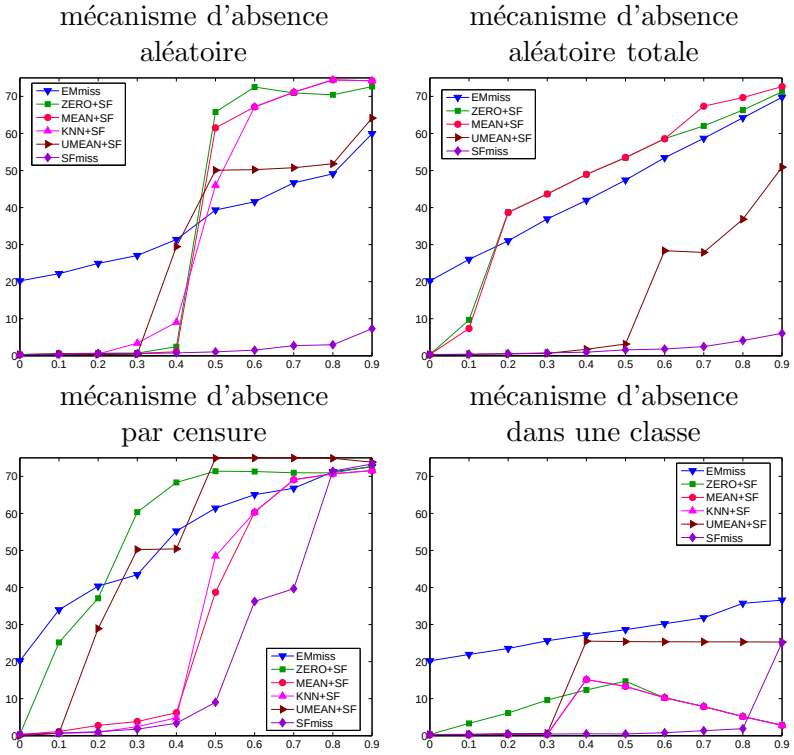


FIG. VIII.7 – Taux d'erreur de classification en fonction du paramètre ρ sur une image synthétique bruitée par un bruit additif gaussien en dimension $D = 10$ de matrice de covariance $\Sigma = 2\mathbb{I}_{10}$.

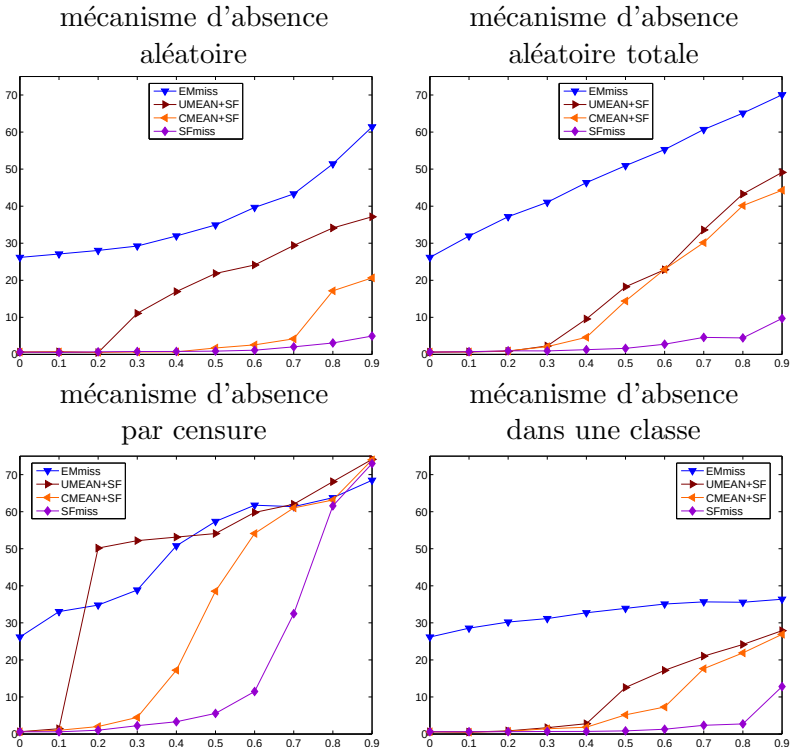


FIG. VIII.8 – Taux d'erreur de classification en fonction du paramètre ρ sur une image synthétique bruitée par un bruit additif gaussien en dimension $D = 4$ de matrice de covariance Σ telle que $\Sigma_{dd} = 0.5$ et $\Sigma_{dd'} = 0.2$ pour $d \neq d'$).

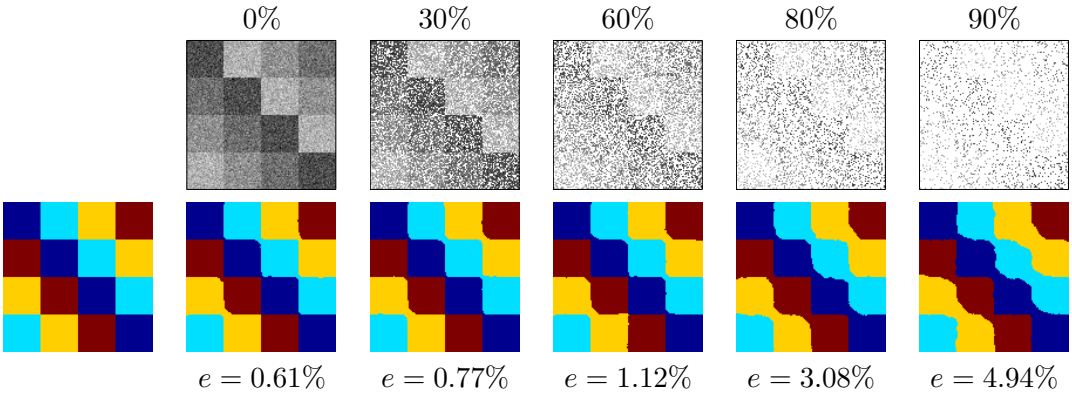


FIG. VIII.9 – Classification et taux d'erreur e en fonction du pourcentage d'observations manquantes en dimension 4 (Σ non diagonale), lorsque le mécanisme d'absence est aléatoire. Les images de la première ligne correspondent à la première dimension des données. Les valeurs manquantes apparaîssent en blanc.

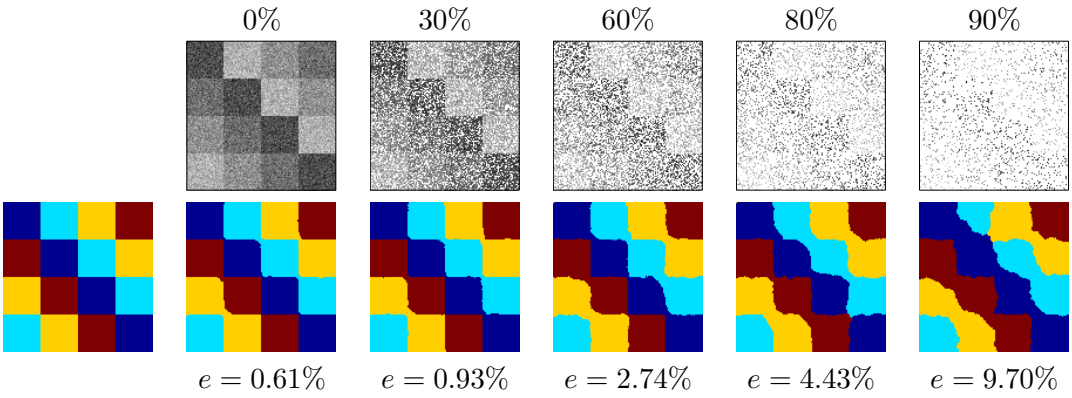


FIG. VIII.10 – Classification et taux d'erreur e en fonction du pourcentage d'observations totalement manquantes en dimension 4 (Σ non diagonale), pour le mécanisme d'absence aléatoire totale. Les images de la première ligne correspondent à la première dimension des données. Les valeurs manquantes apparaîssent en blanc.

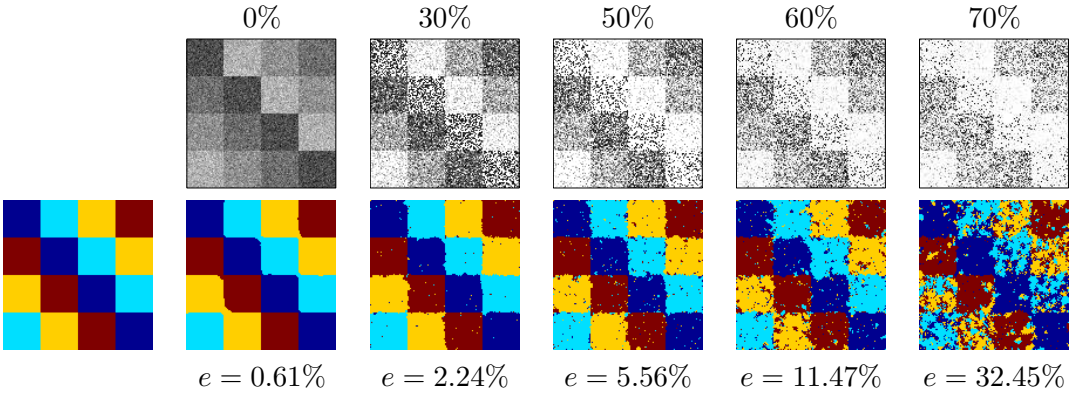


FIG. VIII.11 – Classification et taux d'erreur e en fonction du pourcentage d'observations censurées en dimension 4, pour le mécanisme d'absence par censure. Les images de la première ligne correspondent à la première dimension des données. Les valeurs manquantes apparaîssent en blanc.

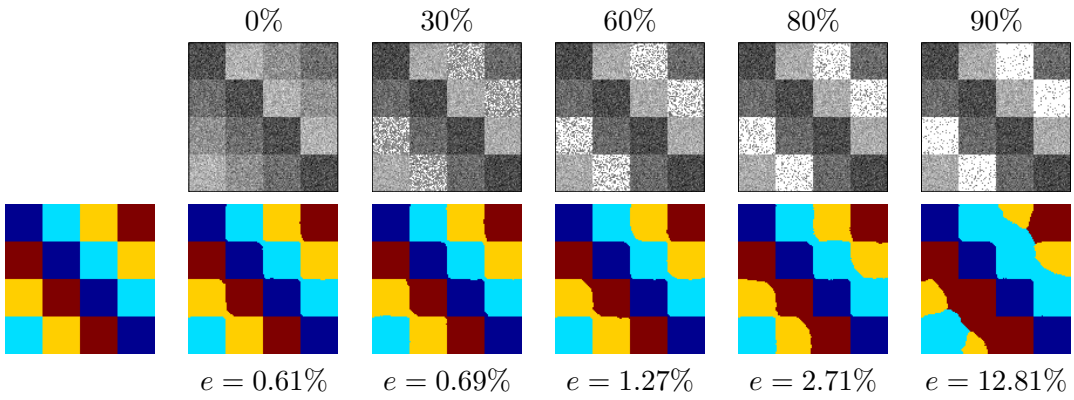


FIG. VIII.12 – Classification et taux d'erreur e en fonction du pourcentage d'observations manquantes dans la classe 3 (en bleu clair) en dimension 4 (Σ non diagonale), pour le mécanisme d'absence dans une classe. Les images de la première ligne correspondent à la première dimension des données. Les valeurs manquantes apparaîssent en blanc.

valeurs qui sont informatives. Enfin, la procédure la plus performante est SFmiss (sans imputation). Les résultats de classification sont très bons, même pour un très grand nombre de données manquantes (50% sur données censurées, 80%, voire 90% pour les autres mécanismes d'absence).

2 Sur données réelles : classification de données d'expression de gènes

Nous présentons dans cette section une application de notre méthode à la classification de gènes à partir d'observations de puces à ADN.

2.1 Problématique

La compréhension des mécanismes sous-jacents au fonctionnement du génome est un problème complexe que les biologistes cherchent à appréhender depuis de longues années. Elle doit permettre à terme de répondre à de très nombreuses problèmatiques, comme de trouver des cibles thérapeutiques, d'identifier des marqueurs pour un diagnostic rapide, de déterminer l'effet de la mutation d'un gène régulateur, d'analyser sous quelles conditions environnementales un gène est activé et comment il évolue, etc... Parmi les approches possibles, les nouvelles technologies telles que les puces à ADN permettent désormais d'accéder à des informations précises sur l'expression du génome. Elles permettent de suivre l'évolution du transcriptome pendant un temps déterminé sur les gènes étudiés. Les puces apportent donc des données quantitatives (une mesure numérique de l'expression de chaque gène à chaque instant). Il s'avère alors utile de regrouper les gènes en fonction de leur profil d'expression, cela dans le but d'identifier les relations qui existent entre ces gènes et donc de comprendre le fonctionnement du réseau de régulation sous-jacent.

D'autre part, l'analyse du protéome (ensemble des protéines produites par un génome) à grande échelle est devenue possible grâce aux évolutions récentes de technologies telles que l'électrophorèse bidimensionnelle et la spectrométrie de masse. Ces méthodes nous

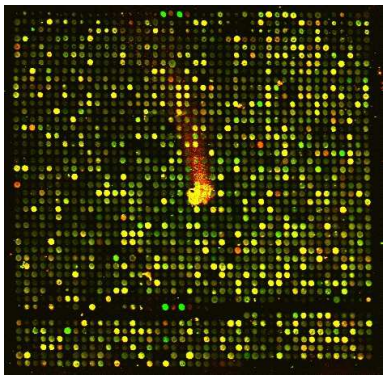


FIG. VIII.13 – Exemple de puce à ADN. Aux spots centraux de l'image correspondront des données manquantes.

permettent d'identifier les différents partenaires d'un complexe, et de ce fait de définir des associations entre protéines. Ces associations peuvent se référer à un lien physique (deux protéines orthologues ou ayant fusionnées dans un organisme par exemple), ou à une interaction plus indirecte, comme la participation à une même voie métabolique ou à un même processus cellulaire. De nombreuses banques de données d'interactions entre protéines sont disponibles sur la Toile. Citons entre autre BIND (`http://bind.ca`), MINT (`http://mint.bio.uniroma2.it/mint`), STRING (`http://string.embl.de/`), DIP (`http://dip.doe-mbi.ucla.edu/`). Or, il est naturel de penser que le fait de disposer de ces données protéomiques à grande échelle peut nous aider à accéder à un niveau de compréhension globale de l'organisme. En particulier, utiliser ces réseaux d'interactions entre protéines peut faciliter la tâche de classification de gènes.

Si les puces à ADN permettent de mesurer et de visualiser très facilement les différences d'expression entre les gènes, plusieurs raisons expliquent le fait que les expériences ne fournissent pas de données complètes [123] : des problèmes lors de l'étape d'hybridation, de la poussière ou des éraflures sur la lame, du bruit sur l'image à analyser ou une résolution insuffisante, une erreur systématique du robot qui dépose les sondes etc... [96] rappelle de plus que ces données manquantes affectent en général de l'ordre de 90% des gènes. Un exemple de puce à ADN est donné en Figure VIII.13. Au centre de l'image, on peut observer une anomalie si bien que les observations des spots centraux seront manquantes.

L'objectif de ce travail est donc de classer des gènes à partir de leurs données d'expression issues de puces ADN, certaines de ces données étant manquantes. Pour ce faire, nous nous aidons de la connaissance *a priori* que nous avons du protéome. Plus exactement, nous supposons les gènes en interaction les uns avec les autres via une structure de graphe : deux gènes dépendent directement l'un de l'autre (ce qui se traduit par la présence d'une arête entre ces deux gènes dans le graphe) si les protéines qu'ils codent sont elles-même associées. Notons que le problème de la classification de gènes sous modélisation markovienne a déjà été abordé, notamment dans [126], mais à partir de données complètes uniquement. Nous proposons ici de compléter cette étude en permettant de prendre en compte les gènes pour lesquels des données manquent, et ce, sans faire appel à des méthodes d'imputation.

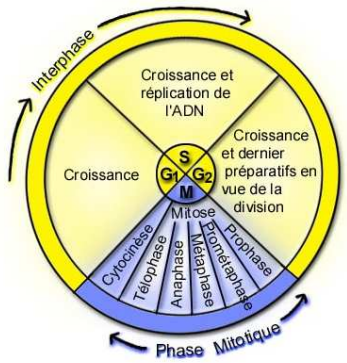


FIG. VIII.14 – Le cycle cellulaire comprend la mitose, durant laquelle les chromosomes sont séparés et le cytoplasme est divisé, et l'interphase, durant laquelle la majorité de la croissance cellulaire, des activités métaboliques et la réplication des chromosomes se fait.

2.2 Les données

Les données sont celles décrites dans [117], correspondant aux données d'expression de gènes du cycle cellulaire de *Saccharomyces cerevisiae* (levure du boulanger). Rappelons que le cycle cellulaire est divisé en plusieurs phases (Figure VIII.14) :

- la phase G1 (*Growth 1*), première phase de croissance pendant laquelle les synthèses d'ARN sont très actives,
- la phase S (*Synthesis*) durant laquelle le matériel génétique est doublé par réplication de chacun des chromosomes,
- la phase G2 (*Growth 2*), deuxième phase de croissance, à l'issue de laquelle chaque chromosome est constitué de deux chromatides morphologiquement et génétiquement identiques et unies par leur centromère,
- la phase M (*Mitosis*), celle de la division cellulaire proprement dite, au cours de laquelle les chromatides se séparent et se répartissent dans chacun des noyaux fils

Ce cycle cellulaire est un phénomène continu aboutissant, à partir d'une cellule mère diploïde, à deux cellules filles diploïdes.

