Analyse statistique d'IRM quantitatives par modèles de mélange : application à la localisation et la caractérisation de tumeurs cérébrales

Alexis Arnaud

Communauté UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques Appliquées

Arrêté ministériel : 25 mai 2016

Présentée par

Alexis ARNAUD

Thèse dirigée par **Florence FORBES**, INRIA
et codirigée par **Emmanuel BARBIER**, INSERM

préparée au sein du **Laboratoire Jean Kuntzmann**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Analyse statistique d'IRM quantitatives par modèles de mélange : Application à la localisation et la caractérisation de tumeurs cérébrales

Statistical analysis of quantitative MRI based on mixture models: Application to the localization and characterization of brain tumors

Thèse soutenue publiquement le **24 octobre 2018**,
devant le jury composé de :

Madame FLORENCE FORBES
DIRECTRICE DE RECHERCHE, INRIA CENTRE DE GRENOBLE RHÔNE-ALPES, Directeur de thèse
Monsieur EMMANUEL BARBIER
DIRECTEUR DE RECHERCHE, INSERM DELEGATION ALPES, Co-directeur de thèse
Monsieur CHARLES BOUVEYRON
PROFESSEUR, UNIVERSITE COTE D'AZUR, Rapporteur
Monsieur FABRICE HEITZ
PROFESSEUR, UNIVERSITE DE STRASBOURG, Rapporteur
Madame CAROLE LARTIZIEN
DIRECTRICE DE RECHERCHE, UNIVERSITE DE LYON, Président
Monsieur OLIVIER SAUT
DIRECTEUR DE RECHERCHE, CNRS DELEGATION AQUITAINE, Examinateur
Monsieur RUSSELL STEELE
PROFESSEUR ASSOCIE, UNIVERSITE MCGILL QUEBEC - CANADA, Examinateur
Monsieur BENJAMIN LEMASSON
CHARGE DE RECHERCHE, INSERM DELEGATION ALPES, Examinateur

TABLE DES MATIÈRES

Un très grand merci à mes chers directeurs de thèse qui m'ont permis de mener à bien cette thèse en étant à mes côtés tout au long de cette aventure !

À Miyulu, ma petite luciole.

Alexis Arnaud

Analyse statistique d'IRM quantitatives par modèles de mélange
Application à la localisation et la caractérisation de tumeurs cérébrales

Résumé

Nous présentons dans cette thèse une méthode générique et automatique pour la localisation et la caractérisation de lésions cérébrales telles que les tumeurs primaires à partir de multiples contrastes IRM. Grâce à une récente généralisation des lois de probabilités à mélange par l'échelle de distributions gaussiennes, nous pouvons modéliser une large variété d'interactions entre les paramètres IRM mesurés, et cela afin de capter l'hétérogénéité présente dans les tissus cérébraux sains et endommagés. En nous basant sur ces lois de probabilités, nous proposons un protocole complet pour l'analyse de données IRM multi-contrastes : à partir de données quantitatives, ce protocole fournit, s'il y a lieu, la localisation et le type des lésions détectées au moyen de modèles probabilistes. Le protocole proposé a été validé sur des données IRM chez le petit animal et publié dans la revue IEEE Transactions on Medical Imaging. Le passage à l'humain est en cours, ainsi que la transposition à d'autres pathologies. Nous proposons également deux extensions de ce protocole. La première extension concerne la sélection automatique du nombre de composantes au sein du modèle probabiliste, sélection réalisée via une représentation bayésienne des modèles utilisés. La seconde extension traite de la prise en compte de la structure spatiale des données IRM par l'ajout d'un champ de Markov latent au sein du protocole développé.

Mots clés : modèle de mélange, détection d'anomalie, caractérisation automatique, analyse bayésienne, sélection de modèles bayésiens, champ de markov aléatoire, algorithme em, approximation variationnelle, imagerie par résonance magnétique, tumeur cérébrale

---

Statistical analysis of quantitative MRI based on mixture models
Application to the localization and characterization of brain tumors

Abstract

We present in this thesis a generic and automatic method for the localization and the characterization of brain lesions such as primary tumor using multi-contrast MRI. From the recent generalization of scale mixtures of Gaussians, we reach to model a large variety of interactions between the MRI parameters, with the aim of capturing the heterogeneity inside the healthy and damaged brain tissues. Using these probability distributions we propose an all-in-one protocol to analyze multi-contrast MRI: starting from quantitative MRI data this protocol determines if there is a lesion and in this case the localization and the type of the lesion based on probability models. The proposed protocol has been validated on small animal MRI data and published in the IEEE Transactions on Medical Imaging journal. Trials in humans are in progress, as the study of other pathologies. We also develop two extensions for this protocol. The first one concerns the selection of mixture components in a Bayesian framework. The second one is about taking into account the spatial structure of MRI data by the addition of a random Markov field to our protocol.

Keywords: mixture model, anomaly detection, automatic characterization, bayesian analysis, bayesian model selection, markov random field, em algorithm, variational approximation, magnetic resonance imaging, brain tumor

CHAPITRE 1

INTRODUCTION

1.1 Limites actuelles dans l'analyse des images IRM dans le contexte des tumeurs cérébrales

Le Centre International de Recherche sur le Cancer ("International Agency for Research on Cancer") estime à 14.1 millions[1] le nombre de nouveaux cas de cancer en 2012 (hors cancer de la peau non mélanome), associé à une mortalité de 8.2 millions de cas pour la même année. Pour le cerveau et le système nerveux, le nombre de nouveaux cas était de 256 000 pour 189 000 décès. L'imagerie par résonance magnétique (IRM) est devenue la modalité d'imagerie de référence pour la détection et le suivi de tumeurs cérébrales (De Angelis [1], Drevelegas and Papanikolaou [2], Wen et al [3]).