Les données sur le cycle cellulaire de [117] ont été obtenues à partir de cultures cellulaires synchronisées par quatre méthodes indépendantes : Alpha, cdc15, cdc28 et Elutriation (voir Figure VIII.15). Nous nous sommes intéressés aux données correspondant à l'expérience "cdc28" de [35], provenant de cultures cellulaires synchronisées au stade G1 tardif de la division cellulaire par l'emploi de souches mutantes cdc28. 612 gènes sont observés toutes les 10 minutes durant 160 minutes. Chaque observation est donc en dimension 17.

Le réseau d'interaction que nous avons choisi d'utiliser a été obtenu à partir de la base de données STRING (`http://string.embl.de/`) [127] et est représenté en Figure VIII.16 (a). Il est constitué de 3530 arêtes (11.5 voisins en moyenne par noeuds). 41 gènes sont non connectés (ils n'ont donc aucun voisin) et un gène possède 120 voisins (voir Figure VIII.16 (b)). Le diamètre (plus grande distance entre deux gènes) est égale à 7. Il y a 5% de données manquantes, ce qui affecte plus de 80% des gènes.

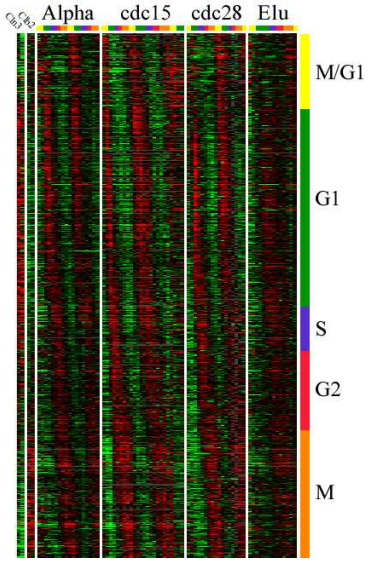
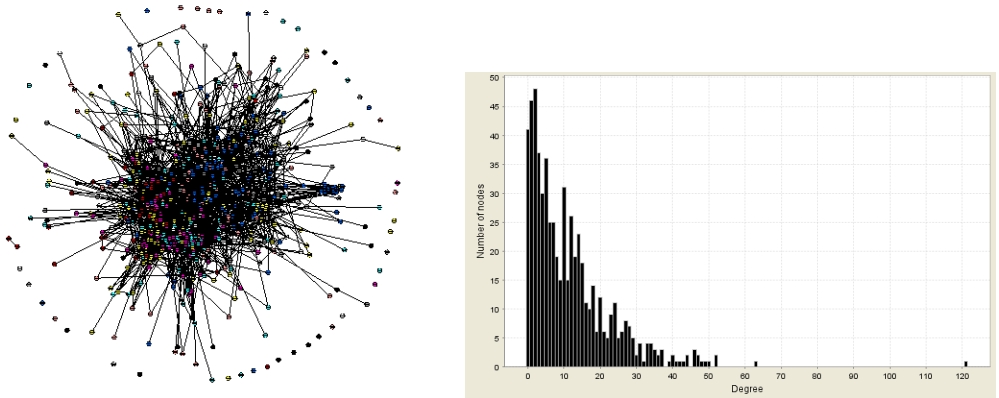


FIG. VIII.15 – Données d'expression de gènes durant le cycle cellulaire de la levure [117]. Chaque ligne correspond à un gène, et chaque colonne à un instant de l'expérience. Une forte couleur rouge (respectivement verte) indique que le gène est fortement exprimé (respectivement réprimé). Les données manquantes sont représentées par la couleur grise. A droite de la figure, la barre de couleur indique à quelle phase de la division cellulaire appartiennent chacun des gènes.



(a) (b)

FIG. VIII.16 – (a) Graphe d'interaction sur les 612 gènes du cycle cellulaire de *Saccharomyces cerevisiae* (b) Nombre de noeuds du graphe en fonction de leurs degrés (nombre de voisins).

2.3 Résultats

2.3.1 Modèle

Les observations $\mathbf{x} = (x_1, \cdots, x_n)$ ($n = 612$) considérées sont les données d'expression des 612 gènes étudiés. Chaque observation x_i est en dimension $D = 17$ correspondant à différents instants de l'expérience. Certaines des valeurs x_{id} ($d \in [\![1, 17]\!]$) sont manquantes.

Dans ce qui suit, nous considèrerons deux modèles :

- Une modélisation par champ de Markov caché à bruit indépendant. Les gènes (ou sites) sont supposés être régis par le graphe de la Figure VIII.16 (a), définissant de ce fait une structure de voisinage. De plus, la distribution markovienne des classes est modélisée par un modèle de Potts paramétré par un unique scalaire β.
- Une modélisation par mélange indépendant sous laquelle les gènes sont supposés indépendants les uns des autres.

L'objectif de la classification est alors d'associer à chaque gène i une classe $z_i \in [\![1, K]\!]$.

2.3.2 Sélection du nombre de classes

La première question à se poser est le nombre de classes à utiliser. Sur cette question, il n'est pas possible de répondre *a priori*, c'est-à-dire sans connaissance particulière sur les groupes que l'on aimerait avoir.

Nous avons donc choisi de sélectionner le nombre de classes le plus approprié à l'aide du critère BIC et de son approximation en champ moyen BIC^w (voir chapitre II, section 5.2). Les courbes correspondant à la valeur de ce critère en fonction du nombre de classes K sont données en Figure VIII.17 pour différentes méthodes d'estimation :

- SFmiss : l'algorithme en champ simulé avec données manquantes sous modélisation par champ de Markov caché et bruit indépendant, comme décrit au chapitre VII, section 3.
- EMmiss : l'algorithme EM avec données manquantes sous modélisation par modèle de mélange indépendant, comme décrit au chapitre VII, section 2.
- KNN+SF : l'algorithme en champ simulé sur données complétées en pré-traitement par la technique des k-plus proches voisins.
- PREV+SF : l'algorithme en champ simulé sur données complétées en pré-traitement par la valeur observé à l'instant précédent, ce qui revient à remplacer une observation manquante à un temps t donné par l'expression de ce gène au temps $t - 1$.

Il est intéressant de remarquer que toutes les méthodes présentent une décroissance, plus ou moins marquée, du critère entre $K = 9$ et $K = 10$, ce qui nous suggère de sélectionner un nombre de classe égal à 9. Notons encore que, s'il fallait choisir une méthode à nombre de classes fixé, c'est toujours l'algorithme SFmiss qui serait préféré puisque, pour toutes les valeurs de K, sa courbe se situe au dessus des autres. L'écart entre les courbes SFmiss et EMmiss était attendu. D'une part on peut penser que prendre en compte le réseau, donc de la connaissance supplémentaire sur des "similarités" entre gènes, aide à la classification. D'autre part la pénalisation dans le critère BIC par $\nu \log n$ (où ν désigne le nombre de paramètres du modèle) défavorise le modèle indépendant par rapport au modèle de Potts caché. En effet, pour K classes, il y a $K - 1$ proportions du mélange a estimer, contre seulement 1 paramètre (le scalaire β) pour le modèle de Potts. L'écart entre la méthode SFmiss sans imputation et les méthodes KNN+SF et PREV+SF nous assurent que les paramètres estimés par notre algorithme sont plus vraisemblables

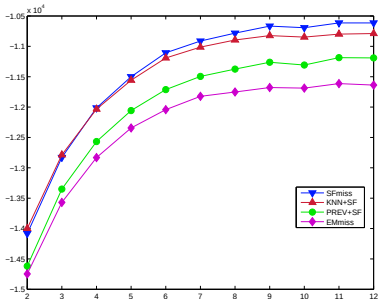


FIG. VIII.17 – Valeur du critère BIC^w en fonction du nombre de classe K (de 2 à 12) pour différents méthodes avec ou sans imputation. Cette courbe nous suggère de sélectionner $K = 9$.

que ceux avec imputation. En effet, pour ces méthodes, le modèle sous-jacent est le même (celui d'une champ de Markov caché à bruit indépendant), si bien que les pénalisations dans le critère BIC sont rigoureusement égales.

2.3.3 Classifications

En classification non supervisée de gènes, juger de la qualité des groupes produits n'est pas chose facile et il n'existe aucun critère qui fasse consensus. Néanmoins, une approche largement utilisée est de s'aider du *Gene Onthology Consortium* (GO) (`http://www.geneonthology.org`). Il s'agit d'une gigantesque base de données standardisée indiquant, pour chaque gène, sa fonction moléculaire, dans quel processus biologique il intervient et dans quel composant cellulaire. Pour chacune de ces 3 catégories, la nomenclature se présente sous la forme d'un arbre (une partie de l'arbre des processus biologiques se trouve en Figure VIII.18). Plus une fonction se trouve à une profondeur élevée, plus elle correspond à un processus particulier. Par exemple, la fonction "Assimilation de sulfate" se trouve à une profondeur de 7, alors que la fonction "Métabolisme", à une profondeur de 3. Cette nomenclature peut notamment être utilisée pour l'analyse des profils d'expression de gènes, via le calcul de p-valeurs. Pour une fonction donnée (par exemple l'assimilation de sulfate), plus la p-valeur est faible pour une classe, plus cette fonction est sur-représentée dans cette classe en comparaison aux autres classes.

La Table VIII.1 donne un résumé des résultats de l'analyse sur les termes GO. Le test utilisé est le FDR (*False Discovery Rate*). Brièvement, ce test calcule la proportion de faux positifs d'une classe k donnée (les gènes présentant cette fonction mais classés dans $k' \neq k$) par rapport aux vrais positifs (les gènes présentant cette fonction et classés dans k). Il y apparaît clairement que la classification obtenue par l'algorithme SFmiss est plus fine. En effet, les p-valeurs des termes GO sont en règle générale plus faibles sur la classification par SFmiss que par les autres algorithmes, signifiant par là que cette procédure regroupe un plus grand nombre de gènes aux fonctionnalités similaires dans un même groupe. Signalons tout de même que, dans la Table VIII.1, 2 termes GO semblent mieux représentés dans la classification par KNN+SF. Néanmoins, les différences de p-valeurs avec SFmiss sont faibles.

Cet aspect est confirmé par une comparaison globale des classes formées et du réseau

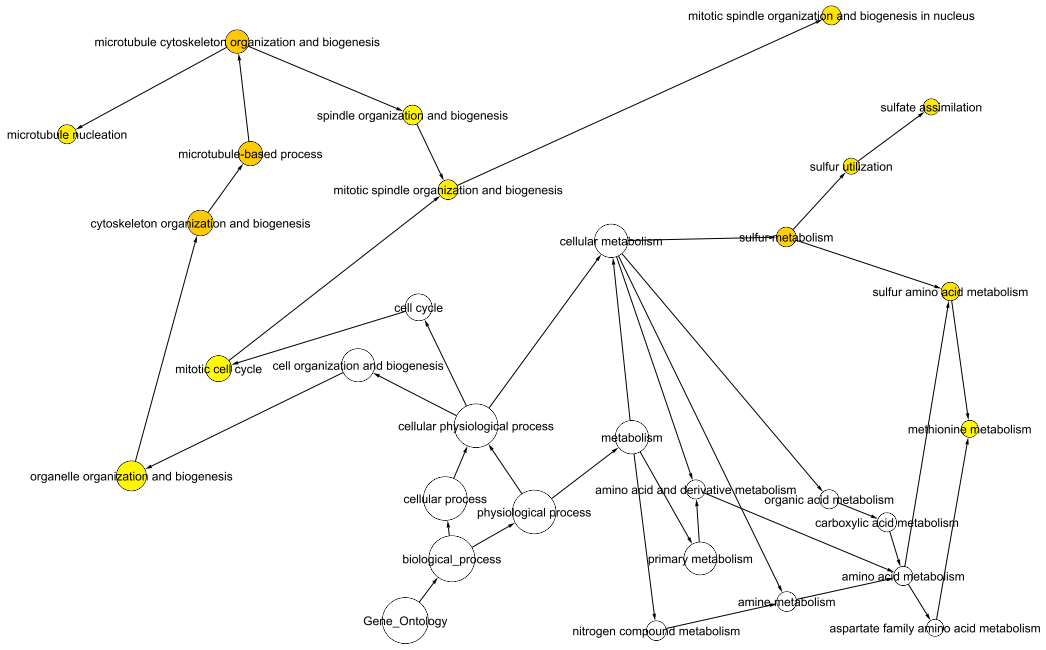


FIG. VIII.18 – Une partie de l'arbre des processus biologiques selon la nomenclature de GO.

d'interaction. En effet, un comportement général de l'algorithme est que la structure du graphe est mieux conservée, c'est-à-dire que des gènes voisins (reliés dans le graphe) ont plus tendance à être groupés dans une même classe. Une illustration de ce phénomène est donnée en Figure VIII.19. Sur le zoom de droite, on observe que 11 gènes sur les 21 appartiennent à une même classe (coloriée en bleue) avec EMmiss, 14 avec KNN+SF et 17 avec SFmiss. Notons que le fait que l'algorithme EMmiss ne regroupe pas spécifiquement les gènes voisins dans une même classe était tout à fait prévisible, le modèle de mélange stipulant que les gènes sont indépendants les uns des autres. Par contre, il est intéressant de remarquer que, lorsque on impute les données au préalable (par KNN), la classification résultante perd en qualité (en comparaison avec SFmiss), bien que le modèle sous jacent soit le même (celui d'une champ de Markov caché).

Notons encore que les classes obtenues peuvent être interprétées en terme de classes temporelles. En effet, considérons les 5 classes de [117] relatifs aux différentes étapes de la division cellulaire, à savoir, chronologiquement, M/G1 (transition entre M et G1), G1, S, S/G2 (transition entre S et G2), G2/M (transition entre G2 et M) La répartition croisée des gènes des classes temporelles M/G1, G1, S, S/G2 et G2/M avec les 9 classes de SFmiss est donnée en Table VIII.2. Il ressort de ce tableau que 6 classes peuvent être interprétées biologiquement, correspondant à un total de 406 gènes (en gras dans le tableau) :

- la classe $k = 1$ regroupe des gènes régulés lors de la transition G2/M
- la classe $k = 2$ regroupe des gènes régulés lors de la phase G1
- la classe $k = 4$ regroupe des gènes régulés lors des phases M et G1
- la classe $k = 5$ regroupe des gènes régulés lors de la phase S large
- la classe $k = 6$ regroupe des gènes régulés lors de la phase G1
- la classe $k = 9$ regroupe des gènes régulés lors de la phase M

n^o de classe	p-valeur pour SFmiss	p-valeur pour EMmiss	p-valeur pour KNN+SF
k=4	GO :0006732, coenzyme met. process		
	$\mathbf{1.1\,10^{-2}}$	> 0.1	> 0.1
k=5	GO :0005819, spindle		
	$\mathbf{4.6\,10^{-9}}$	$6.7\,10^{-7}$	$2.0\,10^{-6}$
	GO :0006790, sulf. met. process		
	$\mathbf{1.1\,10^{-4}}$	$2.4\,10^{-4}$	$8.7\,10^{-4}$
	GO :0000278, mitotic cell cycle		
	$\mathbf{2.2\,10^{-3}}$	$7.7\,10^{-3}$	> 0.1
	GO :0030472, mit. spin. org. & biogen. in nucleus		
	$\mathbf{5.2\,10^{-3}}$	$8.8\,10^{-3}$	$2.0\,10^{-2}$
k=6	GO :0006974, resp. to DNA dam. stim.		
	$\mathbf{1.8\,10^{-3}}$	$3.0\,10^{-3}$	$8.0\,10^{-3}$
	GO :0000724, dbl-str. bk rep. via hom. comb.		
	$\mathbf{1.9\,10^{-2}}$	$2.7\,10^{-2}$	$4.6\,10^{-2}$
	GO :0000030, mannosyltransf. act.		
	$\mathbf{1.1\,10^{-2}}$	$1.2\,10^{-2}$	$2.7\,10^{-2}$
k=9	GO :0042555, MCM cplx		
	$\mathbf{3.4\,10^{-4}}$	$8.3\,10^{-4}$	$4.0\,10^{-4}$
	GO :0008026, ATP-dep. helicase act.		
	$5.5\,10^{-4}$	$1.3\,10^{-3}$	$\mathbf{4.5\,10^{-4}}$
	GO :0006268, DNA unwind. replic.		
	$2.8\,10^{-3}$	$6.7\,10^{-3}$	$\mathbf{1.1\,10^{-3}}$
	GO :0042623, ATPase act. coupl.		
	$\mathbf{4.4\,10^{-3}}$	$1.5\,10^{-2}$	$4.3\,10^{-2}$

TAB. VIII.1 – Quelques termes GO représentatitfs de l'analyse des classes produites par les différentes méthodes. En gras, la meilleure p-valeur (correspondant à la plus faible) de chaque fonction considérée.