Au cours des dernières décennies, de nombreuses approches ont été développées pour l'IRM afin d'obtenir de nouveaux contrastes pour cartographier des informations nouvelles sur la nature et le fonctionnement des tissus observés. Ces nouvelles méthodes permettent également de mettre en évidence davantage de caractéristiques tumorales. Aujourd'hui, lors d'un examen IRM classique pour un patient porteur d'une tumeur cérébrale, plusieurs images, ou contrastes, sont acquises ([4]). Une acquisition standard comporte une image pondérée T1 (anatomie cérébrale), une image pondérée T2 (anatomie cérébrale, sensible notamment à l'œdème), une image FLAIR (Fluid-Attenuated Inversion Recovery : séquence qui permet de bien séparer le liquide céphalorachidien des zones de tissu œdémateux), une image de perfusion (évaluation de l'angiogenèse), et une image pondérée T1 après injection d'un agent de contraste (évaluation des régions où la paroi micro-

1. Site internet de l'IARC pour l'outil *Cancer today* : http ://gco.iarc.fr

vasculaire est altérée). On peut parler ici d'IRM multiparamétrique (IRMm). En effet, à chaque voxel imagé est associé plusieurs paramètres physiques ou physiologiques. Pour caractériser au mieux un tissu, on peut souhaiter acquérir le maximum d'information possible. Techniquement, les vitesses d'acquisition de données progressant régulièrement, il est possible d'acquérir de plus en plus d'information au cours d'une session IRM typique, soit environ 30 min. Toutefois, l'acquisition, au cours d'une même session, de nombreuses images IRM différentes pose le problème de l'interprétation conjointe de ces différents contrastes. Aujourd'hui, un radiologue est formé à interpréter, par inspection visuelle et en quelques minutes, 4 à 5 contrastes simultanément. L'augmentation du nombre de contrastes proposés au radiologue entrainerait une forte augmentation du temps d'interprétation et une plus grande variabilité dans l'élaboration des diagnostics, que cela soit dû aux écarts de lecture entre les radiologues ou à la subjectivité des analyses. Enfin, dans les rares cas où il est nécessaire de quantifier une information, le dessin d'une région d'intérêt est réalisé sur un des contrastes, au mieux avec une correction à partir d'un autre contraste. Pour un radiologue, contourer une région en tenant compte de plus d'un contraste serait une opération beaucoup trop chronophage. Dans le cas d'une lésion hétérogène, un unique contourage ne permet pas d'interpréter l'hétérogénéité de la lésion. En résumé, même s'il est techniquement possible d'acquérir davantage d'information sur les tissus, il n'existe pas aujourd'hui d'outil, utilisé en routine clinique, qui permettrait d'exploiter plus efficacement la richesse d'information contenue dans des protocoles d'IRMm.

Dans le cas des tumeurs cérébrales de hauts grades (les glioblastomes), l'imagerie joue plusieurs rôles. Le premier est de contribuer à poser le diagnostic. L'IRM permet de déterminer grossièrement le type de lésion. Le second rôle de l'IRM est de guider une éventuelle biopsie. Cette biopsie permettra aux anatomopathologistes de réaliser le diagnostic de la tumeur. C'est ce diagnostic, réalisé à partir de la biopsie, qui fait aujourd'hui référence dans la prise en charge du patient. Enfin, en fonction du diagnostic, l'IRM contribuera à guider soit une intervention chirurgicale (résection) soit une radiothérapie. Dans les deux cas, la définition du contour de la zone tumorale est un enjeu très sensible. Pour réaliser cette opération, plusieurs contrastes qui font consensus sont acquis en routine clinique ([4]). Les analyses d'images recommandées par la communauté médicale se résument à l'exploitation de quelques caractéristiques des données, telles que le volume d'une lésion ou son plus grand diamètre ([3]) ou encore le rehaussement éventuel de la lésion après injection d'un agent de contraste. La communauté des chercheurs propose un grand nombre de méthodes d'analyse plus avancées mais, comme il n'existe pas de consensus méthodologique sur ces questions, ces méthodes ne sont pas exploitées en clinique. Pourtant, les méthodes d'analyse cliniques actuelles sont limitées. Par exemple, l'hétérogénéité d'une tumeur n'est pas quantifiée. L'hétéro-

généité des lésions en imagerie IRM est simplement prise en compte pour guider la biopsie vers les zones de forte agressivité afin d'établir le diagnostic grâce à une analyse histologique. Il serait plus intéressant de connaître tous les sous-types tissulaires qui composent l'ensemble de la tumeur pour optimiser la prise en charge thérapeutique.

L'approche par biopsie pose également des problèmes. La biopsie, c'est-à-dire le prélèvement d'un petit volume de tissu d'environ 1 mm^3, n'est pas toujours faisable du fait de la position critique de la tumeur, et est difficilement répétable dans le temps (soit on re-biopsie ailleurs, et dans ce cas, on ne caractérise pas l'évolution de la tumeur, soit on re-biopsie au même endroit et alors le tissu obtenu est probablement cicatriciel). De plus, bien que la biopsie apporte une information très détaillée au nivau cellulaire ou sub-cellulaire (avec une information de type génétique ou protéomique par exemple), ce prélèvement n'est qu'un fragment de la tumeur, et les informations recueillies sont nécessairement très localisées. La tumeur étant hétérogène, il est connu que le diagnostic variera en fonction de la localisation de la biopsie ([5]). De plus, l'inspection des lames histologiques est visuelle. En effet, les anatomopathologistes n'utilisent pas de méthodes d'analyse d'images automatiques. Une conséquence est que le diagnostic peut être très variable en fonction du médecin qui lit les lames (critères RECIST 1.1 [6]). La communauté médicale est très consciente des limites du diagnostic par biopsie et manque d'un outil plus performant.

Nous avons vu plus haut que l'IRM permet l'acquisition de nombreuses informations, couvre tout un organe, et permet un suivi temporel. Même si l'IRM ne permet pas d'atteindre la résolution spatiale de la biopsie et ne permet pas non plus d'obtenir les mêmes informations que la biopsie, l'IRM apparait depuis longtemps comme la modalité d'imagerie idéale pour proposer un complément de diagnostic à l'approche par biopsie. Ce potentiel s'est encore renforcé au cours des deux dernières décennies grâce au développement de nouvelles techniques IRM permettant de cartographier de nouvelles informations. Comme indiqué plus haut, le problème qui se pose depuis plus de vingt ans porte sur l'interprétation des données. Avec la multiplication des informations disponibles, la complexité du problème n'a fait que se renforcer.