Les réseaux internes à ces 6 "bonnes" classes sont visibles en Figure VIII.21. Les couleurs données aux arcs sont fonctions du type d'interaction protéine-protéine considérée pour la création du graphe. Enfin, on donne en Figure VIII.20 les profils des gènes de chacune des 9 classes. Les classes $k = 7$ et $k = 8$ peuvent difficilement être interprétées car elles contiennent des gènes de toutes les classes temporelles. Notons que le profil moyen de ces classes (en noire sur la Figure VIII.20) est relativement plat. Pour comparaison, remarquons que les 6 classes 1, 2, 4, 5, 6 et 9 présentent un profil à 2 "périodes" caractéristique du fait que l'expérience corresponde à 2 cycles cellulaires. Enfin, la classe $k = 3$ présente un pic en $t = 10$ (ce qui correspond à la phase S) mais le reste du profil de la classe est relativement plat.

Remarque. *Nous avons également testé les méthodes d'imputation par des zéros ou la moyenne. Les résultats obtenus par ces méthodes sont très mauvais, même sous hypothèse markovienne, car tous les gènes (ou presque) ayant une valeur manquante selon la 15ème variable sont classés dans une même classe, indépendamment de leurs profils (voir Figure VIII.22). La 15ème variable (correspondant au 15ème temps de l'expérience) est celle*

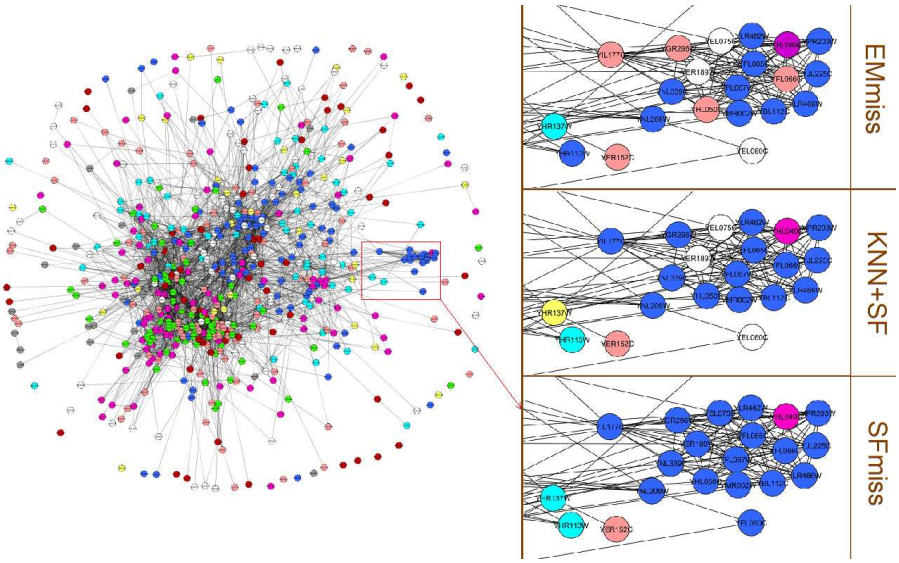


FIG. VIII.19 – A gauche : le graphe dans son intégralité, colorié en fonction des groupes de SFmiss. A droite : zoom sur une partie du graphe et comparaison des groupements effectués avec les algorithmes EMmiss et KNN+SF.

	k = 1	**k = 2**	$k = 3$	**k = 4**	**k = 5**	**k = 6**	$k = 7$	$k = 8$	**k = 9**
M/G1	**0**	**2**	28	**27**	**0**	**2**	1	11	**13**
G1	**0**	**69**	34	**14**	**20**	**84**	10	9	**0**
S	**1**	**0**	7	**1**	**29**	**5**	13	4	**0**
S/G2	**2**	**0**	14	**3**	**38**	**0**	31	0	**0**
G2/M	**33**	**0**	18	**16**	**1**	**0**	20	6	**46**

TAB. VIII.2 – Tableau croisé de répartition des gènes des classes temporelles M/G1, G1, S, S/G2 et G2/M selon les 9 classes.

possédant le plus de valeurs manquantes, 207 *au total.*

2.3.4 Etude de stabilité

Afin d'étudier le comportement de ces méthodes lorsque plus de données viennent à manquer, nous avons retiré aléatoirement un certain nombre d'observations (ce qui correspond à un processus MCAR). En Figure VIII.23 sont présentées les courbes du pourcentage de gènes dont la classification a changé en fonction du pourcentage d'observations supplémentaires manquantes. La comparaison est faite par rapport aux classifications obtenues selon les mêmes méthodes sur les données initiales. De ce fait, toutes les courbes commencent à 0. Les courbes "s'arrêtent" lorsque les classifications sont trop différentes des classifications initiales pour pouvoir être comparées avec elles (les "erreurs" de classification sont alors $> 30\%$). On observe que la méthode la moins stable est la méthode où les gènes sont supposés indépendants les uns des autres (notée EMmiss). Sous modélisation markovienne, la méthode la plus stable est celle sans imputation (notée SFmiss). Avec imputation, après 4% de données manquantes supplémentaires (soit 9% au

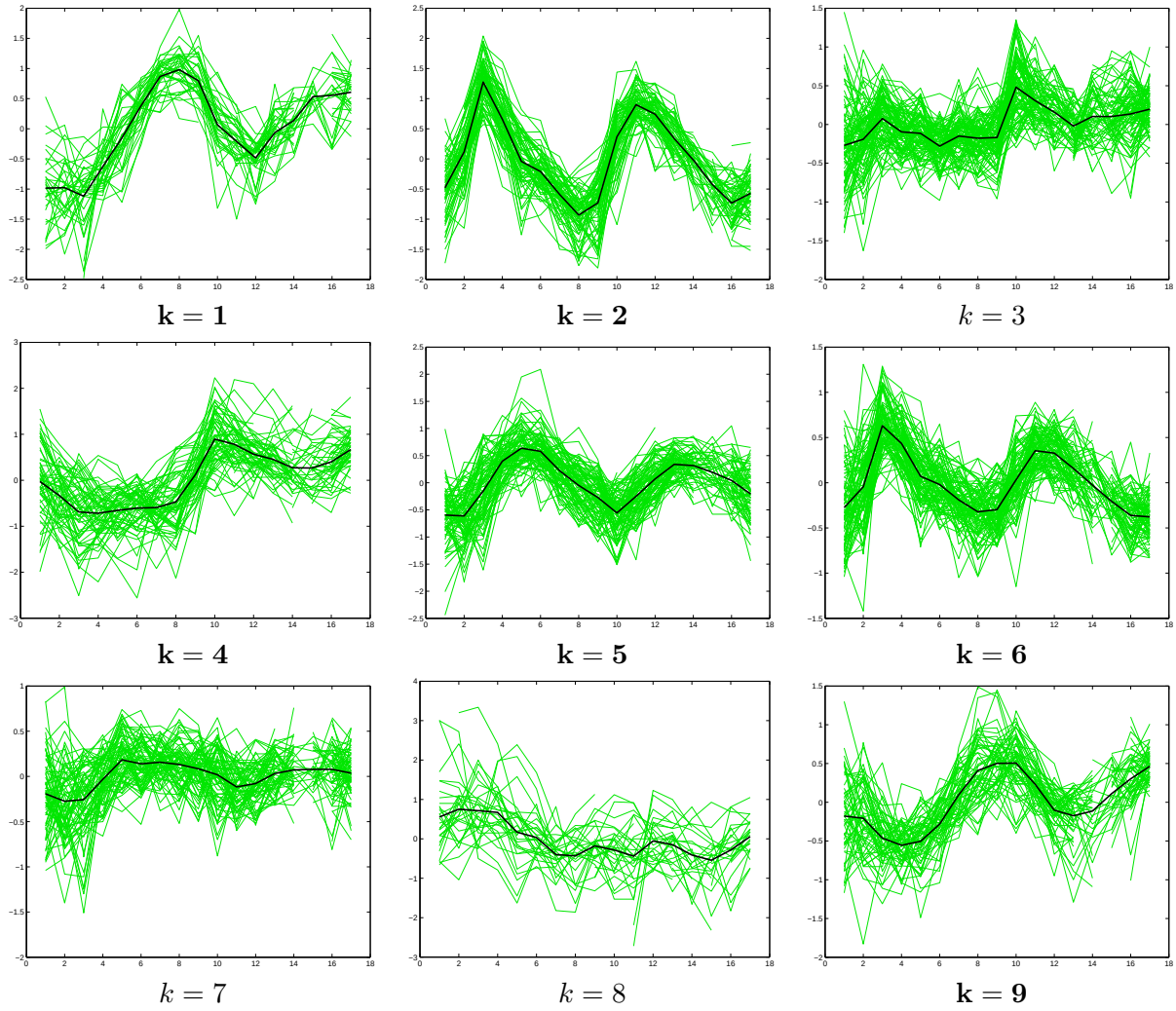
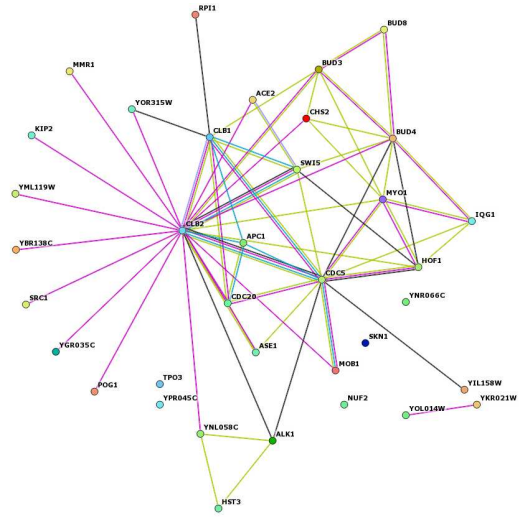
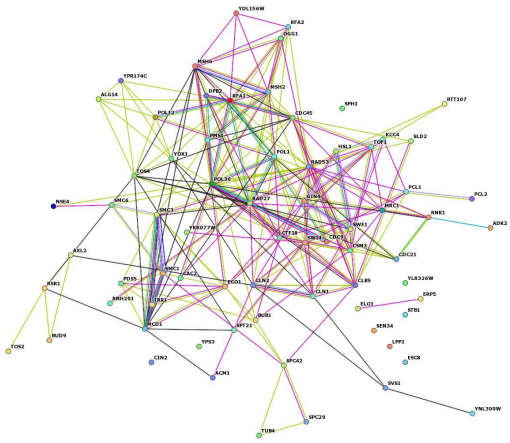


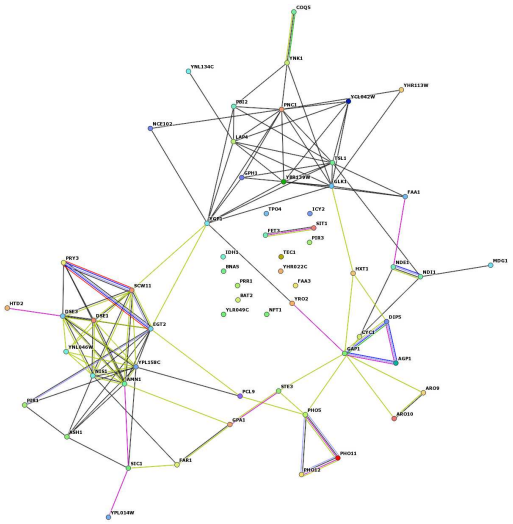
FIG. VIII.20 – Profil des expressions au cours du temps pour les 9 classes. En noire, le profil moyen de ces classes.



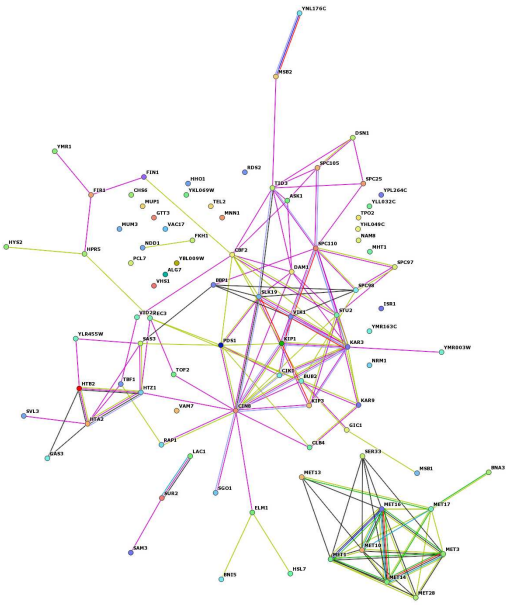
$k = 1$



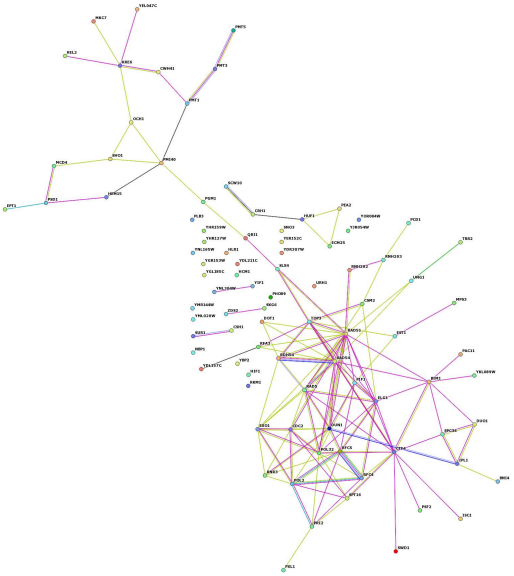
$k = 2$



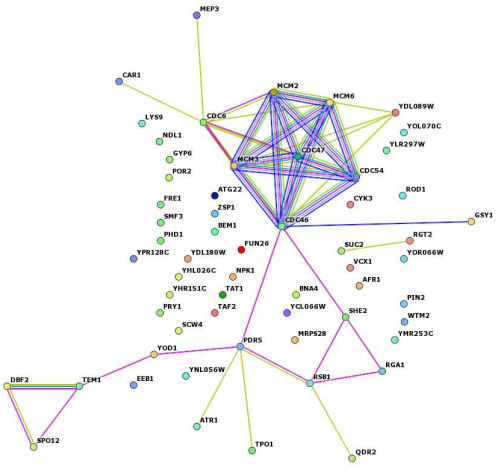
$k = 4$



$k = 5$



$k = 6$



$k = 9$

FIG. VIII.21 – Réseau interne aux 6 classes.

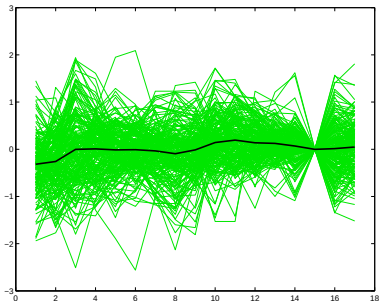


FIG. VIII.22 – Profil d'expression des gènes d'une classe avec imputation selon la moyenne. Cette classe contient tous (ou presque tous) les gènes dont la 15ème variable est manquante, indépendamment de leur profil.

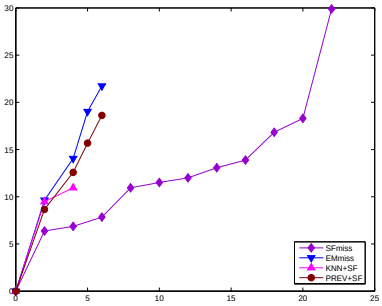


FIG. VIII.23 – Pourcentage de gènes dont la classification a changé par rapport à la classification sur les données initiales, en fonction du pourcentage de données manquantes aléatoirement (processus MCAR).

total), la méthode selon les k-plus proches voisins (KNN+SF) ne peut plus être utilisée car trop peu de gènes sont entièrement observés (c'est-à-dire sans valeur manquante). L'imputation selon la valeur précédente (PREV+SF), si elle donne des classifications similaires au cas sans imputation sur les données initiales, est beaucoup moins stable que cette dernière. Après 6% de données manquantes supplémentaires (soit 11% au total), la classification change tout à fait. On comprend en effet que, pour un gène de profil non plat, plus le nombre de composantes manquantes de ce dernier est important, plus le profil imputé en sera modifié. Entre autre, si le gène a un pic d'expression en t et si cette composante t manque, la valeur imputée correspondra à celle à $t-1$ qui peut être très loin de la valeur du pic en t.

2.4 Conclusion

On peut conclure de l'ensemble de ces expériences que la prise en compte, via une structure de voisinage, des dépendances entre gènes ainsi que l'utilisation d'un modèle pour les données manquantes permet d'améliorer la classification. De plus, sous hypothèse de markovianité, la méthode développée au chapitre VII, section 3, est la plus stable. Un prolongement intéressant à ce travail serait d'étudier si l'introduction de poids sur les arêtes (donnant plus ou moins d'importance à celles-ci) permettrait d'améliorer la qualité

des classifications. En effet, sur les résultats présentés, toutes les arêtes ont le même poids, donc la même importance. Or, lors de la création du graphe à partir de la base de données STRING, différents types d'associations sont considérés (si les gènes ont un lien physique, s'ils participent à une même voie métabolique, à un même processus cellulaire...). Deux protéines seront alors associées (ce qui se traduit par un arc dans le réseau entre les deux gènes correspondants) si une des ces associations est jugée "sûre" (le niveau de confiance pouvant être réglé par l'utilisateur). Il pourrait alors être intéressant de donner un poids plus grand aux arêtes pour lesquelles plusieurs associations ont été jugées sûres. Notons que de tels poids peuvent facilement être pris en compte dans une modélisation markovienne. Il suffit en effet d'introduire un coefficient de pondération w_{ij} dans le terme de l'énergie correspondant à l'interaction entre des sites i et j voisins.