En parallèle du développement des méthodes d'acquisition de données IRM, les méthodes statistiques permettant d'analyser les images médicales ont fortement progressé. La littérature distingue généralement deux étapes dans l'analyse d'images. La première est la détection de la lésion, ou segmentation. Manuellement, le radiologue contoure à l'aide de sa souris d'ordinateur une région ("Region Of Interest" : ROI). Ce contourage (ou un ensemble de contourages réalisés par plusieurs experts) est utilisé comme vérité terrain par les équipes qui développent des méthodes de segmentation automatique. La seconde étape est l'interprétation.

À partir du contenu en information de la (ou les) ROI, des méthodes mathématiques identifient des types tissulaires (tumeur, nécrose etc.) par rapport à des informations de référence (atlas, connaissances d'experts, apprentissage etc.).

Pour l'étape de localisation, des méthodes basées sur des modèles génératifs (tels que des modèles physiques de croissance de tumeurs ou des modèles statistiques de mélange) ou discriminants (tels que les méthodes k-means) ont été proposées afin de segmenter la tumeur au sein d'une ou plusieurs images IRM. Cependant, soit l'apprentissage se fait de façon supervisée et nécessite alors un jeu de données entièrement étiqueté, soit l'apprentissage est non supervisé et il est alors nécessaire d'ajouter des contraintes de façon à différencier les zones tumorales des zones saines. Ces contraintes sont issues de connaissances d'experts et sont donc potentiellement entachées de subjectivité.

La partie caractérisation est généralement appliquée à une segmentation manuelle de la zone tumorale. En présence de contrastes IRM différents, la segmentation manuelle est réalisée sur un des contrastes, en prenant éventuellement en compte un second contraste pour corriger le tracé. De plus, les contrastes pour lesquels le radiologue est expert sont les images acquises en routine clinique. Les nouveaux contrastes issus de calculs ou de développements méthodologiques sont donc plus difficiles à intégrer par l'expert dans le contourage manuel.

Aujourd'hui, la très grande majorité des études portent sur l'une ou l'autre de ces étapes : les méthodes de segmentation des tumeurs exploitent uniquement quelques cartes d'IRM standard, tandis que les méthodes de caractérisation de la tumeur reposent généralement sur l'hypothèse qu'il existe une segmentation manuelle préalable. Il existe pourtant une dépendance entre les deux étapes. L'étape de segmentation repose sur des hypothèses : on pourra segmenter la partie de la tumeur qui prend le contraste, ou celle qui présente un œdème, ou encore prendre en compte toute anomalie dans les images. Le choix des critères de contourage a donc des conséquences sur la partie de caractérisation. En séparant les deux étapes dans les études, on considère à tort les deux étapes comme indépendantes l'une de l'autre.

1.2 Résumé des contributions et publications

Nous résumons ci-dessous les principaux apports de cette thèse en matière de modèles statistiques, d'estimation de modèles, ainsi que l'application de ces différentes approches dans le domaine de l'aide au diagnostic en imagerie médicale.

Modélisation de l'interaction entre plusieurs cartes IRM.
Nous avons évalué l'utilisation de lois à mélange d'échelles multiples (MST, "Multiple Scaled T-distribution", et MSP "Multiple Scaled Pearson type VII dis-

tribution", Forbes et Wraith [7]) pour la caractérisation de données IRM. Ces lois permettent de modéliser une large variété d'interactions entre les paramètres IRM, notamment lorsque seule une partie de ces paramètres présentent des valeurs extrêmes. C'est notamment le cas des tumeurs où les tissus tumoraux s'écartent radicalement des tissus sains.

Comparaison des lois à mélange d'échelles multiples et des lois gaussiennes.

La comparaison avec le modèle classique de mélange gaussien montre l'importance de l'utilisation des lois de Student à mélange d'échelles multiples (MMST) pour représenter les valeurs IRM mesurées dans les tissus tumoraux. La détection de la zone pathologique est plus fine, tandis que la reconnaissance du type de pathologie est meilleure sur le jeu de test utilisé d'IRM cérébrale petit animal.

Développement d'une chaîne d'analyse pour le diagnostic automatique d'anomalies.

Nous avons développé un protocole original pour l'analyse de données IRM multiparamétriques et quantitatives. Ce protocole se veut modulaire, entièrement automatique et guidé par les données. Il est composé de 5 étapes successives, chaque étape étant basée sur un modèle d'apprentissage statistique pouvant également être changé. Les paramètres de ces différents modèles sont systématiquement estimés pour maximiser un critère statistique, sans intervention de l'utilisateur. De plus, ces critères prennent en compte les données étudiées de façon à obtenir une représentation à la fois correcte et parcimonieuse des données IRM. Enfin, le protocole repose sur une connaissance experte pré-existante limitée. En effet, il ne requiert pas l'emploi d'atlas ou d'une description des valeurs IRM au sein des pathologies mais un simple étiquetage de chaque examen (i.e. sujet sain, ou sujet porteur de telle pathologie).

Sélection automatique de modèles statistiques d'apprentissage.

La plupart des modèles statistiques utilisés dans le protocole mis au point sont des modèles de mélange de lois de probabilité, or ceux-ci nécessitent de fixer à l'avance le nombre de lois qui sont présentes au sein du mélange. Ce choix de paramètre peut être reformulé en terme de choix de modèles, et ainsi être traité par des techniques de sélection de modèle. Au sein du protocole développé, cette sélection de modèle est fait grâce à une heuristique de pente basée sur celle développée par Mauguis et Michel [8] et adaptée au cas du mélange de lois de Student à mélange d'échelles multiples (MMST). Cependant cette approche étant couteuse en temps de calcul, une extension bayésienne du modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP) a été mise au point afin de sélectionner

le nombre de lois du mélange au cours d'une procédure plus rapide.

Prise en compte de la dépendance spatiale des données IRM.
Le protocole d'analyse développé suppose que tous les voxels d'une acquisition IRM sont indépendants. Malgré cette supposition forte, les segmentations obtenues au cours des analyses sont cohérentes par rapport aux structures anatomiques. Afin de prendre en compte l'information spatiale disponible, une extension markovienne des modèles de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP) a été développée.

Outils numériques.
Un paquet R pour l'estimation des modèles de mélange de lois de Student à mélange d'échelles multiples a été développé sous licence publique générale GNU (v.3) en partenariat avec Stéphane Despréaux [2], à partir du code R développé par Daren Wraith [3].

Publications et communications.

Journal international.

— Alexis Arnaud, Florence Forbes, Nicolas Coquery, Nora Collomb, Benjamin Lemasson, et Emmanuel Barbier. *"Fully Automatic Lesion Localization and Characterization : Application to Brain Tumors Using Multiparametric Quantitative MRI Data."* IEEE Transactions on Medical Imaging, vol. 37, no. 7, pp. 1678-1689, July 2018.

Conférences internationales.

— Alexis Arnaud, Florence Forbes, Benjamin Lemasson, Emmanuel Luc Barbier. *"Tumor classification and prediction using robust multivariate clustering of multiparametric MRI."* International Society for Magnetic Resonance in Medicine (ISMRM), Toronto, Canada, May 2015.
— Benjamin Lemasson, Nora Collomb, Alexis Arnaud, Florence Forbes, Emmanuel Barbier. *"Monitoring brain tumor evolution using multiparametric MRI."* IEEE International Symposium on Biomedical Imaging (ISBI), Melbourne, Australia, April 2017.

2. Ingénieur de recherche, Laboratoire Jean-Kuntzmann, Grenoble, France
3. Senior lecturer, Queensland University of Technology, Brisbane, Australie

— Benjamin Lemasson, Nora Collomb, Alexis Arnaud, Emmanuel Luc Barbier, Florence Forbes. *"Monitoring glioma heterogeneity during tumor growth using clustering analysis of multiparametric MRI data."* International Society for Magnetic Resonance in Medicine (ISMRM), Honolulu, United States, April 2017.

Conférences nationales.

— Alexis Arnaud, Florence Forbes, Nicolas Coquery, Emmanuel Barbier, Benjamin Lemasson. *"Mélanges de lois de Student multivariées généralisées : application à la caractérisation de tumeurs par IRM multiparamétrique."* 2ème congrès de la Société Française de Résonance Magnétique en Biologie et Médecine (SFRMBM), Grenoble, France, March 2015.
— Alexis Arnaud, Florence Forbes, Benjamin Lemasson, Emmanuel Barbier, Nicolas Coquery. *"Mélanges de lois de Student à Échelles Multiples pour la caractérisation de tumeurs par IRM multiparamétrique."* 47èmes Journées de Statistique de la Société Française de Statistique (JdS), Lille, France, June 2015.
— Alexis Arnaud, Florence Forbes, Benjamin Lemasson, Emmanuel Barbier. *"Paquet R pour l'estimation d'un mélange de lois de Student multivariées à échelles multiples."* Quatrièmes Rencontres R, Grenoble, France, June 2015.
— Benjamin Lemasson, Nora Collomb, Alexis Arnaud, Florence Forbes, Emmanuel Barbier. *"Suivi de l'hétérogénéité de la croissance des gliomes par IRM multiparamétrique analysée par clustering."* 3ème congrès de la Société Française de Résonance Magnétique en Biologie et Médecine (SFRMBM), Bordeaux, France, March 2017.
— Felana Andriatsitoaina, Nora Collomb, Alexis Arnaud, Florence Forbes, Jean-Paul Issartel, Claire Loussouarn, Emmanuel Garcion, Emmanuel Barbier, et Benjamin Lemasson *"Suivi de l'hétérogénéité de la croissance de 4 modèles de gliomes par IRM multiparamétrique analysée par clustering."* Congrès national de l'imagerie du vivant (CNIV), Paris, France, November 2017.

1.3 Plan du manuscrit

Le chapitre 2 est consacré à une présentation des méthodes d'analyse actuelles dans le cadre des tumeurs cérébrales, suivant le type de données IRM considéré. En effet, les données IRM peuvent être quantitatives ou non, ce qui donne lieu à différentes stratégies d'analyse possibles.

Le chapitre 3 détaille la chaîne d'analyse mise au point afin de localiser et de caractériser automatiquement des zones anormales au sein d'un jeu de données d'intérêt. Il est illustré dans le cas de tumeurs cérébrales chez le rat, les données IRM ont été fournies par l'équipe Neuroimagerie Fonctionnelle et Perfusion Cérébrale, dirigée par Emmanuel Barbier, du Grenoble Institut des Neurosciences (GIN). Une première classification sur un jeu de données de référence (issu d'animaux sains) permet de réaliser la localisation de valeurs anormales présentes dans un second jeu de données (issu d'animaux porteurs de pathologies) ; une deuxième classification est alors réalisée sur les régions détectées comme anormales dans le but de caractériser les pathologies du second jeu de données.

Le chapitre 4 décrit une extension réalisée sur le modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP). Ce travail a été initié à l'Université McGill à Montréal (Canada), dans le cadre d'une bourse de recherche Mitacs Globalink-Inria. L'obtention de cette bourse a permis de réaliser un project de recherche de 5 mois au sein du laboratoire du professeur Russell Steele. Au cours de ce projet, une approche bayésienne a été ajoutée au modèle de mélange de lois à mélange d'échelles multiples afin de sélectionner le nombre de composantes du mélange.

Le chapitre 5 présente une seconde extension du modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP) pour la prise en compte de dépendances spatiales au sein des données IRM. Il s'agit cette fois d'exploiter la structure spatiale de l'acquisition IRM afin de renforcer les modèles statistiques utilisés. Cette dépendance spatiale est modélisée par un champ de Markov latent sur les variables de classe au sein du mélange de lois de Pearson type VII à mélange d'échelles multiples.

Le chapitre 6 résume les différents développements informatiques (en R et C++) réalisés au cours de la thèse pour l'estimation des différents modèles statistiques mis en jeu et leurs applications aux données IRM.

CHAPITRE 2

INTRODUCTION À L'ANALYSE DE DONNÉES IRM

Ce chapitre résume les principales études portant sur la localisation et la caractérisation des tumeurs cérébrale menées ces dernières années et permet d'évaluer les principales limites des méthodes actuelles.