Conclusion et perspectives

Bilan

Les travaux présentés dans ce mémoire abordent le problème de la classification d'individus sous des hypothèses non classiques. Tout d'abord, les individus considérés ne sont pas indépendants les uns des autres. Plus précisément, les dépendances sont définies à l'aide d'un système de voisinage ou encore un graphe d'interaction. L'approche adoptée est alors une approche probabiliste fondée sur une modélisation markovienne. Nous avons abordé le problème de la classification lorsque les individus sont soumis à trois sources de complexité.

La première source de complexité abordée concerne la grande dimension des observations, comme c'est de plus en plus le cas dans nombre d'applications modernes. A partir du modèle pour la grande dimension de C. Bouveyron [23], nous avons proposé une procédure d'estimation et classification basée sur l'approximation en champ moyen et l'algorithme EM. Cette procédure permet la prise en compte des dépendances entre composantes d'une observation tout en ne demandant que l'estimation d'un nombre raisonnable de paramètres.

La deuxième source de complexité abordée est plus générale. Elle concerne l'hypothèse très largement utilisée de bruit indépendant qui, dans de nombreuses applications (notamment la segmentation d'images texturées ou d'images non stationnaires) ne peut être un modèle efficace. A partir des modèles de Markov triplets de D. Benboudjema et W. Pieczynski [6], nous avons proposé une famille de modèles triplets adaptés à la classification supervisée et permettant de lever l'hypothèse de bruit indépendant lorsqu'elle est trop restrictive. Le schéma de classification que nous proposons est général, tous les traitement classiques pouvant être utilisés pour l'estimation des paramètres. En particulier, nous avons illustré les étapes d'apprentissage et de test en nous basant sur des approximations de type champ moyen et l'algorithme NREM, originellement proposé dans [29] pour le modèle classique de champ de Markov caché à bruit indépendant. Nous avons adapté cet algorithme à nos modèles de Markov triplets. Nous avons encore étudié la possibilité de sélectionner, parmi la famille de champs de Markov triplets proposée, le modèle le plus adapté aux données. A l'aide de simulations, nous avons également mis en évidence un phénomène de transition de phase associé à ces modèles. Enfin, nous avons appliqué notre schéma à la classification supervisée d'images de textures décrites à l'aide de descripteurs locaux de grande dimension. Les résultats obtenus ont mis en évidence les performances de notre modèle, sa flexibilité, ainsi que l'intérêt d'un modèle pour la grande dimension. Mentionnons qu'une approche alternative permettant de lever l'hypothèse de bruit indépendant serait l'utilisation de champs conditionnels (*Conditional Random Fields*) [79],

largement utilisés ces dernières années, notamment dans le domaine de la vision par ordinateur. Ces modèles peuvent être rangés dans la catégorie des *modèles discriminatifs* alors que les champs de Markov triplets appartiennent à la famille des *modèles génératifs*. En effet, les champs conditionnels modélisent directement la loi *a posteriori* des classes sachant les observations comme une distribution markovienne alors que les champs de Markov triplets modélisent la loi jointe de telle sorte que la loi *a posteriori* soit la marginale d'une distribution markovienne. Un argument en faveur des champs conditionnels est que lorsqu'on veut faire de la classification, en général seule la loi *a posteriori* est vraiment utile. Plus précisément, il arrive que les lois $P(\mathbf{x}|\mathbf{z})$ décrivant le bruit contiennent certes beaucoup de structure mais avec peu d'effet sur la loi *a posteriori* (voir [14] p.144). Inversement, un avantage des modèles génératifs est lié à la probabilité $P(\mathbf{x})$ (la vraisemblance des observations). Celle-ci peut être utile pour détecter des individus qui ont une probabilité faible sous le modèle et pour lesquels le décision de classification risque d'être peu fiable. L'approche générative présente ainsi un avantage certain pour ce qui est de la détection des données abérrantes (*outliers*). De plus, il parait difficile d'accéder à des propriétés théoriques des estimateurs basés par exemple sur le principe du maximum de vraisemblance, donc sur la loi des observations, si celle-ci n'est pas modélisée. Dans cette thèse, nous avons montré que les modèles de Markov triplets pouvaient être utilisés efficacement à l'aide d'algorithmes classiques pour prendre en compte des modèles de bruit complexes, montrant ainsi que l'approche générative peut être efficace.

La troisième source de complexité abordée est celle des observations incomplètes. Nous avons proposé un modèle et une procédure pour traiter de telles données sans faire appel aux méthodes d'imputation qui, bien que largement utilisées, introduisent un biais dans les données et, en remplaçant brutalement les valeurs manquantes, ne permettent pas de prendre en compte l'incertitude sur les valeurs imputées elles-même. Nous avons illustré notre démarche sur des images simulées, ainsi que sur un problème réel de classification de gènes. Ces expériences ont mis en évidence la performance de notre méthode en comparaison aux méthodes d'imputation les plus classiques.

Perspectives

Dans la continuité de ce travail, nous voyons plusieurs prolongements possibles.
Le premier concerne la recherche de techniques permettant d'étendre l'utilisation de notre famille de modèle de Markov triplets au cadre d'une classification non supervisée. En effet, notre modèle souffre d'un possible problème de non identifiabilité en non supervisé, qui n'apparaît pas lorsqu'on se place en supervisé. Il serait intéressant d'étudier plus en détail les techniques existantes et utilisées dans une approche bayésienne pour résoudre le problème de *label-switching* des méthodes MCMC et notamment les algorithmes dits de "relabelisation" (*relabelling algorithm*, voir notamment [74]). Rappelons que des modèles de Markov triplets ont été utilisés dans des applications pratiques [6, 7] mais sous une hypothèse restrictive supplémentaire sur le modèle de bruit par rapport à notre modèle. Adapter des techniques bayésiennes pour résoudre ce problème d'identifiabilité permettrait d'utiliser, en non supervisé, un modèle plus général que celui de [6, 7] et donc probablement d'obtenir des modèles plus performants.

Un second prolongement serait d'étudier théoriquement la transition de phase du champ de Markov (Y, Z) du chapitre VI, section 1, que nous avons mis en évidence par des simulations. Dans les cas extrêmes (par exemple b et c suffisamment grands), la réalisation générée par simulation d'un tel modèle sera unicolore. Or les distributions de Gibbs étant incalculables, de nombreuses méthodes d'approximation markovienne font appel à des techniques de simulation. Dès lors, l'étude de la transition de phase pourrait permettre de définir un espace recommandé pour les paramètres b et c pour les simulations, espace à l'intérieur duquel on serait assuré que les réalisations générées ne soient pas dégénérées (unicolores).
Une troisième perspective à ce travail serait d'introduire une notion d'incertitude dans les graphes sous-jacents aux modèles markoviens en général. En effet, en génomique par exemple, les réseaux d'association disponibles (entre protéines, entre enzymes catalysant des réactions successives...) ne concernent qu'une poignée de gènes sur l'ensemble des gènes existants. A titre d'exemple, les réseaux d'association protéines-protéines actuels concernent environ 5000 gènes alors que le génome humain contient entre 60 000 et 150 000 gènes. D'un côté, se limiter à l'étude de ces 5000 gènes est extrêment restrictif. De l'autre, prendre en compte les gènes dont les dépendances possibles sont inconnues, implique à l'heure actuelle de les supposer indépendants des autres gènes (car hors du graphe d'interaction), même sous modélisation markovienne. Il pourrait être beaucoup plus intéressant d'introduire de l'incertitude sur le graphe plutôt que de le supposer donné comme une vérité. Une approche possible pourrait alors être d'introduire une variable aléatoire partiellement observée représentant la présence, ou non, d'une arête entre 2 individus donnés. Cette variable serait observée pour les arêtes connues (celles associées aux 5000 gènes du réseau d'association de protéines par exemple) et non observée, donc manquante pour les autres. On retrouve alors un problème similaire à celui abordé en Partie C avec la classification d'observations incomplètes.
Dans la même idée, une quatrième extension serait d'introduire la possiblité de sélectionner de manière adaptative le graphe au cours des itérations de l'algorithme de classification. On pourrait par exemple penser à définir les voisins d'un site en fonction de la classe à laquelle il appartient. Dans l'application sur les images texturées du chapitre VI, il serait fondé de ne pas définir le voisinage de toutes les textures de la même manière (par le graphe de Delaunay), mais de permettre à l'algorithme de définir différemment le graphe selon la texture. Une telle problématique a déjà était abordée dans [83]. L'approche utilisée est d'activer, ou non, les voisins d'un site en fonction de la classe à laquelle il appartient. Cette activation est modélisée par l'utilisation d'un champ auxiliaire. Une approche similaire pourrait être adoptée en utilisant un modèle de Markov Triplet comme décrit en Partie B.
Enfin, il serait intéressant de comparer les champs de Markov couples avec les *champs conditionnels* (*Conditional Random Field*, voir notamment [79]) très utilisés en vision par ordinateur. Ce type de modèle appartient à la famille des modèles dits *discriminatifs* par opposition aux modèles *génératifs* des champs de Markov couples et plus généralement de tous les modèles présentés jusqu'à présent. L'objectif de base des champs conditionnels et des champs de Markov couples est la même : faire en sorte que la loi *a posteriori* soit markovienne. Les champs conditionnels travaillent alors directement sur une distribution markovienne *a posteriori* sans même définir de loi jointe, alors que les champs de Markov couples partent d'une loi jointe markovienne, impliquant de ce fait la markovianité de la loi *a posteriori*. Les champs conditionnels semblent très performants en pratique et leur

utilisation est en réelle expansion dans le domaine de la vision par ordinateur. Néanmoins, les procédures d'estimation des paramètres semblent très coûteuses. De plus, on peut se demander si le fait de ne pas définir entièrement le modèle (il n'y a pas de loi jointe, ni de loi sur les observations), n'entraine pas des problèmes théoriques, même si en pratique les résultats semblent très satisfaisants.

Annexes

1 Identifiabilité des mélanges

L'estimation d'un jeu de paramètres $\boldsymbol{\psi}$ à partir d'observations $x_1, \cdots x_n$ n'a de sens que si $\boldsymbol{\psi}$ est identifiable. En général, une famille paramétrique de distributions $f(x|\boldsymbol{\psi})$ est identifiable si des valeurs distinctes de $\boldsymbol{\psi}$ correspondent à des membres distincts de la famille de distributions.

Pour les mélanges de distributions, il est nécessaire de définir l'identifiabilité de ces mélanges un peu différemment. En effet, supposons que

$$f(x|\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k)$$

est un mélange fini à K composants. Si tous les composants du mélange appartiennent à la même famille paramétrique de distributions alors $f(x|\boldsymbol{\psi})$ est invariant par chacune des $K!$ permutations des indices des composantes de $\boldsymbol{\psi}$. Notons alors

$$f'(x|\boldsymbol{\psi}') = \sum_{k=1}^{K'} \pi'_k f_k(x|\theta'_k)$$

un deuxième mélange de la même famille paramétrique de mélanges et S_K l'ensemble des permutations de $[\![1, K]\!]$. Nous dirons que cette classe de mélanges finis est identifiable si :

$$\begin{aligned} & f(x|\boldsymbol{\psi}) = f'(x|\boldsymbol{\psi}') \quad \text{presque sûrement} \\ \Leftrightarrow \quad & \begin{cases} K = K' \\ \exists\, \sigma \in S_K \text{ t.q. } \pi_k = \pi'_{\sigma(k)} \text{ et } \theta_k = \theta'_{\sigma(k)} \end{cases} \end{aligned}$$

La plupart des mélanges finis de distributions continues sont identifiables. Une exception classique est les mélanges de lois uniformes.

Dans ce qui suit nous nous intéressons à une famille paramétrique particulière que nous qualifierons de *mélanges de mélanges*. Il s'agit du cas où les distrubutions des composantes $f_k(x|\theta_k)$ sont elles-mêmes des mélanges finis de distributions d'une même famille paramétrique. Nous montrons ci-dessous que même si on considère une classe de mélange identifiable, les mélanges de ces derniers ne le sont pas.

Notons donc

$$g(x|\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k)$$

avec

$$\sum_{k=1}^{K} \pi_k = 1$$

$$\text{et} \qquad f_k(x|\theta_k) = \sum_{l=1}^{L_k} q(k,l) f_{(k,l)}(x|\theta_{(k,l)})$$

où $L_k > 1$ pour au moins un k et

$$\sum_{l=1}^{L_k} q(k,l) = 1$$

Un tel mélange peut être vu comme un mélange à $G = \sum_{k=1}^{K} L_k$ composantes,

$$g(x|\boldsymbol{\psi}) = \sum_{(k,l)} \pi_k q(k,l) f_{(k,l)}(x|\theta_{(k,l)}) \quad \text{(VIII.1)}$$

Noter que l'on a bien $\sum_{(k,l)} \pi_k q(k,l) = 1$.

Pour simplifier les écritures on supposera dans la suite que $L_k = L$ pour tout k avec $L > 1$. Le cas général se traite de la même manière. Renumerotons alors les indices (k,l) de 0 à $KL-1$ en posant $j = j(k,l) = (k-1)L+l-1$. Inversement on peut retrouver (k,l) à partir de j par $k = j[L]+1$ et $l = j-j[L]\times L+1$ où [.] désigne l opérateur modulo. Posons ensuite $\alpha(k,l) = \pi_k\, q(k,l)$ soit en renumérotant $\alpha_j = \pi_{j[L]+1}\, q(j[L]+1, j-j[L]\times L+1)$. Le mélange (VIII.1) s'écrit alors

$$g(x|\boldsymbol{\psi}) = \sum_{j=0}^{KL-1} \alpha_j f_j(x|\theta_j).$$

Considérons alors un second mélange de mélanges.

$$g'(x|\boldsymbol{\psi}') = \sum_{j=0}^{K'L'-1} \alpha'_j f_j(x|\theta'_j).$$

à $G' = K'L'$ composantes. Pour simplifier, on supposera néanmoins que $L = L'$. Ce cas particulier suffit à montrer la non-identifiabilité. Ainsi, si la classe de mélanges de composantes les $f_j(x|\theta_j)$ est identifiable,

$$g(x|\boldsymbol{\psi}) = g'(x|\boldsymbol{\psi}') \quad \text{presque surement}$$

est équivalent à dire que $K = K'$ et qu'il existe $\sigma \in S_{KL}$ telle que

$$\begin{aligned} \alpha_j &= \alpha'_{\sigma(j)} \\ \theta_j &= \theta'_{\sigma(j)} \end{aligned} \quad \text{(VIII.2)}$$

Cependant, pour montrer que nos mélanges de mélanges sont identifiables il nous faut entre autre, montrer qu'il existe $\sigma \in S_K$ telle que pour tout $k \in [\![1, K]\!]$, $\pi_k = \pi_{\sigma(k)}$.

Or

$$\begin{aligned} \sum_{j=(k-1)L}^{(k-1)L+L-1} \alpha_j &= \sum_{j=(k-1)L}^{(k-1)L+L-1} \pi_{j[L]+1}\, q(j[L]+1, j-j[L]\times L+1) \\ &= \sum_{j=(k-1)L}^{(k-1)L+L-1} \pi_k\, q(k, j-j[L]\times L+1) \end{aligned}$$

car pour tout $j = (k-1)L$ à $(k-1)L+L-1$, on a $j[L]+1 = k$. D'où

$$\sum_{j=(k-1)L}^{(k-1)L+L-1} \alpha_j = \pi_k$$

car pour tout $k \in [\![1, K]\!]$

$$\sum_{j=(k-1)L}^{(k-1)L+L-1} q(k, j - j[L] \times L + 1) = \sum_{l=1}^{L} q(k, l) = 1$$

Ainsi par (VIII.2) il vient,

$$\begin{aligned} \pi_k &= \sum_{j=(k-1)L}^{(k-1)L+L-1} \alpha'_{\sigma(j)} \\ &= \sum_{j=(k-1)L}^{(k-1)L+L-1} \pi'_{\sigma(j)[L]+1}\, q'(\sigma(j)[L] + 1, \sigma(j) - \sigma(j)[L] \times L + 1) \end{aligned}$$

mais pour $j = (k-1)L$ à $(k-1)L + L - 1$ et $\sigma \in S_{KL}$, $\sigma(j)[L] + 1$ n'est en général pas constant d'où l'impossibilité d'avoir $\pi_k = \pi_{\sigma(k)}$ en général.

2 Approche variationnelle : Approximation en champ moyen

L'objectif de cette annexe est de donner un moyen d'approcher une distribution de Gibbs par une distribution plus simple. Nous présentons l'approche variationnelle conduisant à l'approximation en champ moyen et permettant d'approcher une distribution markovienne par une distribution factorisée optimale au sens de la divergence de Kullback-Leibler. On pourra également se référer aux ouvrages [72] et [133].

2.1 Divergence de Kullback-Leibler et énergie libre

Définition 16 (Divergence de Kullback-Leibler)**.** *La divergence de Kullback-Leibler entre deux distributions P et Q est définie par :*

$$KL(Q||P) = \sum_{\mathbf{z}} Q(\mathbf{z}) \log \frac{Q(\mathbf{z})}{P(\mathbf{z})} = \left\langle \log \frac{Q(\mathbf{Z})}{P(\mathbf{Z})} \right\rangle_Q$$

où on a noté $\langle . \rangle_Q$ l'espérance par rapport à la loi Q.

Cette divergence n'est pas une distance mathématique : elle est non symétrique et elle ne satisfait pas l'inégalité triangulaire. Elle présente néanmoins la caractéristique intéressante d'être toujours non négative et nulle si et seulement si les deux distributions P et Q sont égales, Q presque partout.