2.1 Images quantitatives et non-quantitatives

Les images IRM acquises en routine clinique mesurent l'intensité d'un signal. Celui-ci dépend de nombreux paramètres physiques, biologiques et instrumentaux. Ainsi, l'intensité d'un signal obtenu chez une même personne et au même point d'un organe peut varier au-delà du bruit d'acquisition, indépendamment de l'état biologique de la personne. Cette variation du signal est essentiellement d'origine physique et liée soit à des biais d'acquisition soit à de petites fluctuations des paramètres d'acquisition.

Un premier biais important est lié à la distribution du champ magnétique (B0) qui n'est pas identique entre tous les voxels : la valeur moyenne et la dispersion de B0 varient entre les voxels. Si la variation de valeur moyenne de B0 est en général trop faible pour conduire à des variations de signal détectable, la dispersion du champ magnétique dans un voxel influe sur la vitesse de décroissance du signal. Comme les signaux de tous les voxels sont mesurés en même temps, on perçoit une variation d'intensité entre des voxels contenant des tissus identiques mais dont les vitesses de décroissance diffèrent. Un second biais important provient du fait que les distributions spatiales des deux champs électromagnétiques (B1), celui qui excite les spins des protons et celui qui est émis pendant la relaxation de l'aimantation, présentent également des variations spatiales d'intensité. Ainsi, l'antenne qui capte le signal ne produit pas un signal spatialement homogène. Une personne

installée deux jours de suite dans l'IRM ne placera pas sa tête dans la même position et donc le signal IRM variera. Cette sensibilité aux biais B0 et B1 dépend du type de séquence IRM utilisée. Une autre source de variabilité importante est liée aux variations des valeurs de paramètres d'acquisition. Ces derniers sont très nombreux en IRM (plus d'une centaine), et un même contraste au sens radiologique (par exemple l'imagerie pondérée T2) peut être obtenu avec des jeux de paramètres légèrement différents. On constate que les protocoles sauvegardés sur les équipements IRM pour un même contraste sont aussi très nombreux. La modification de certains paramètres qui n'ont a priori pas de lien avec le contraste (par exemple le nombre de coupes) peuvent entraîner une légère variation de la valeur de certains paramètres d'acquisition. Ces légères modifications ne sont pas aisément décelables par l'opérateur et contribuent à la variabilité du signal. De même, les équipements IRM diffèrent d'un site à l'autre (différents constructeurs, différentes versions du logiciel d'acquisition etc.), ce qui contribue aussi à la variabilité des signaux IRM. Toutes ces sources de variabilité posent des problèmes importants dans la standardisation des essais cliniques multicentriques et plus généralement dans la standardisation de la pratique radiologique.

À ces variations d'origine physique et instrumentale s'ajoutent des biais physiologiques potentiels. Par exemple, la fréquence des battements cardiaques, le débit sanguin ou la température du tissu peut influer sur le signal. En pratique, les biais physiologiques peuvent être négligés pour les images anatomiques.

Ce mode d'acquisition, qui produit des images non quantitatives, est le mode d'acquisition standard des IRM en clinique. En effet, même si les valeurs absolues des signaux varient, les contrastes, c'est-à-dire les différences relatives de signal entre deux tissus de nature différentes, peuvent être rendus suffisamment stables pour permettre une interprétation des images par un radiologue.

La variabilité rencontrée sur ces images non quantitatives pose cependant des problèmes pour l'analyse automatique de ces données. En effet, comment distinguer les variations d'origine physique ou instrumentale des variations d'origine biologique ? Une approche pour réduire ces variations physiques et instrumentales consiste à obtenir des images quantitatives. Par exemple, en faisant des acquisitions d'une image non quantitative pour plusieurs valeurs d'un paramètre d'acquisition, il est possible d'estimer un modèle biophysique afin d'obtenir des cartes dites paramétriques ou quantitatives. Par exemple, pour les images en écho de spin, si on répète l'acquisition en faisant varier la valeur du temps d'écho, il est possible de calculer une carte paramétrique T2 à partir de l'ensemble des images à l'aide d'un modèle biophysique de décroissance exponentielle. Pour obtenir des cartes paramétriques T1, on peut ajouter une impulsion radiofréquence d'inversion de l'aimantation et faire varier le temps d'inversion, c'est-à-dire le délai entre l'impulsion d'inversion et l'acquisition de l'image (Figure 2.1).

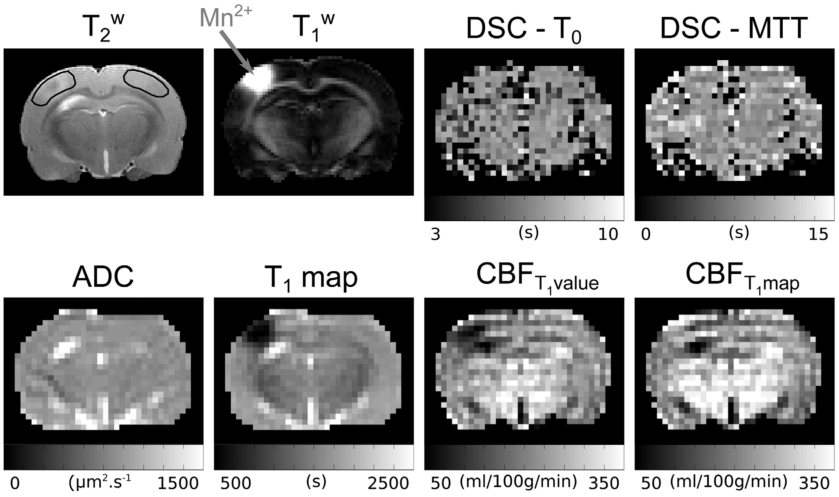


FIGURE 2.1 – Exemple d'images non quantitatives et de cartes paramétriques quantitatives, obtenues par IRM chez le rat, après injection intra-corticale d'un agent de contraste (Mn2+, Flèche). Les images pondérées T2 (T2w) et T1 (T1w) sont non quantitatives. Les autres cartes sont quantitatives. La carte paramétrique T1 (T1 map) donne la valeur du temps de relaxation en ms. La réduction de T1, visible dans le cortex comme une zone noire, est à l'origine du contraste visible sur la l'image pondérée T1 (T1w). Figure extraite de Debacker et al 2017.