Cas d'une distribution de Gibbs Soit P_G une distribution de Gibbs d'énergie H :

$$P_G(\mathbf{z}) = \frac{1}{W} \exp(-H(\mathbf{z}))$$

La divergence de Kullback-Leibler entre une distribution Q quelconque et P_G est :

$$KL(Q||P_G) = \langle H(\mathbf{Z}) \rangle_Q + \langle \log Q(\mathbf{Z}) \rangle_Q + \log W$$

où $\langle . \rangle_Q$ désigne l'espérance sous la loi Q. La divergence de Kullback-Leibler sera nulle, et donc la distribution Q sera égale à la distribution de Gibbs P_G, lorsque :

$$KL(Q||P_G) = 0 \quad \Leftrightarrow \quad F(Q||P_G) = -\log W$$

où $F(Q||P_G) = \langle H(\mathbf{Z}) \rangle_Q + \langle \log Q(\mathbf{Z}) \rangle_Q$ est généralement appelée l'*énergie libre de Gibbs* et $-\log W$ est l'*énergie libre de Helmotz* (elle ne dépend que de P_G). Étant donnée une distribution P_G, trouver la distribution Q minimisant la divergence de Kullback-Leibler revient à trouver Q minimisant l'énergie libre $F(.||P_G)$.

2.2 Approximation d'une distribution de Gibbs par un produit

2.2.1 Cas général

Soit P une distribution quelconque. Cherchons, non pas le vrai minimum de l'énergie libre (ou de manière équivalente le minimum de la divergence de Kullback-Leibler) par

rapport à P, mais le minimum parmi l'ensemble des distributions de probabilités Q qui ont la propriété de factorisation :

$$Q(\mathbf{z}) = \prod_i Q_i(z_i) \tag{VIII.3}$$

Minimiser $KL(Q||P)$ par rapport à Q sous la contrainte $\forall i \in \mathcal{I},\ \sum_{z_i} Q_i(z_i) = 1$ revient, en notant λ_i le multiplicateur de Lagrange, à résoudre :

$$\frac{\partial}{\partial Q_i(z_i)} \left[KL(Q||P) - \lambda_i \left(\sum_{z_i} Q_i(z_i) - 1 \right) \right] = 0$$

Notons $\mathbf{Z}_{\neq i} = \{Z_j, j \neq i\}$ et $Q_{\neq i} = \prod_{j \neq i} Q_j$. En utilisant les relations $P(\mathbf{Z}) = P(\mathbf{Z}_{\neq i})P(Z_i|\mathbf{Z}_{\neq i})$ et $Q(\mathbf{Z}) = Q_{\neq i}(\mathbf{Z}_{\neq i})Q_i(Z_i)$, il vient :

$$KL(Q||P) = \left\langle \log \frac{Q_{\neq i}(\mathbf{Z}_{\neq i})}{P(\mathbf{Z}_{\neq i})} \right\rangle_{Q_{\neq i}} + \langle \log Q_i(Z_i) \rangle_{Q_i} - \langle \log P(Z_i|\mathbf{Z}_{\neq i}) \rangle_Q$$

Seuls les deux derniers termes de l'expression dépendent de Q_i. Or,

$$\frac{\partial \langle \log Q_i(Z_i) \rangle_{Q_i}}{\partial Q_i(z_i)} = \log Q_i(z_i) + 1$$

$$\frac{\partial \langle \log P(Z_i|\mathbf{Z}_{\neq i}) \rangle_Q}{\partial Q_i(z_i)} = \sum_{\mathbf{z}_{\neq i}} \left(\prod_{j \neq i} Q_j(z_j) \right) \log P(z_i|\mathbf{z}_{\neq i}) = \langle \log P(z_i|\mathbf{Z}_{\neq i}) \rangle_{Q_{\neq i}}$$

et donc :

$$\begin{aligned} & \frac{\partial}{\partial Q_i(z_i)} \left[KL(Q||P) - \lambda \left(\sum_{z_i} Q_i(z_i) - 1 \right) \right] = 0 \\ \Leftrightarrow \quad & \log Q_i(z_i) + 1 - \langle \log P(z_i|\mathbf{Z}_{\neq i}) \rangle_{Q_{\neq i}} - \lambda = 0 \\ \Leftrightarrow \quad & Q_i(z_i) = \frac{\exp \langle \log P(z_i|\mathbf{Z}_{\neq i}) \rangle_{Q_{\neq i}}}{\exp(1-\lambda)} \end{aligned} \tag{VIII.4}$$

où la constante de normalisation peut être calculée facilement par :

$$\exp(1-\lambda) = \sum_{z_i} \exp \langle \log P(z_i|\mathbf{Z}_{\neq i}) \rangle_{Q_{\neq i}}$$

2.2.2 Cas d'une distribution de Gibbs

Soit P_G une distribution de Gibbs de fonctions potentiels V_c sur les cliques $c \in \mathcal{C}$:

$$P_G(\mathbf{z}) = \frac{1}{W} \exp(-\sum_{c \in C} V_c(z_c)) \tag{VIII.5}$$

On a donc :

$$P_G(z_i|\mathbf{z}_{\neq i}) = P_G(z_i|\mathbf{z}_{N_i}) \propto \exp(-\sum_{c\ni i} V_c(z_i, \mathbf{z}_{c\setminus i}))$$

D'après (VIII.4), la distribution Q vérifie donc :

$$Q_i(z_i) \propto \exp\left\langle -\sum_{c\ni i} V_c(z_i, \mathbf{Z}_{c\setminus i})\right\rangle_{Q_{\neq i}} = \prod_{c\ni i} \exp\left\langle -V_c(z_i, \mathbf{Z}_{c\setminus i})\right\rangle_{Q_{c\setminus i}} \tag{VIII.6}$$

On en déduit la proposition :

Proposition 17. *Soit P_G une distribution de Gibbs définie par (VIII.5). Parmi l'ensemble des distributions Q ayant la propriété de factorisation (VIII.3), celle minimisant la divergence de Kullback-Leibler avec P_G est :*

$$Q(\mathbf{z}) = \prod_i \frac{\exp\left(-\sum_{c\ni i}\left\langle V_c(z_i, \mathbf{Z}_{c\setminus i})\right\rangle_{Q_{c\setminus i}}\right)}{\sum_{z_i'}\exp\left(-\sum_{c\ni i}\left\langle V_c(z_i', \mathbf{Z}_{c\setminus i})\right\rangle_{Q_{c\setminus i}}\right)} \tag{VIII.7}$$

Supposons que les cliques soient d'ordre 1 et 2 uniquement. Notons E l'espace d'état des Z_i et $\mathcal{E}$ l'espace d'état des moyennes $\langle Z_i\rangle$. Soit c la clique sur la paire (i,j) de sites voisins. Supposons que la fonction potentiel V_c (définie sur $E\times E$) soit bilinéaire et puisse être étendue à $\mathcal{E}\times\mathcal{E}$. Alors $\langle V_c(z_i, Z_j)\rangle = V_c(z_i, \langle Z_j\rangle)$ et l'expression (VIII.7) revient à fixer les voisins j du site i à leur moyenne $\mu_j = \langle \mathbf{Z}_j\rangle_{Q_j}$. On retrouve alors l'approximation en champ moyen [31] de la physique statistique et l'expression (VIII.7) s'écrit :

$$Q(\mathbf{z}) = \prod_i \frac{\exp\left(-\sum_{c\ni i} V_c(z_i, \boldsymbol{\mu}_{c\setminus i})\right)}{\sum_{z_i'}\exp\left(-\sum_{c\ni i} V_c(z_i, \boldsymbol{\mu}_{c\setminus i})\right)} = \prod_i P_G(z_i|\boldsymbol{\mu}_{N_i}) \tag{VIII.8}$$

Notons que dans le cas où les Z_i sont discrets, $Z_i \in [\![1, K]\!]$ par exemple, on peut toujours se ramener à un tel cas en définissant, sur les cliques c d'ordre 2, les matrices $\mathbb{V}_c$ de dimension $K\times K$ telles que $V_c(z_i, z_j) = \mathbf{e}'_{z_i}\mathbb{V}_c\ \mathbf{e}_{z_j}$ où $\mathbf{e}_k$ désigne le k-ième vecteur de la base canonique en dimension K. L'espace d'état des $\mathbf{e}_{Z_i}$ est alors $E = \{\mathbf{e}_1, \cdots, \mathbf{e}_K\}$ et celui des $\langle \mathbf{e}_{Z_i}\rangle$ est $\mathcal{E} = \{\mathbf{u}\in[0,1]^K, \sum_{k=1}^K u_k = 1\}$. La fonction de $E\times E \to \mathbb{R}$ qui, à un couple $(\mathbf{e}_k, \mathbf{e}_l)$ associe $\mathbf{e}'_k\mathbb{V}_c\ \mathbf{e}_l$ est bilinéaire et peut être étendue à $\mathcal{E}\times\mathcal{E}$. En notant μ_j le vecteur des K probabilités $P_G(Z_j = k)$ et $\mathbf{z}_i$ sous forme vectorielle ($z_i = k \Leftrightarrow \mathbf{z}_i = \mathbf{e}_k$), l'équation (VIII.8) est valable.

Remarque. *Dans le cas général, minimiser la divergence de Kullback-Leibler n'est pas toujours équivalent à remplacer le champ des voisins par leur moyenne.*

C'est toutefois le cas dans beaucoup de cas courant, notamment lorsque les Z_i sont discrets et les cliques d'ordre ≥ 3 nulles. En particulier, les modèles classiques de la physique statistique (modèles d'Ising, de Potts) sont discrets et définis sur les singletons et paires uniquement. Le terme "champ moyen" [31] fait référence au remplacement des voisins du site i considéré par leurs moyennes.

2.2.3 Condition de cohérence

Remarquons que les expressions (VIII.7) et (VIII.8) ne fournissent pas de solution explicite puisque les expressions de droite dépendent d'espérances calculées sous la loi Q. Une approche possible est alors de résoudre le problème itérativement. Dans le cas particulier où l'équation (VIII.8) est valable (le champ moyen "usuel"), cette approche peut être remplacée par une équation de cohérence sur les moyennes. En effet, pour que l'approximation de $P_G(z_i)$ au site i par $P_G(z_i|\boldsymbol{\mu}_{N_i})$ soit cohérente, il est nécessaire que l'espérance de Z_i sous la loi $P(z_i|\boldsymbol{\mu}_{N_i})$ soit égale à la valeur μ_i utilisée pour l'approximation, c'est-à-dire que :

$$\mu_i = \langle Z_i \rangle_{P(.|\boldsymbol{\mu}_{N_i})} \tag{VIII.9}$$

Le terme de droite est une fonction Λ_i de $\boldsymbol{\mu}_{N_i} = \{\mu_j, j \in N_i\}$. Cette condition de cohérence sur l'ensemble des sites $\mathcal{I} = \{1, \cdots, n\}$ s'écrit alors :

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \Lambda_1(\boldsymbol{\mu}_{N_1}) \\ \vdots \\ \Lambda_n(\boldsymbol{\mu}_{N_n}) \end{pmatrix}$$

soit, en notation vectorielle,

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}(\boldsymbol{\mu}) \tag{VIII.10}$$

L'approximation revient donc à résoudre un problème de point fixe.

2.2.4 Exemple : cas du modèle de Potts étendu

Soit $\mathbf{Z}$ un modèle de Potts étendu (voir chapitre I, section 1.3.2) à K couleurs (les Z_i sont à valeurs dans $[\![1, K]\!]$), d'énergie H paramétrée par une matrice symétrique $\boldsymbol{\beta}$ de dimension $K \times K$:

$$H(\mathbf{z}) = \sum_{c \in C} V_c(\mathbf{z}_c) = -\sum_{i \sim j} \beta_{z_i z_j} \tag{VIII.11}$$

Avec la notation vectorielle $\mathbf{z}_i = \mathbf{e}_{z_i}$ où $(\mathbf{e}_1, \cdots, \mathbf{e}_K)$ désigne la base canonique en dimension K, l'énergie (VIII.11) s'écrit encore :

$$H(\mathbf{z}) = -\beta \sum_{i \sim j} \mathbf{z}_i' \boldsymbol{\beta} \mathbf{z}_j = -\frac{1}{2} \sum_i \sum_{j \in N_i} \mathbf{z}_i' \boldsymbol{\beta} \mathbf{z}_j \tag{VIII.12}$$

Sous cette forme, l'énergie est bilinéaire et chercher une distribution factorisée minimisant la divergence de Kullback-Leibler revient à appliquer le principe du champ moyen [31]. Au site i, la distribution marginale $P_G(\mathbf{z}_i)$ est remplacée par $P_G(\mathbf{z}_i|\boldsymbol{\mu}_{N_i})$ d'énergie :

$$H(\mathbf{z}_i|\boldsymbol{\mu}_{N_i}) = -\sum_{j \in N_i} \mathbf{z}_i' \boldsymbol{\beta} \boldsymbol{\mu}_j \tag{VIII.13}$$

et de fonction de partition :

$$W_i(\boldsymbol{\mu}) = \sum_{\mathbf{z}_i} \exp\Big(\sum_{j \in N_i} \mathbf{z}_i' \boldsymbol{\beta} \boldsymbol{\mu}_j\Big) = \sum_{k=1}^{K} \exp\Big(\sum_{j \in N_i} \sum_{l=1}^{K} \beta_{kl} \mu_{jl}\Big) \tag{VIII.14}$$

L'équation de cohérence (VIII.10) au site i devient :

$$\boldsymbol{\mu}_i = \begin{pmatrix} W_i(\boldsymbol{\mu})^{-1} \exp(\sum_{j \in N_i} \sum_{l=1}^{K} \beta_{1l} \mu_{jl}) \\ \vdots \\ W_i(\boldsymbol{\mu})^{-1} \exp(\sum_{j \in N_i} \sum_{l=1}^{K} \beta_{Kl} \mu_{jl}) \end{pmatrix} = W_i(\boldsymbol{\mu})^{-1} \sum_{\mathbf{z}_i} \mathbf{z}_i \exp(\sum_{j \in N_i} \mathbf{z}_i' \boldsymbol{\beta} \boldsymbol{\mu}_j) \quad \text{(VIII.15)}$$

Remarque. *Soit un modèle de Potts classique d'énergie paramétrée par le scalaire β (c'est-à-dire tel que $\boldsymbol{\beta} = \beta \mathbb{I}_K$ avec $\mathbb{I}_K$ la matrice identité de dimension K). Alors l'approximation en champ moyen est peu intéressante puisqu'il s'agit de la configuration uniforme $\mathbf{u}^K$ en chaque site, indépendamment du paramètre spatial β : $\forall i \in \mathcal{I}$, $\mathbf{u}_i^K = \frac{1}{K}(1, \cdots, 1)$. Une telle approximation est en pratique beaucoup plus intéressante pour approximer une distribution markovienne a posteriori $P(\mathbf{z}|\mathbf{x})$ (voir chapitre II, section 4.2).*

3 Résultats généraux sur les champs de Markov

On se place dans le cadre d'une modélisation par champ de Markov $\mathbf{Z}$ de distribution de Gibbs paramétrée par un ensemble de paramètres $\boldsymbol{\phi}$:

$$\begin{aligned} P_G(\mathbf{z}) &= W(\boldsymbol{\phi})^{-1} \exp\left(-H(\mathbf{z};\boldsymbol{\phi})\right) \\ \text{où} \quad W(\boldsymbol{\phi}) &= \sum_{\mathbf{z}} \exp(-H(\mathbf{z};\phi)) \end{aligned}$$

On donne dans cette annexe différents résultats sur les dérivées de $\log W$ et $\log P_G$ utilisés à différents endroits de la thèse.