Dans une carte paramétrique, une valeur quantitative, c'est-à-dire avec une unité physique, est ainsi associée à chaque voxel et possède une signification physique ou chimique (Pierpaoli [9]). En théorie, tous les voxels deviennent ainsi comparables entre eux, c'est-à-dire que les impacts des biais B0 et B1 ainsi que l'impact des variations des paramètres d'acquisition ont été éliminés. Il devient ainsi possible de comparer directement les voxels de différents patients, et même de différents centres d'acquisition. En pratique, les biais B0 et B1 sont fortement réduits mais pas toujours totalement éliminés. Par exemple, dans le cas d'une carte paramétrique T1 obtenue avec une approche multi-angle, la carte de T1 sera sensible au biais B1. Dans notre exemple de carte T2, l'approche n'est pas sensible au biais B0, et le biais B1 peut être totalement éliminé. Un autre intérêt des cartes quantitatives provient du fait que les valeurs mesurées sont associées à un seul paramètre alors que les images pondérées sont la résultante de l'influence plusieurs paramètres, différemment exprimés. La stabilité et la variété des informations accessibles des cartes IRM quantitatives apparaissent comme des atouts attractifs pour développer des biomarqueurs en IRM (Wu, Dijkhuizen and Sorensen [10], Waldman et al [11]).

2.2 Étude des données non quantitatives

Les images non quantitatives ont fait l'objet de nombreuses études et leurs utilisation en clinique fait consensus. Ce sont donc celles qui sont actuellement le plus utilisées pour développer des méthodes d'analyse.

2.2.1 Localisation de tumeurs

Les recommandations actuelles pour quantifier la présence de tumeurs sont principalement basées sur des mesures anatomiques unidimensionnelles (RECIST, Eisenhauer et al [6]), de même pour la réponse des tumeurs aux traitements qui reposent sur l'évolution de mesures bidimensionnelles au cours du temps (RANO, Wen et al [3]). Même si elles sont disponibles, des mesures plus complexes ne sont pas actuellement recommandées du fait du manque de standardisation de ces mesures (les travaux de standardisation sont en cours), et d'une validation clinique rigoureuse. De plus, les méthodes développées au cours de ces dernières années étaient souvent évaluées sur des jeux de données avec de petits effectifs, et des types d'images IRM et des types de métriques variables selon les études. La mise en place du challenge "Multimodal Brain Tumor Segmentation" (BraTS) a permis de répondre en partie au besoin de données communes pour comparer et quantifier les performances des nouvelles méthodes de segmentation. Depuis 2012, le challenge BraTS a ainsi mis à disposition un jeu de données publiques d'IRM de patients porteurs de tumeurs cérébrales, ainsi qu'une méthode d'évaluation automatique des algorithmes soumis. Nous allons détailler les derniers résultats connus du challenge BraTs afin d'avoir une bonne vision des méthodes de segmentation qui existent.

L'objectif du challenge BraTs est de délimiter 4 structures pour chaque tumeur cérébrale (haut-grade ou bas-grade) : œdème (en jaune sur la Figure 2.2), région tumorale ne prenant pas le contraste (en rouge), centre nécrosé (en vert), centre actif (en bleu). Les données disponibles sont issues de l'acquisition d'images IRM non quantitatives : image pondérée T1, image pondérée T1 après injection de gadolinium (Gd), image pondérée T2, et image FLAIR (contraste T2 mais dont la contribution du liquide céphalorachidien a été supprimée). Une délinéation manuelle des 4 structures de la tumeur, réalisée par 1 à 4 experts, est également fournie pour servir de référence lors de l'évaluation des segmentations automatiques (voir Figure 2.2). Cette évaluation se fait en comparant la concordance des segmentations en terme de surface de recouvrement (coefficient de DICE), et en terme de distance (distance de Hausdorff). L'annexe A détaille les expressions des différentes métriques utilisées dans ce manuscrit : le coefficient DICE, l'indice ARI, et la distance de Hausdorff. Les méthodes soumises à ce challenge sont comparées entre elles via un test non paramétrique de signification (test de Wilcoxon au

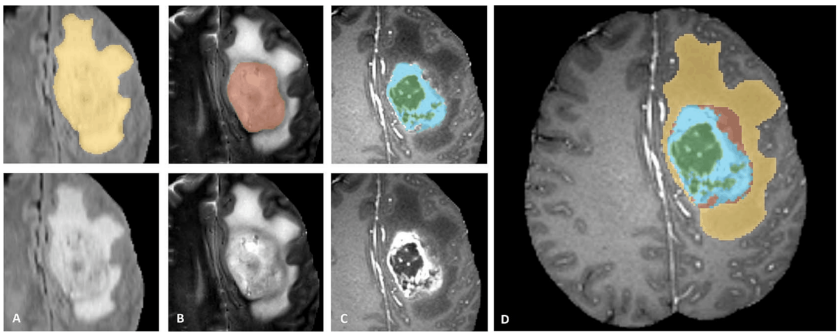


FIGURE 2.2 – Contourage manuel des experts. En haut à gauche, (A) image FLAIR avec le contour de la tumeur entière et de l'œdème, (B) image pondérée T2 avec le contour du cœur de la tumeur, (C) image pondérée T1 après injection d'agent de contraste avec les contours de la tumeur prenant le contraste (en bleu) et des régions nécrotiques/cystiques (en vert). En bas à gauche, mêmes images mais sans le contour. (D) Image pondérée T1 après agent de contraste avec les 4 régions superposées : en jaune, l'œdème, en rouge, la tumeur qui ne prend pas le contraste, en vert les régions nécrotiques/cystiques, et en bleu la tumeur qui prend le contraste. Figure extraite de Menze et al 2015.

niveau 5 %), et par rapport à la variabilité des segmentations manuelles.