Proposition 18. *En notant ∇_ϕ le gradient par rapport à $\boldsymbol{\phi}$, on a :*

$$\begin{aligned} \nabla_\phi \log W(\boldsymbol{\phi}) &= -\langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle \\ \nabla_\phi \log P_G(\mathbf{z};\boldsymbol{\phi}) &= -\nabla_\phi H(\mathbf{z};\boldsymbol{\phi}) + \langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle \\ \nabla_\phi P_G(\mathbf{z};\phi) &= P_G(\mathbf{z};\boldsymbol{\phi})(-\nabla_\phi H(\mathbf{z};\boldsymbol{\phi}) + \langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle) \end{aligned}$$

Preuve :

$$\begin{aligned} \nabla_\phi \log W(\boldsymbol{\phi}) &= -\nabla_\phi \log(\sum_{\mathbf{z}} \exp(-H(\mathbf{z};\phi))) \\ &= -\sum_{\mathbf{z}} \nabla_\phi H(\mathbf{z};\phi) P_G(\mathbf{z};\phi) \\ &= -\langle \nabla_\phi H(\mathbf{Z};\phi)\rangle \end{aligned}$$

$$\begin{aligned} \nabla_\phi \log P_G(\mathbf{z};\phi) &= -\nabla_\phi H(\mathbf{z};\phi) - \nabla_\phi \log W(\boldsymbol{\phi}) \\ &= -\nabla_\phi H(\mathbf{z};\phi) + \langle \nabla_\phi H(\mathbf{Z};\phi)\rangle \end{aligned}$$

$$\begin{aligned} \nabla_\phi P_G(\mathbf{z};\phi) &= \nabla_\phi \log P_G(\mathbf{z};\phi)\ P_G(\mathbf{z};\phi) \\ &= P_G(\mathbf{z};\phi)(-\nabla_\phi H(\mathbf{z};\phi) + \langle \nabla_\phi H(\mathbf{Z};\phi)\rangle) \end{aligned}$$

Corollaire 19. *En notant ∇^2_ϕ la hessienne par rapport à $\boldsymbol{\phi}$, on a :*

$$\begin{aligned} \nabla^2_\phi \log W(\boldsymbol{\phi}) &= -\langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle + Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})) \\ \nabla^2_\phi \log P_G(\mathbf{z};\boldsymbol{\phi}) &= -\nabla^2_\phi H(\mathbf{z};\boldsymbol{\phi}) + \langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle - Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})) \end{aligned}$$

Preuve : En dérivant $\nabla_\phi \log W(\boldsymbol{\phi})$ par rapport à $\boldsymbol{\phi}$, il vient :

$$\nabla^2_\phi \log W(\boldsymbol{\phi}) = -\langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle - \sum_{\mathbf{z}} \nabla_\phi P_G(\mathbf{z}|\boldsymbol{\phi}) \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})$$

En utilisant la formule l'expression de $\nabla_\phi P_G(\mathbf{z}|\boldsymbol{\phi})$, on a :

$$\begin{aligned} \sum_{\mathbf{z}} \nabla_\phi P_G(\mathbf{z}|\boldsymbol{\phi}) \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}) &= \sum_{\mathbf{z}} P_G(\mathbf{z};\boldsymbol{\phi})(-\nabla_\phi H(\mathbf{z};\boldsymbol{\phi}) + \langle \nabla_\phi H(\mathbf{Z};\phi)\rangle) \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}) \\ &= -\langle (\nabla_\phi H(\mathbf{Z};\phi))^2\rangle + \langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle^2 \\ &= -Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})) \end{aligned}$$

et donc

$$\begin{aligned}\nabla_\phi^2 \log W(\boldsymbol{\phi}) &= -\langle \nabla_\phi^2 H(\mathbf{Z};\phi)\rangle + Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}))\\ \nabla_\phi^2 \log P_G(\mathbf{z};\boldsymbol{\phi}) &= -\nabla_\phi^2 H(\mathbf{z};\boldsymbol{\phi}) + \langle \nabla_\phi^2 H(\mathbf{Z};\boldsymbol{\phi})\rangle - Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}))\end{aligned}$$

Corollaire 20. *Si l'énergie $H(\mathbf{z};\boldsymbol{\phi})$ s'écrit sous la forme $H(\mathbf{z};\boldsymbol{\phi}) = \boldsymbol{\phi}.G(\mathbf{z})$ où . désigne le produit scalaire et $G(\mathbf{z})$ est une fonction, alors*

$$\begin{aligned}\nabla_\phi^2 \log W(\boldsymbol{\phi}) &= Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}))\\ \nabla_\phi^2 \log P_G(\mathbf{Z};\boldsymbol{\phi}) &= -Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}))\end{aligned}$$

et l'estimateur de maximum de vraisemblance de P_G est unique.

Preuve : Sous cette forme, $\nabla_\phi^2 H(\mathbf{z};\boldsymbol{\phi}) = 0$ donc $\nabla_\phi^2 \log P_G(\mathbf{z};\boldsymbol{\phi}) = -Var(\nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}))$ et la hessienne de $\log P_G(\mathbf{Z};\boldsymbol{\phi})$ est une matrice symétrique négative. La vraisemblance est alors concave.

Remarque. *C'est en particulier le cas du modèle de Potts étendu (voir chapitre I, section 1.3.2), d'énergie $H(\mathbf{z}) = -\sum_i \alpha_{z_i} - \sum_{i\sim j} \beta_{z_i z_j}$, en notant :*

$$\begin{array}{lcccccccc}\boldsymbol{\phi} & = & (\alpha_1, & \cdots & \alpha_K, & \beta_{11}, & \cdots & \beta_{kh}, & \cdots & \beta_{KK})\\ G(\mathbf{z}) & = & (\sum\limits_i 1_{z_i=1}, & \cdots & \sum\limits_i \mathbb{1}_{z_i=K}, & \sum\limits_{i\sim j} \mathbb{1}_{z_i=1}\mathbb{1}_{z_j=1}, & \cdots & \sum\limits_{i\sim j} \mathbb{1}_{z_i=k}\mathbb{1}_{z_j=h}, & \cdots & \sum\limits_{i\sim j} \mathbb{1}_{z_i=K}\mathbb{1}_{z_j=K})\end{array}$$

4 Estimation des paramètres du champ de Markov

On se place dans le cadre d'une modélisation par champ de Markov caché avec hypothèse de bruit indépendant (voir chapitre II, section 4.1). On désire estimer les paramètres $\boldsymbol{\psi} = (\boldsymbol{\phi}, \boldsymbol{\theta})$ du champ de Markov caché par application de l'algorithme NREM (chapitre II, section 4.4.2). A l'étape (E), le calcul des probabilités *a posteriori* ne pose pas de problème (équation II.32). A l'étape (M), la mise à jour des paramètres $\boldsymbol{\theta}$ nécessite les même calculs que dans le cas de l'algorithme EM pour modèle de mélange indépendant. En particulier dans le cas gaussien, les formules de mise à jour sont explicites (voir chapitre II, section 3.4.4). Cette annexe a pour objet l'estimation des paramètres $\boldsymbol{\phi}$ du champ de Markov $\mathbf{Z}$ lors de l'étape (M). Nous donnons dans un premier temps l'expression générale du gradient à calculer avec l'algorithme EM, avant d'en déduire celle pour l'algorithme NREM, c'est-à-dire sous approximation de type champ moyen de la distribution de Gibbs.

Proposition 21. *Soit* $Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\boldsymbol{\psi}^{(q)}) = \langle \log P_G(\mathbf{Z}|\boldsymbol{\phi})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle$ *la fonction de* $\boldsymbol{\phi}$ *à maximiser à l'itération* (q) *de EM. Son gradient est donné par :*

$$\nabla_\phi Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)}) = -\langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})|\mathbf{x}, \psi^{(q)}\rangle + \langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle$$

Preuve :

$$\begin{aligned}
\nabla_\phi Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\boldsymbol{\psi}^{(q)}) &= \langle \nabla_\phi \log P_G(\mathbf{Z}|\boldsymbol{\phi})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle \\
&= \langle -\nabla_\phi H(\mathbf{Z};\phi) + \langle \nabla_\phi H(\mathbf{Z};\phi)\rangle \,|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle \\
&= -\langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})|\mathbf{x}, \psi^{(q)}\rangle + \langle \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle
\end{aligned}$$

Proposition 22. *Soit* $Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\boldsymbol{\psi}^{(q)}) = \langle \log P_G(\mathbf{Z}|\boldsymbol{\phi})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle$ *la fonction de* $\boldsymbol{\phi}$ *à maximiser à l'itération* (q) *de EM. Sa matrice hessienne est donnée par :*

$$\nabla^2_\phi Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)}) = -\langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})|\mathbf{x}, \psi^{(q)}\rangle + \langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle - Var(\nabla_\phi H(\mathbf{Z};\phi))$$

Preuve : Dérivons par rapport à $\boldsymbol{\phi}$ le gradient $\nabla_\phi Q_{\boldsymbol{\phi}}$ de la Proposition 21.

$$\begin{aligned}
\nabla^2_\phi Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)}) &= -\langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})|\mathbf{x}, \psi^{(q)}\rangle + \langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle + \sum_{\mathbf{z}} \nabla_\phi P_G(\mathbf{z}|\boldsymbol{\phi}) \nabla_\phi H(\mathbf{Z};\boldsymbol{\phi}) \\
&= -\langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})|\mathbf{x}, \psi^{(q)}\rangle + \langle \nabla^2_\phi H(\mathbf{Z};\boldsymbol{\phi})\rangle - Var(\nabla_\phi H(\mathbf{Z};\phi))
\end{aligned}$$

Corollaire 23. *Si l'énergie* $H(\mathbf{z};\boldsymbol{\phi})$ *s'écrit sous la forme* $H(\mathbf{z};\boldsymbol{\phi}) = \boldsymbol{\phi}.G(\mathbf{z})$ *où . désigne le produit scalaire et* $G(\mathbf{z})$ *est une fonction, alors*

$$\exists!\ \ \hat{\phi} = \arg\max_{\boldsymbol{\phi}} Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)})$$

Preuve : Sous cette forme, $\nabla^2_\phi H(\mathbf{z};\boldsymbol{\phi}) = 0$ et, d'après la proposition 22,

$$\nabla^2_\phi Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)}) = -Var(\nabla_\phi H(\mathbf{Z};\phi)),$$

qui est une matrice définie négative.

Remarque. *C'est en particulier le cas du modèle de Potts étendu (voir chapitre I, section 1.3.2), d'énergie* $H(\mathbf{z}) = -\sum_i \alpha_{z_i} - \sum_{i\sim j} \beta_{z_i z_j}$*, en notant :*

$$\begin{array}{lcccccccc} \boldsymbol{\phi} & = & (\alpha_1, & \cdots & \alpha_K, & \beta_{11}, & \cdots & \beta_{kh}, & \cdots & \beta_{KK}) \\ G(\mathbf{z}) & = & (\sum\limits_i \mathbb{1}_{z_i=1}, & \cdots & \sum\limits_i \mathbb{1}_{z_i=K}, & \sum\limits_{i\sim j} \mathbb{1}_{z_i=1}\mathbb{1}_{z_j=1}, & \cdots & \sum\limits_{i\sim j} \mathbb{1}_{z_i=k}\mathbb{1}_{z_j=h}, & \cdots & \sum\limits_{i\sim j} \mathbb{1}_{z_i=K}\mathbb{1}_{z_j=K}) \end{array}$$

Proposition 24. *Soit* $\tilde{Q}_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\boldsymbol{\psi}^{(q)}) = \langle \log P_{\tilde{\mathbf{z}}^{\mathbf{x}}}(\mathbf{Z}|\boldsymbol{\phi})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle$ *la fonction de* $\boldsymbol{\phi}$ *à maximiser à l'itération* (q) *de NREM. Alors :*

$$\begin{aligned} \nabla_\phi \tilde{Q}_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)}) &= \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{L}} (\tilde{t}_{ik}^{(q)} - \tilde{\pi}_{ik}) \nabla_\phi \tilde{H}_i(k;\boldsymbol{\phi}) \\ \nabla_\phi^2 \tilde{Q}_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)}) &= \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{L}} \left((\tilde{t}_{ik}^{(q)} - \tilde{\pi}_{ik}) \nabla_\phi^2 \tilde{H}_i(k;\boldsymbol{\phi}) + \tilde{\pi}_{ik}(1-\tilde{\pi}_{ik})(\nabla_\phi \tilde{H}_i(k;\boldsymbol{\phi}))^2 \right) \end{aligned}$$

Preuve : Il suffit de remplacer dans la Proposition 21, l'énergie $H(\mathbf{z}, \boldsymbol{\phi})$ par son approximation en champ moyen $\sum_i \tilde{H}_i(z_i;\boldsymbol{\phi})$, puis de dériver, en notant que $\nabla_\phi \tilde{\pi}_{ik} = \tilde{\pi}_{ik}(-\nabla_\phi^2 \tilde{H}_i(k) + \langle \tilde{H}_i(Z_i)\rangle)$

Corollaire 25. *Si l'énergie* $H(\mathbf{z};\boldsymbol{\phi})$ *s'écrit sous la forme* $H(\mathbf{z};\boldsymbol{\phi}) = \boldsymbol{\phi}.G(\mathbf{z})$ *où* $.$ *désigne le produit scalaire et* $G(\mathbf{z})$ *est une fonction, alors*

$$\exists!\ \ \hat{\phi} = \arg\max_{\boldsymbol{\phi}} \tilde{Q}_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)})$$

Preuve : Sous cette forme, $\nabla_\phi^2 \tilde{H}_i(k;\boldsymbol{\phi}) = 0$ et $\nabla_\phi^2 \tilde{Q}_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\psi^{(q)}) = -\sum_i Var(\nabla_\phi \tilde{H}_i(Z_i;\phi))$, qui est une matrice définie négative.

Corollaire 26. *Dans le cas du modèle de Potts étendu, l'estimation des paramètres* $\boldsymbol{\phi}$ *à l'étape (M) de l'algorithme NREM peut se faire par descente de gradient. Les dérivées sont données par :*

$$\begin{aligned} \frac{\partial \tilde{Q}_{\boldsymbol{\phi}}}{\alpha_k} &= \sum_{i\in\mathcal{I}} (\tilde{t}_{ik}^{(q)} - \tilde{\pi}_{ik}) \\ \frac{\partial \tilde{Q}_{\boldsymbol{\phi}}}{\beta_{kk'}} &= \begin{cases} \sum_i (\tilde{t}_{ik}^{(q)} - \tilde{\pi}_{ik})\tilde{S}_{ik} & si\ k = k' \\ \sum_i (\tilde{t}_{ik}^{(q)} - \tilde{\pi}_{ik})\tilde{S}_{ik'} + \sum_i (\tilde{t}_{ik'}^{(q)} - \tilde{\pi}_{ik'})\tilde{S}_{ik} & sinon \end{cases} \end{aligned}$$

où on a noté $\tilde{S}_{ik} = \sum_{j\in N_i} \tilde{z}_{jk}^{\mathbf{x}}$ *et où* $\tilde{z}_{jk}^{\mathbf{x}}$ *désigne la composante* k *du champ* $\tilde{z}_j^{\mathbf{x}}$ *au site* j*.*

Preuve : On a, pour tout $i \in \mathcal{I}$ et $z_i \in \mathcal{K}$, $\tilde{H}_i(z_i;\boldsymbol{\phi}) = \alpha_{z_i} + \sum_{j\in N_i}\sum_h \beta_{z_i h}\tilde{z}_{jh}^{\mathbf{x}}$. Donc,

$$\begin{aligned} \frac{\partial \tilde{H}_i(z_i;\phi)}{\alpha_k} &= \begin{cases} 1 & \text{si } z_i = k \\ 0 & \text{sinon} \end{cases} \\ \frac{\partial \tilde{H}_i(z_i;\phi)}{\beta_{kk'}} &= \begin{cases} \tilde{S}_{ik} & \text{si } z_i = k \\ \tilde{S}_{ik'} & \text{si } z_i = k' \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

et il suffit de remplacer dans le gradient de la Proposition 24.

5 Estimation des paramètres du champ de Markov triplet

On se place dans le cadre d'une modélisation par champ de Markov Triplet $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ comme défini au chapitre V, section 2. Nous détaillons les formules d'estimation des paramètres $\mathbb{B}_{kk'}$ et $\mathbb{C}$, lors de la phase d'apprentissage et de test, par l'algorithme NREM.

5.1 Etape d'apprentissage

Comme décrit au chapitre V, section 3.1, la mise à jour des paramètres $\mathbb{B} = \{\mathbb{B}_{kk'}\}$ est donnée par :

$$\begin{aligned} \mathbb{B}^{(q+1)} &= \arg\max_{\mathbb{B}} \sum_{i\in\mathcal{I}_1}\sum_{l\in\mathcal{L}} \tilde{t}_{il}^{(q)} \log \tilde{\pi}_{il} \\ \text{avec } \tilde{\pi}_{il} &= P(Y_i = l|\tilde{y}_{N_i}, \mathbf{z}, \mathbb{B}) \propto \exp(\sum_{j\in N_i}\sum_{l'\in\mathcal{L}} B_{z_i z_j}^{ll'} \tilde{y}_{jl'}) \end{aligned}$$

où on a noté $B_{kk'}^{ll'}$ la composante (l, l') de la matrice $\mathbb{B}_{kk'}$ et et $\tilde{y}_{jl}$ la composante l du champ $\tilde{y}_j$ au site j. Soit

$$\tilde{H}_i(l; \mathbb{B}) = \sum_{j\in N_i}\sum_{l'\in\mathcal{L}} B_{z_i z_j}^{ll'} \tilde{y}_{jl'}$$

Soit $\tilde{Q}(\mathbb{B}|\boldsymbol{\psi}^{(q)}) = \langle \log P_{\tilde{\mathbf{y}}}(\mathbf{Y}|\mathbf{z}; \mathbb{B})|\mathbf{x}, \boldsymbol{\psi}^{(q)}\rangle$ la fonction de $\mathbb{B}$ à maximiser à l'itération (q) de NREM et $\boldsymbol{\psi}^{(q)}$ la valeur des paramètres $(\mathbb{B}, \boldsymbol{\theta})$ courants. D'après la Proposition 24, le gradient de $\tilde{Q}$ par rapport à $\mathbb{B}$ est donné par :

$$\nabla\tilde{Q}(\mathbb{B}|\boldsymbol{\psi}^{(q)}) = \sum_{i\in\mathcal{I}_1}\sum_{l\in\mathcal{L}} (\tilde{t}_{il}^{(q)} - \tilde{\pi}_{il})\nabla\tilde{H}_i(l; \mathbb{B})$$

Proposition 27. *L'estimation des paramètres $\mathbb{B}$ à l'étape (M) de l'algorithme NREM peut se faire par descente de gradient. Les dérivées sont données par :*

$$\frac{\partial\tilde{Q}}{B_{kk'}^{ll'}} = \begin{cases} \sum_{i/z_i=k} (\tilde{t}_{il}^{(q)} - \tilde{\pi}_{il})\tilde{S}_{ilk} & si\ k = k'\ et\ l = l' \\ \sum_{i/z_i=k} [(\tilde{t}_{il}^{(q)} - \tilde{\pi}_{il})\tilde{S}_{il'k} + (\tilde{t}_{il'}^{(q)} - \tilde{\pi}_{il'})\tilde{S}_{ilk}] & si\ k = k'\ et\ l \neq l' \\ \sum_{i/z_i=k} (\tilde{t}_{il}^{(q)} - \tilde{\pi}_{il})\tilde{S}_{ilk'} + \sum_{i/z_i=k'} (\tilde{t}_{il}^{(q)} - \tilde{\pi}_{il})\tilde{S}_{ilk} & si\ k \neq k'\ et\ l = l' \\ \sum_{i/z_i=k} [(\tilde{t}_{il'}^{(q)} - \tilde{\pi}_{il'})\tilde{S}_{ilk'} + (\tilde{t}_{il}^{(q)} - \tilde{\pi}_{il})\tilde{S}_{il'k'}] & \\ + \sum_{i/z_i=k'} [(\tilde{t}_{il}^{(q)} - \tilde{\pi}_{il})\tilde{S}_{il'k} + (\tilde{t}_{il'}^{(q)} - \tilde{\pi}_{il'})\tilde{S}_{ilk}] & sinon \end{cases}$$

où on a noté $\tilde{S}_{ilk} = \sum_{j\in N_i/z_j=k} \tilde{y}_{jl}$.

Preuve : on a, pout tout $i \in \mathcal{I}$ et tout $y_i \in \mathcal{L}$,

$$\frac{\partial \tilde{H}_i(y_i;\mathbb{B})}{B^{ll'}_{kk'}} = \begin{cases} \sum\limits_{j\in N_i/z_j=k'} \tilde{y}_{jl'} & \text{si } z_i = k \text{ et } y_i = l \\ \sum\limits_{j\in N_i/z_j=k'} \tilde{y}_{jl} & \text{si } z_i = k' \text{ et } y_i = l \\ \sum\limits_{j\in N_i/z_j=k'} \tilde{y}_{jl} & \text{si } z_i = k \text{ et } y_i = l' \\ \sum\limits_{j\in N_i/z_j=k} \tilde{y}_{jl} & \text{si } z_i = k' \text{ et } y_i = l' \\ 0 & \text{sinon} \end{cases}$$

5.2 Etape de classification

Comme décrit au chapitre V, section 3.2, la mise à jour des paramètres $\mathbb{V} = \{(\mathbb{B}_{kk'})_{k,k'\in\mathcal{K}}, \mathbb{C}\}$ s'écrit :

$$\begin{aligned} \mathbb{V}^{(q+1)} &= \arg\max_{\mathbb{V}} \sum_{i\in\mathcal{I}_2}\sum_{l\in\mathcal{L}}\sum_{k\in\mathcal{K}} \tilde{t}^{(q)}_{ilk} \log \tilde{\pi}_{ilk} \\ \text{avec } \tilde{\pi}_{ilk} &= P(U_i = e''_{lk}|\tilde{\mathbf{u}},\mathbb{V}) \propto \exp(\sum_{j\in N_i}\sum_{l'\in\mathcal{L}}\sum_{k'\in\mathcal{K}} (B^{ll'}_{kk'} + c_{kk'})\tilde{u}_{jl'k'}) \end{aligned}$$

où on a noté $c_{kk'}$ la composante (l,l') de la matrice $\mathbb{C}$ et $\tilde{u}_{jlk}$ la composante (l,k) du champ $\tilde{u}_j$ au site j. Notons :

$$\tilde{H}_i(l,k;\mathbb{V}) = \sum_{j\in N_i}\sum_{l'\in\mathcal{L}}\sum_{k'\in\mathcal{K}} (B^{ll'}_{kk'} + c_{kk'})\tilde{u}_{jl'k'}$$

Soit $\tilde{Q}(\mathbb{V}|\boldsymbol{\psi}^{(q)}) = \langle \log P_{\tilde{y}}(\mathbf{Y}|\mathbf{z};\mathbb{V})|\mathbf{x},\boldsymbol{\psi}^{(q)}\rangle$ la fonction de $\mathbb{V}$ à maximiser à l'itération $(q+1)$ de NREM et $\boldsymbol{\psi}^{(q)}$ la valeur des paramètres $(\mathbb{V},\boldsymbol{\theta})$ courants. D'après la Proposition 24, le gradient de $\tilde{Q}$ par rapport à $\mathbb{V}$ est donné par :

$$\nabla\tilde{Q}(\mathbb{V}|\boldsymbol{\psi}^{(q)}) = \sum_{i\in\mathcal{I}_2}\sum_{l\in\mathcal{L}} (\tilde{t}^{(q)}_{il} - \tilde{\pi}_{il})\nabla\tilde{H}_i(l,k;\mathbb{V})$$

Proposition 28. *L'estimation des paramètres $\mathbb{V} = (\mathbb{B},\mathbb{C})$ à l'étape (M) de l'algorithme NREM peut se faire par descente de gradient. Les dérivées sont données par :*

$$\frac{\partial\tilde{Q}}{B^{ll'}_{kk'}} = \begin{cases} \sum\limits_{i\in\mathcal{I}_2} (\tilde{t}^{(q)}_{ilk} - \tilde{\pi}_{ilk})\tilde{S}_{ilk} & \textit{si } k = k' \textit{ et } l = l' \\ \sum\limits_{i\in\mathcal{I}_2} [(\tilde{t}^{(q)}_{ilk} - \tilde{\pi}_{ilk})\tilde{S}_{il'k} + (\tilde{t}^{(q)}_{il'k} - \tilde{\pi}_{il'k})\tilde{S}_{ilk}] & \textit{si } k = k' \textit{ et } l \neq l' \\ \sum\limits_{i\in\mathcal{I}_2} (\tilde{t}^{(q)}_{ilk} - \tilde{\pi}_{ilk})\tilde{S}_{ilk'} + \sum\limits_{i\in\mathcal{I}_2} (\tilde{t}^{(q)}_{ilk'} - \tilde{\pi}_{ilk'})\tilde{S}_{ilk} & \textit{si } k \neq k' \textit{ et } l = l' \\ \sum\limits_{i\in\mathcal{I}_2} [(\tilde{t}^{(q)}_{il'k} - \tilde{\pi}_{il'k})\tilde{S}_{ilk} + (\tilde{t}^{(q)}_{ilk} - \tilde{\pi}_{ilk})\tilde{S}_{il'k'}] & \\ + \sum\limits_{i\in\mathcal{I}_2} [(\tilde{t}^{(q)}_{ilk'} - \tilde{\pi}_{ilk'})\tilde{S}_{il'k} + (\tilde{t}^{(q)}_{il'k'} - \tilde{\pi}_{il'k'})\tilde{S}_{ilk}] & \textit{sinon} \end{cases}$$

$$\frac{\partial\tilde{Q}}{c_{kk'}} = \begin{cases} \sum\limits_{i\in\mathcal{I}_2}\sum\limits_{l\in\mathcal{L}} (\tilde{t}^{(q)}_{ilk} - \tilde{\pi}_{ilk})(\sum_{l'\in\mathcal{L}} \tilde{S}_{il'k}) & \textit{si } k = k' \\ \sum\limits_{i\in\mathcal{I}_2} [\sum\limits_{l\in\mathcal{L}} (\tilde{t}^{(q)}_{ilk} - \tilde{\pi}_{ilk})(\sum_{l'\in\mathcal{L}} \tilde{S}_{il'k}) + \sum\limits_{l\in\mathcal{L}} (\tilde{t}^{(q)}_{ilk'} - \tilde{\pi}_{ilk'})(\sum_{l'\in\mathcal{L}} \tilde{S}_{il'k})] & \textit{sinon} \end{cases}$$

où on a noté $\tilde{S}_{ilk} = \sum_{j \in N_i} \tilde{u}_{jlk}$

Preuve : On a, pout tout $i \in \mathcal{I}_2$ et tout $u_i \in \mathcal{L} \times \mathcal{K}$,

$$\frac{\partial \tilde{H}_i(u_i; \mathbb{B})}{B^{ll'}_{kk'}} = \begin{cases} \sum_{j \in N_i} \tilde{u}_{jl'k'} & \text{si } u_i = (l,k) \\ \sum_{j \in N_i} \tilde{u}_{jl'k} & \text{si } u_i = (l,k') \\ \sum_{j \in N_i} \tilde{u}_{jlk'} & \text{si } u_i = (l',k) \\ \sum_{j \in N_i} \tilde{u}_{jlk} & \text{si } u_i = (l',k') \\ 0 & \text{sinon} \end{cases}$$

$$\frac{\partial \tilde{H}_i(u_i; \mathbb{B})}{c_{kk'}} = \begin{cases} \sum_{l' \in \mathcal{L}} \sum_{j \in N_i} \tilde{u}_{jl'k'} & \text{si } u_i \in \{(l,k), l \in \mathcal{L}\} \\ \sum_{l' \in \mathcal{L}} \sum_{j \in N_i} \tilde{u}_{jl'k} & \text{si } u_i \in \{(l,k'), l \in \mathcal{L}\} \\ 0 & \text{sinon} \end{cases}$$

6 Le logiciel SpaCEM^3

Mentionnons que les résultats expérimentaux présentés dans cette thèse s'appuient sur un travail de programmation, le logiciel SpaCEM^3 (*Spatial Clustering with EM and Markov Models*), disponible à l'adresse :

`http://mistis.inrialpes.fr/software/SpaCEM3.html`.

Cette librairie écrite en C++ propose une variété d'algorithmes pour la classification, supervisée ou non supervisée d'individus en interaction, sur lesquels sont mesurés des données uni ou multidimensionnelles. Ceci inclut la segmentation d'images, avec comme structure de dépendance sous-jacente des grilles régulières de pixels. Plus généralement, les algorithmes implémentés permettent de classer des données multimodales et dépendantes du fait de leur localisation spatiale ou du fait d'autres types de relations décrites par des structures graphiques quelconques.

L'approche principale se fonde sur l'utilisation de l'algorithme EM (ou sur approximation en champ moyen NREM) pour une classification *floue* et sur les modèles de champs de Markov pour la modélisation des dépendances. Les fonctionnalités de SpaCEM^3 incluent les points suivants :

- Classification non supervisée d'individus, basée sur une description des dépendances à l'aide d'un graphe non nécessairement régulier et un traitement basé sur les champs de Markov cachés. Les modèles markoviens disponibles incluent diverses extensions du modèle de Potts standard, notamment avec la possibilité d'utiliser des modèles d'interaction plus généraux.

- Classification supervisée d'individus, basée sur la famille de modèles de Markov triplets décrit au chapitre V, avec des phases d'apprentissage et de test.

- Critère de sélection de modèles (BIC, ICL et leurs approximations en champ moyen) permettant de sélectionner "le meilleur" modèle de champs de Markov cachés en fonction des données.

- Simulation de modèles de champs de Markov généraux avec interactions d'ordre 1 et 2 (modèle de Potts et ses diverses extensions).

- Simulation de modèles de champs de Markov triplets généraux avec interactions d'ordre 1 et 2 (chapitre V).

- Simulation de modèles de champs de Markov caché à bruit indépendant (gaussien)

- Simulation de modèles de champs de Markov triplet à bruit indépendant (gaussien)

- Possibilité de traiter le cas de données de très grande dimension dans un cadre markovien, comme décrit au chapitre III.

- Possibilité de traiter le cas d'observations manquantes avec imputation *off-line* (par

les KNN, la moyenne), *on-line* (au cours de l'algorithme) ou sans imputation (algorithme décrit au chapite VII).

De manière générale, la librairie se présente sous la forme de classes héritant les unes des autres. En l'état actuel des choses, le logiciel n'est pas doté d'interface utilisateur. On donne ci-après un exemple d'utilisation pour l'estimation et la classification d'observations éventuellement manquantes sous modélisation markovienne.

```
// crée des données spatiales 'sdat' par lecture du fichier des observations
// (certaines peuvent être manquantes) et du graphe de voisinage
Spatial_Data *sdat = new Spatial_Data();
sdat -> ReadFromFile("mes_donnees");

// retourne le graphe de voisinage dans 'nei'
Neighborhood_System *nei = new Neighborhood_System();
nei = sdat -> Get_NS();

// retourne le nombre de sites 'N' and la dimension 'D' des donnees
uint N = sdat->Get_N();
uint D = sdat->Get_D();

// crée un vecteur 'gauss' de K=4 gaussiennes diagonales en dimension D
uint K=4;
vector<Diag_Normal*> gauss(K);
for (uint k=0;k<K;k++){
  gauss[k] = new Diag_Normal(D);
}

// crée un champ de Markov (modèle de Potts sans champ externe) 'mrf'
// à K=4 couleurs (classes) et legraphe de voisinage 'nei' des données
B_MRF * mrf = new B_MRF(nei,K);

// crée un champ de Markov caché 'hmrf'
HMRF * hmrf = new HMRF(mrf,gauss);

// crée un algorithme de segmentation 'algo' (champ simulé)
Simulated_Field_EM *algo = new  Simulated_Field_EM(hmrf,sdat);

// initialise l'algorithme par K-means
algo->Init_KMeans();

// 10 iterations de l'algorithme
algo->Run(100);

// retourne la classification MAP dans le vecteur vector 'map_labels'
vector<uint> map_labels;
algo->Compute_MAP_Labels(map_labels);
```

```
// sauve la classification dans le fichier 'mes_classes.cls'
ofstream file_labels("mes_classes.cls");
for (uint i=0;i<N;i++) file_labels << map_labels[i] << endl;
file_labels.close();

// écrit à l'écran la valeur du critère BICw
double bic=algo->BICw();
cerr << "valeur du critère BICw = " << bic << endl;

// sauve les paramètres du modèle dans le fichier 'mes_params.par'
hmrf->WriteToFile("mes_params.par");
```

Bibliographie générale

[1] T. Aach, A. Kaup, and R. Mester, *Statistical model-based change detection in moving video*, Signal Processing, 31 (1993), pp. 165–180.

[2] K. Abend, T. Harley, and L. Kanal, *Classification of binary patterns*, IEEE Transactions on Information Theory, 11 (1965), pp. 538–544.

[3] M. Aitkin and D. B. Rubin, *Estimation and hypothesis testing in Finite Mixture Models*, Journal of the Royal Statistical Society, Series B (Statistical Methodology), 47 (1985), pp. 67–75.

[4] H. Akaike, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, 19 (1974), pp. 716–723.

[5] J. Banfield and A. Raftery, *Model-based Gaussian and non-Gaussian clustering*, Biometrics, 49 (1993), pp. 803–821.

[6] D. Benboudjema and W. Pieczynski, *Unsupervised image segmentation using triplet Markov Fields*, Computer Vision and Image understanding, 99 (2005), pp. 476–498.

[7] D. Benboudjema and W. Pieczynski, *Unsupervised statistical segmentation of non stationary images using triplet Markov Fields*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (2007), pp. 367–1378.

[8] H. Bensmail and G. Celeux, *Regularized Gaussian discrimant analysis through eigenvalue decomposition*, Journal of the American Statistical Association, 91 (1996), pp. 1743–1748.

[9] J. Besag, *Spatial interaction and the statistical analysis of lattice systems*, Journal of the Royal Statistical Society, 35 (1974), pp. 192–236.

[10] ——, *Spatial analysis of dirty pictures*, Journal of the Royal Statistical Society, 48 (1986), pp. 259–302.

[11] J. C. Bezdek, *Numerical taxonomy with fuzzy sets*, Journal of Mathematical Biology, 1 (1974), pp. 57–71.

[12] C. Biernacki, *Précision sur les données et coude de la vraisemblance pour trouver le nombre de classes dans un mélange*, Revue de Statistique Appliquée, 47 (1999), pp. 47–62.

[13] C. Biernacki, G. Celeux, and G. Govaert, *Assessing a Mixture model for clustering with the integrated completed likelihood*, Transactions on Pattern Analysis and Machine Intelligence, 22 (2000), pp. 719–725.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[15] J. Blanchet and F. Forbes, *Triplet markov field designed for supervised classification of texture images*, in Proceedings in computational statistics (COMPSTAT), 2006.

[16] ——, *Triplet markov fields for the supervised classification of complex structured data*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (Submitted).

[17] J. Blanchet and M. Vignes, *Combined expression data with missing values and gene interaction network analysis : a Markovian integrated approach*, in Proceedings IEEE International Symposium on BioInformatics and BioEngineering (BIBE), 2007.

[18] T. H. Bø, B. Dysvik, and I. Jonassen, *LSimpute : accurate estimation of missing values in microarray data with least square methods*, Nucleic Acids Research, 32 (2004), p. e34.

[19] D. Bouchaffra and J. Meunier, *A Markovian Random Field approach to information retrieval*, in Proceedings of the International Conference on Document Analysis and Recognition, 1995, pp. 997– 1002.

[20] G. Bouchard and G. Celeux, *Model selection in supervised classification*, Transactions on Pattern Analysis and Machine Intelligence, 28 (2005), pp. 544–554.

[21] C. Bouman and M. Shapiro, *A Multiscale Random Field model for Bayesian image segmentation*, IEEE Transaction on Image Processing, 3 (1994), pp. 162–177.

[22] P. Bouthemy and E. François, *Motion segmentation and qualitative dynamic scene analysis from an image sequence*, International Journal of Computer Vision, 10 (1993), pp. 157–182.