Au cours des éditions 2012 et 2013 ([12]), les méthodes de segmentation proposées étaient issues de deux grandes familles : les méthodes probabilistes génératives, et celles discriminatives. Dans le premier cas, il s'agit de modèles explicitant le lien entre les valeurs mesurées et l'étiquette de segmentation, que ce soit d'un point de vue anatomique ou d'un point de vue des contrastes IRM. Ces modèles permettent ainsi une intégration facile de connaissances a priori du domaine d'application. On retrouve notamment l'utilisation de modèles de croissance de lésion, de détection d'anomalie, de modèle de mélange ou encore de champs de Markov aléatoires. Cependant, ces méthodes nécessitent de définir ce qu'est une lésion, afin d'en expliciter le lien avec les valeurs mesurées. Une limitation majeure réside dans l'interprétation des segmentations obtenues via la sémantique du radiologue : quel est le sens physiologique de l'appartenance d'un voxel à une certaine classe ? Les méthodes discriminatives, à l'inverse, estiment directement le lien entre les étiquettes de segmentation et les valeurs mesurées ; cet apprentissage se base sur l'extraction de propriétés (ou caractéristiques) locales des images, puis la recherche des propriétés pertinentes pour la tâche de segmentation. On y classe dedans les forêts aléatoires, les arbres de décisions, les séparateurs à vaste marge, les réseaux de neurones à convolution, ou encore les champs aléatoires conditionnels. Cependant, ces méthodes nécessitent un volume important de données pour permettre une bonne généralisation de la segmentation apprise. De plus, une limitation majeure

persiste dans la dépendance de la discrimination par rapport aux intensités mesurées, dans le cas de données IRM non quantitatives ; les données doivent ainsi avoir été acquises suivant le même protocole (même équipement IRM, même séquences, etc.), l'étape de calibration étant particulièrement cruciale. Une approche hybride est possible, dans laquelle une méthode générative est utilisée afin de fournir des données stables à une seconde étape réalisant la discrimination.

Les méthodes ayant donné les meilleurs résultats aux éditions 2012 et 2013 sont principalement les méthodes discriminantes ayant une régularisation spatiale, y compris celles hybrides incorporant un modèle génératif préalable. En particulier, les méthodes de forêts aléatoires ont été très populaires pour l'édition de 2013. Si aucune méthode n'est apparue uniformément meilleure que les autres (sur les 4 structures à détecter), la fusion de ces algorithmes a permis d'obtenir par la suite une segmentation plus robuste et plus performante que chacune des méthodes individuelles. Inversement, les méthodes ayant eu la moins bonne performance sont celles ayant peu ou pas de régularisation spatiale sur les cartes IRM considérées. À partir de 2016, quasiment toutes les méthodes étaient à base de réseaux de neurones à convolution.

Une difficulté classique de la segmentation automatique est la distinction entre classes. En effet, les méthodes non supervisées permettent un partitionnement des données en plusieurs classes, mais sans que les classes n'aient de sens intrinsèque : une classe est simplement définie par la loi de probabilité associée, celle-ci devant être estimée. Il est ainsi nécessaire d'identifier ce que représente chaque classe après l'ajustement du modèle (par exemple œdème, nécrose etc.). Cette identification nécessite généralement l'intervention d'un expert.

À l'étude BraTs, nous pouvons ajouter les travaux de Juan-Albarracín et al. [13] qui s'appuient sur les données du challenge BraTs de 2013. Les auteurs résolvent le problème de l'identification des classes dans le cadre d'une étude chez l'homme en choisissant de détecter les tissus non pathologiques, isolant ainsi les tissus tumoraux (Figure 2.3). Cependant, cette détection se base sur un atlas de référence dont la création présente plusieurs difficultés telles que le recalage de tissus pathologiques. En outre, certains paramètres doivent être fixés par l'utilisateur, comme par exemple le nombre de tissus ou le nombre de classes à obtenir dans la tumeur. Enfin, les paramètres IRM utilisés sont non quantitatifs, ce qui limite la généralisation de la méthode à des données issus d'un même centre d'acquisition IRM. Les résultats sont toutefois intéressants puisque l'approche proposée produit de meilleurs résultats (en terme de DICE par exemple) que ceux publiés dans le contexte du challenge BraTs 2013.

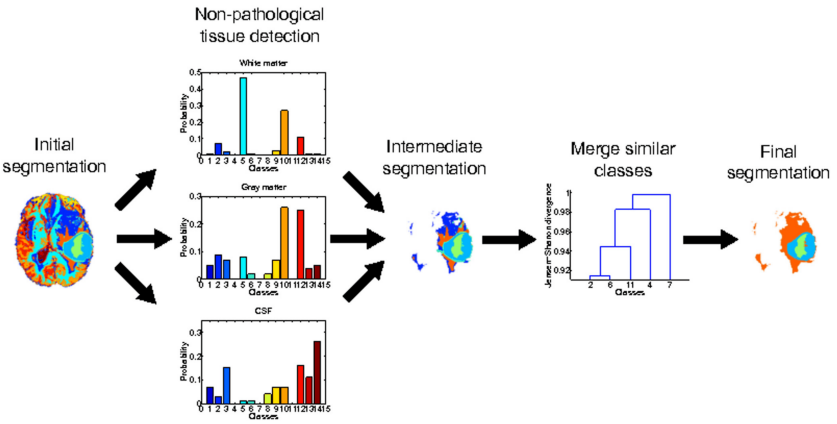


FIGURE 2.3 – Principales étapes de l'approche proposée par Juan-Albarracín et al. Après une segmentation (une approche de type Kmean avec de nombreuses initialisations), les données sont recalées sur un atlas probabilistique du cerveau humain sain (ICBM atlas). Pour chaque voxel, on dispose alors d'une probabilité que le tissu soit de la substance grise (Grey matter), de la substance blanche (White matter) ou du liquide céphalorachidien (CSF). À partir des histogrammes des images FLAIR et T1c, on peut également filtrer les régions présentant de fortes intensités et obtenir un masque très grossier de la lésion. Les classes contenant du tissu sain peuvent ainsi être identifiées et retirées. Les classes restantes sont enfin fusionnées au moyen d'une classification ascendante hiérarchique afin d'obtenir 4 classes. Ce nombre de classes correspondant au nombre de régions contourées par les experts impliqués dans ce challenge BraTs. Figure extraite de Juan-Albarracín et al 2015.