[23] C. Bouveyron, *Modélisation et classification des données de grande dimension. Application à l'analyse d'images*, PhD thesis, Université Joseph Fourier, Grenoble I, 2006.

[24] E. Bredensteiner and K. Bennett, *Multicategory classification by Support Vector Machines*, Computational Optimization and Applications, 12 (1999), pp. 53–79.

[25] R. Catell, *The scree test for the number of factors*, Multivariate Behavioral Research, 1 (1966), pp. 245–276.

[26] G. Celeux, *Bayesian inference for Mixtures : the label-switching problem*, in COMPSTAT 98, 1998, pp. 227–232.

[27] G. Celeux and J. Diebolt, *The SEM algorithm : A probabilistic teacher algorithm derived from the EM algorithm for the Mixture problem*, Computational Statistics Quarterly, 2 (1985), pp. 73–82.

[28] ——, *A stochastic approximation type EM algorithm for the Mixture problem*, Stochastics and Stochastics Reports, 41 (1992), pp. 119–134.

[29] G. Celeux, F. Forbes, and N. Peyrard, *EM procedures using Mean Field-like approximations for Markov model-based image segmentation*, Pattern Recognition, 36 (2003).

[30] B. Chalmond, *An iterative Gibbsian technique for reconstruction of m-ary images*, Pattern Recognition, 22 (1989), pp. 747–762.

[31] D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, 1987.

[32] R. Chellappa, *Two-dimensional discrete Gaussian Markov Random Field models for image processing and analysis*, in Proceedings of SPIE, Digital Image Processing Applications, Y.-W. Lin and R. Srinivasan, eds., vol. 1075, Avril 1989, pp. 336–+.

[33] R. Chellappa, S. Chatterjee, and R. Bagdazian, *Texture synthesis and compression using Gaussian-Markov Random Field models*, IEEE Transactions on Systems, Man and Cybernetics, 15 (1985), pp. 298–303.

[34] R. Chellappa and A. Jain, *Markov Random Fields : theory and applications*, Academic Press, inc., 1993.

[35] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, *A genome-wide transcriptional analysis of the mitotic cell cycle*, Molecular cell, 2 (1998), pp. 65–73.

[36] F. S. Cohen and D. B. Cooper, *Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian Fields*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, (1987), pp. 195–219.

[37] L. Collins, J. Schafer, and C. Kam, *A comparison of inclusive and restrictive strategies in modern missing data procedures*, Psychological Methods, 6 (2001), pp. 330–351.

[38] K. Crammer and Y. Singer, *On the algorithmic implementation of multiclass kernel-based vector machines*, Journal of Machine Learning Research, 2 (2001), pp. 265–292.

[39] G. R. Cross and A. K. Jain, *Markov Random Field texture models*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, 5 (1983), pp. 25–39.

[40] A. Cutler and M. Windham, *Information-based validity functionals for Mixture analysis*, in Proceedings of the first US-Japan Conference on the Frontiers of Statistical Modeling, 1993, pp. 149–170.

[41] M. V. Dang, *Classification de Données Spatiales : Modèles Probabilistes et Critères de Partitionnement*, PhD thesis, Université de Technologie de Compiègne, 1998.

[42] I. de S. Pool and M. Kochen, *Contacts and influence*, Social Networks, 1 (1978), pp. 1–48.

[43] J. Delmas, *An equivalence of the em and ice algorithm for exponential family*, IEEE transactions on signal processing, 45 (1997), pp. 2613–2615.

[44] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, Journal of the Royal Statistical Society B, 39 (1977), pp. 1–38.

[45] H. Derin and H. Elliott, *Modeling and segmentation of noisy and textured images using Gibbs Random Fields*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 9 (1987), pp. 39–55.

[46] E. Diday, *La méthode des nuées dynamiques*, Revue de Statistiques Appliquées, 19 (1971), pp. 19–34.

[47] P. Erdös and A. Rényi, *On random graphs*, Publicationes Mathematicae, 6 (1959), pp. 290–297.

[48] M. Ernst, *A multivariate generalized Laplace distribution*, Computational Statistics, 13 (1998), pp. 227–232.

[49] X. Feng, C. Williams, and S. Felderhof, *Combining Belief Networks and Neural Networks for scene segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2002), pp. 467–483.

[50] F. Forbes and G. Fort, *Combining Monte Carlo and Mean Field like methods for inference in Hidden Markov Random Fields*, IEEE Transactions on Image Processing, 16 (2007), pp. 824–837.

[51] F. Forbes and N. Peyrard, *Hidden Markov Random Field selection criteria based on Mean Field-like approximations*, Transactions on Pattern Analysis and Machine Intelligence, 25 (2003), pp. 1089–1101.

[52] F. Forbes and M. Vignes, *Integrated Markov models for clustering genes combining individual features and pairwise relationships*, in 4th workshop on Statistical methods for post-genomic data, 2006.

[53] C. Fraley and A. E. Raftery, *How many clusters ? Which clustering method ? Answers via model-based cluster analysis*, Computer Journal, 41 (1998), pp. 578–588.

[54] J. Friedman, *Regularized dicriminant analysis*, Journal of the American Statistical Association, 84 (1989), pp. 165–175.

[55] ——, *Another approach to polychotomous classification*, tech. rep., Statistics Department, Stanford University, 1996.

[56] K. R. Gabriel and R. R. Sokal, *A new statistical approach to geographic variation analysis*, Systematic Zoology, 18 (1969), pp. 259–278.

[57] D. Geiger and F. Girosi, *Parallel and deterministic algorithms from MRFs : Surface reconstruction*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13 (1991), pp. 401–412.

[58] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6 (1984), pp. 721–741.

[59] S. Geman and C. Graffigne, *Markov Random Field image models and their applications to computer vision*, in Proceedings of the International Congress of Mathematicians, 1987, pp. 1496–1517.

[60] H. O. Georgii, *Gibbs measures and phase transitions*, Walter de Gruyter, 1988.

[61] Z. Ghahramani and M. I. Jordan, *Supervised Learning from incomplete data via an EM approach*, Advances in Neural Information Processing Systems, 6 (1994), pp. 120–127.

[62] M. Gilloux, *Reconnaissance de chiffres manuscrits par modèle de Markov pseudo-2d*, Traitement du signal, 12 (1995), pp. 561–566.

[63] J. Goutsias, *Markov Random Fields : Interacting particle systems for image modelling and analysis*, tech. rep., Department of Electrical and Computer Engineering, Image Analysis and Communications Laboratory, The Johns Hopkins University, Baltimore, MD, USA, 1996.

[64] G. Govaert, *Classification binaire et modèles*, Revue de Statistique Appliquée, XXXVIII (1990), pp. 67–81.

[65] J. W. Graham, S. M. Hofer, and D. P. MacKinnon, *Maximizing the usefulness of data obtained with planned missing value patterns : An application of maximum likelihood procedures*, Multivariate Behavioral Research, 31 (1996), pp. 197–218.

[66] G. Gravier, M. Sigelle, and G. Cholle, *A Markov Random Field model for speech recognition*, in International Conference on Pattern Recognition, 2000.

[67] T. Hastie, A. Buja, and R. Tibshirani, *Penalized dicriminant analysis*, Annals of Statistics, 23 (1995), pp. 73–102.

[68] T. Hastie and R. Tibshirani, *Discriminant analysis by Gaussian mixtures*, Journal of the Royal Statistical Society series B, 58 (1996), pp. 158–176.

[69] D. Hedeker and R. Gibbons, *Application of random-effects pattern-mixture models for missing data in longitudinal studies*, Psychologocal Methods, 2 (1997), pp. 64–78.

[70] Y. Hung, D. Cooper, and B. Cernuschi-Frias, *Asymptotic Bayesian surface estimation using an image sequence*, International Journal of Computer Vision, 6 (1991), pp. 105–132.

[71] E. Ising, *Beitrag zur theorie des ferromagnetismus*, Zeitschrift für Physik, 31 (1925), pp. 253–258.

[72] T. Jaakkola, *Tutorial on variational approximation methods*, Advanced mean field methods : theory and practice, (2000).

[73] A. K. Jain, *Advances in mathematical models for image processing*, IEEE Proceedings, 69 (1981), pp. 502–528.

[74] A. Jasra, C. C. Holmes, and D. A. Stephens, *Markov Chain Monte Carlo methods and the label switching problem in Bayesian Mixture modeling*, Statistical Science, 20 (2005), pp. 50–67.

[75] N. Jean-Pierre and G. Mina, *Physique statistique de phénomènes collectifs en sciences économiques et sociales*, Mathématiques et sciences humaines, 172 (2005), pp. 67–89.

[76] L. O. Jimenez and D. A. Landgrebe, *Hyperspectral data analysis and supervised feature reduction via projection pursuit*, IEEE Transactions on Geoscience and Remote Sensing, 37 (1999), pp. 2653–2667.

[77] S. C. Johnson, *Hierarchical clustering schemes*, Psychometrika, 2 (1967), pp. 241–254.

[78] S. Kumar and M. Hebert, *Man-Made structure detection in natural images using a Causal Multiscale Random Field*, in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2003.

[79] S. Kumar and M. Hebert, *Discriminative Random Fields*, International Journal of Computer Vision, 68 (2006), pp. 179–201.

[80] P. M. Lankford, *Regionalization : theory and alternative algorithms*, Geographical Analysis, 1 (1969), pp. 196–212.

[81] S. Lauritzen, A. Dawid, B. Larsen, and H.-G. Leimer, *Independence properties of directed Markov Fields*, Networks, 20 (1990), pp. 491–505.

[82] S. Lazebnik, C. Schmid, and J. Ponce, *Affine-invariant local descriptors and neighborhood statistics for texture recognition*, in Proc. ICCV, 2003.

[83] S. Le Hegarat-Mascle, A. Kallel, and X. Descombes, *Ant colony optimization for image regularization based on a non-stationary Markov modeling*, IEEE Transactions on Image Processing, 16 (2007).

[84] L. P. Lefkovitch, *Dynamic Programming*, Princeton University Press, 1957.

[85] T. Lindeberg, *Feature detection with automatic scale selection*, International Journal of Computer Vision, 30 (1998), pp. 77–116.

[86] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, Wiley Series In Probability And Statistics, John Wiley & Sons, New York, 1986.

[87] D. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, 60 (2004), pp. 91–110.

[88] J. MacQueen, *Some methods for classification and analysis of multivariate observations*, in Proceedings Fifth Berkeley Symposium on Mathematical Statistics and Probability, L. M. LeCam and J. N. editors, eds., vol. 1, University of California Press, 1967, pp. 281–297.

[89] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley Series in Probability and Statistics, 2000.

[90] G. J. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing microarray gene expression data*, Probability and Statistics, Wiley, august 2004.

[91] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equations of state calculations by fast computing machine*, Journal of Chemical Physics, 21 (1953), pp. 1087–1091.

[92] D. Metzler and W. B. Croft, *A Markov Random Field model for term dependencies*, in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 472–479.

[93] K. Mikolajczyk and C. Schmid, *Scale and affine invariant interest point detectors*, International Journal of Computer Vision, 60 (2004), pp. 63–86.

[94] M. Newman, *The structure and function of complex networks*, SIAM Review, 45 (2003), pp. 167–256.

[95] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Physical Review E, 69 (2004), p. 026113.

[96] M. Ouyang, W. J. Welsh, and P. Georgopoulos, *Gaussian Mixture clustering and imputation of microarray data*, Bioinformatics, 20 (2004), pp. 917–923.

[97] N. Peyrard, *Approximations de type champ moyen des modèles de champ de Markov pour la segmentation de données spatiales*, PhD thesis, Université Joseph Fourier - Grenoble I, 2001.

[98] J. Pickens and C. Iliopoulos, *Markov Random Fields and Maximum Entropy modeling for music information retrieval*, in Proceedings of the 6th International Conference on Music Information Retrieval, 2005, pp. 207–214.

[99] W. Pieczynski, *Champs de Markov cachés et estimation conditionnelle itérative*, Traitement du Signal, 11 (1994), pp. 141–153.

[100] W. Pieczynski and A. Tebbache, *Pairwise Markov Random Fields and segmentation of textured images*, Machine Graphics and Vision, 9 (2000), pp. 705–718.

[101] W. Qian and D. Titterington, *Stochastic relaxations and EM algorithms for Markov Random Fields*, Journal of Statistical Computation and Simulation, 40 (1992), pp. 55–69.

[102] W. Qian and D. M. Titterington, *Estimation of parameters in Hidden Markov Models*, Royal Society of London Philosophical Transactions Series A, 337 (1991), pp. 407–428.

[103] ——, *Estimation of parameters in Hidden Markov Models*, Royal Society of London Philosophical Transactions Series A, 337 (1991), pp. 407–428.

[104] G. Rellier, X. Descombes, F. Falzon, and J. Zerubia, *Classification de textures hyperspectrales fondée sur un modèle markovien et une technique de poursuite de projection*, Traitement du Signal, 20 (2003), pp. 25–42.

[105] C. Robert and G. Casella, *Monte Carlo statistical methods*, Springer, 1999.

[106] C. R.R. Sokaland Michener, *A statistical method for evaluating systematic relationships*, University of Kansas science bulletin, 38 (1958), pp. 1409–1438.

[107] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, 1987.

[108] D. B. Rubin, *Inference and missing data*, Biometrika, 63 (1976), pp. 581–592.

[109] G. Saporta, *Probabilités, analyse des données et statistique*, 2, editions technip ed., 2006.

[110] S. Saquib, C. Bouman, and K. Sauer, *ML parameter estimation for Markov Random Fields, with applications to Bayesian tomography*, IEEE Transactions on Image Processing, 7 (1998), pp. 1029–1044.

[111] J. Schafer, *Analysis Of Incomplete Multivariate Data*, Chapman & Hall, London, 1997.

[112] J. L. Schafer and J. W. Graham, *Missing data : our view of the state of art*, Psychologocal Methods, 7 (2002), pp. 147–177.

[113] B. Scherrer, M. Dojat, F. Forbes, and C. Garbay, *LOCUS : LOcal Cooperative Unified Segmentation of MRI brain scans*, in MICCAI, Brisbane, Australia, 2007.

[114] G. Schwarz, *Estimating the dimension of a model*, The Annals of Statistics, 6 (1978), pp. 461–464.

[115] D. W. Scott and J. R. Thompson, *Probability density estimation in higher dimensions*, in Computer Science and Statistics : Proceedings of the Fifteenth Symposium on the Interface, 1983, pp. 173–179.

[116] B. Snel, P. Bork, and M. A. Huynen, *The identification of functional modules from the genomic association of genes*, Proceedings of the National Academy of Sciences of the United States of America, 99 (2002), pp. 5890–5895.

[117] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, *Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization*, Molecular Biology of the Cell, 9 (1998), pp. 3273–3297.

[118] M. Stephens, *Bayesian Methods for Mixtures of Normal Distributions*, PhD thesis, University of Oxford, 1997.

[119] ——, *Dealing with label switching in Mixture Models*, Journal of the Royal Statistical Society, Series B (Statistical Methodology), 62 (2000), pp. 795–809.

[120] D. Strauss, *Clustering on coloured lattices*, Journal of Applied Probability, 14 (1977), pp. 135–143.

[121] ——, *On a general class of models for interaction*, SIAM Review, 28 (1986), pp. 513–527.

[122] G. Toussaint, *The relative neighborhood graph of a finite planar set*, Pattern Recognition, 12 (1980), pp. 261–268.

[123] O. Troyanskaya, M. Cantor, GavinSherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, *Missing values estimation methods for DNA microarrays*, Bioinformatics, 17 (2001), pp. 520–525.

[124] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1999.

[125] M. Verleysen, *Learning high-dimensional data*, Limitations and future trends in neural computation, ios press ed., 2004.

[126] M. Vignes and F. Forbes, *Gene clustering via integrated Markov models combining individual and pairwise features*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, en révision (2007).

[127] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, *STRING : known and predicted protein-protein associations, integrated and transferred across organisms.*, Nucleic Acids Res, 33 (2005).

[128] J. H. Ward, *Hierarchical grouping to optimize an objective function*, Journal of the American Statistical Association, 58 (1963), pp. 236–244.

[129] G. Wei and M. Tanner, *A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms*, Journal of the American Statistical Association, 85 (1990), pp. 699–704.

[130] C. Wu, *On the convergence properties of the EM algorithm*, Annals of Statistics, 11 (1983), pp. 95–103.

[131] C. Wu and P. Doerschuk, *Cluster Expansions for the deterministic computation of Bayesian estimators based on Markov Random Fields*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 17 (1995), pp. 275–293.

[132] S. Yakowitz and J. D. Spragins, *On the identifiability of finite Mixtures*, Annals of Mathematics and Statistics, 39 (1968), pp. 209–214.

[133] J. S. Yedidia, W. T. Freeman, and Y. Weiss, *Understanding belief propagation and its generalizations*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[134] L. Younes, *Estimation and annealing for Gibbsian Fields*, Annales de l'Institut Henri Poincaré (B), Probabilités et Statistique, 24 (1988), pp. 269–294.

[135] ——, *Parametric inference for imperfectly observed Gibbsian Fields*, Probability Theory and Related Fields, 82 (1989), pp. 625–645.

[136] J. Zhang, *The Mean Field theory in EM procedures for Markov Random Fields*, IEEE Transactions on signal processing, 40 (1992), pp. 2570–2583.