2.2.2 Caractérisation de tumeurs

Si l'on peut détecter des zones anormales au sein de données saines grâce aux données IRM non quantitatives, la question suivante est naturellement de savoir si l'on peut également utiliser ces données de façon à caractériser cette anomalie. Ainsi, de nombreuses études ont vu le jour pour appliquer des méthodes de traitement d'image afin d'extraire une information autre que spatiale, comme par exemple prédire la survie de patients ou le type de tumeurs à partir de l'imagerie IRM.

Macyszyn et al. [14] cherchent à déterminer s'il existe des paramètres IRM qui permettent de prédire la survie des patients ainsi que le type moléculaire des tumeurs. Les données IRM obtenues chez 134 patients avec un champ magnétique à 3T sont classiques (images pondérée T1 avant et après injection de produit de contraste, pondérée T2, diffusion, perfusion). Les auteurs utilisent une approche

de segmentation automatique précédemment publiée qui contoure trois zones dans la tumeur (tumeur qui prend le contraste, tumeur qui ne prend pas le contraste, œdème) et les ventricules. La segmentation automatique dans l'espace des paramètres IRM est réalisée avec un nombre de classes fixe dont l'interprétation est connue par avance : l'objectif est ici d'obtenir des caractéristiques globales sur les images plutôt qu'une analyse fine de l'hétérogénéité intra-tumorale. Pour l'aspect caractérisation de la tumeur, les auteurs utilisent un algorithme de séparateurs à vaste marge (Support Vector Machines : SVM) avec 2 classes (nombre fixés par les auteurs). Ils peuvent ainsi prédire trois type de survie (courte, moyenne ou longue) et 4 phénotypes moléculaires des tumeurs considérées (neural, proneural, classique, et mésenchymateux).

Zacharaki et al. [15] proposent une classification de différentes tumeurs cérébrales via l'extraction de caractéristiques de type géométrique et de textures à partir des données IRM obtenues chez 98 patients, précédé par une étape de contourage manuel de ROIs. La Figure 2.4 illustre les filtres de Gabor utilisés dans cette étude. Les auteurs ont utilisé des images pondérées T1 avant et après injection de contraste, des images pondérées T2, des images FLAIR et des images de volume sanguin relatif. Les approches SVM de classification de textures apparaissent ici prometteuses. La précision de la classification est de 85%, la sensibilité de 87% et la spécificité de 79%. Les auteurs indiquent néanmoins que la qualité de la classification devrait pouvoir être améliorée par l'ajout d'autres informations comme le spectre RMN de la tumeur ou une carte de coefficient de diffusion.

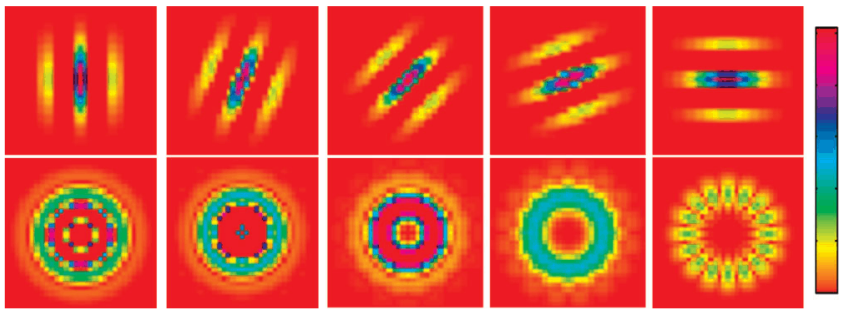


FIGURE 2.4 – Exemples de filtres de Gabor utilisés pour extraire des images des caractéristiques fréquentielles localisées (ou textures). La première ligne montre des filtres de Gabor de même fréquence et d'orientations différentes, la seconde ligne les filtres invariants par rotation. Figure extraite de Zacharaki et al 2009.

Les approches par analyse de textures font partie des approches dites "radiomiques". On distingue les éléments de forme (par exemple la taille), les éléments de premier ordre (moyenne, kurtosis etc.) et les éléments de second ordre (relation

des voxels entre eux ou texture). On peut extraire de nombreuses caractéristiques des images et il peut être nécessaire d'opérer des étapes de réduction de dimensions. Ensuite, on cherche à relier les caractéristiques des images à un indicateur médical ou biologique (par exemple la survie ou le sous-type moléculaire). Même si certaines caractéristiques géométriques ont un sens, les textures ne représentent en général pas une information pathophysiologique lisible par un médecin ou un biologiste. Enfin, les analyses d'images non quantitatives sont par nature dépendantes des biais d'acquisition, comme évoqué plus haut, ce qui rend l'analyse de texture peu robuste, notamment lors d'études multicentriques.

2.3 Étude des données quantitatives

Les études qui portent sur des données quantitatives en IRM sont peu nombreuses et se sont surtout intéressées à l'aspect caractérisation de la tumeur. Ces études se basent souvent sur une délinéation manuelle préalable. C'est le cas des trois étude que nous décrivons ci-après [16], [17] et [18].

2.3.1 Localisation de tumeurs

Concernant la localisation de tumeurs, il est à noter l'étude de Rasmussen et al. [19] qui s'intéresse à la question du choix des paramètres d'imagerie (fluorodéoxyglucose imagé par tomographie par émission de positon, carte de coefficient de diffusion obtenue par IRM) pour discriminer au mieux des volumes et sous-volumes tumoraux (Figure 2.5). Les auteurs rapportent que les volumes tumoraux vus par PET et par IRM correspondent assez bien mais pas parfaitement : le volume tumoral vu en IRM de diffusion est plus petit que celui vu en PET et n'est pas totalement inclus dans la lésion vue en PET. Le volume vu en PET et celui en IRM de diffusion sont par contre totalement inclus dans le volume de lésion vu en IRM pondérée T2. Cette étude pose donc la question de la fiabilité des méthodes d'imagerie pour déterminer la zone qui sera irradiée par radiothérapie.

2.3.2 Caractérisation de tumeurs

De la même façon que pour les données IRM non quantitatives, des études ont été menées afin d'exploiter l'IRM quantitative à des fin de caractérisation des types de tumeurs.

À la frontière de la segmentation et de la caractérisation tumorale, Katiyar et al. [18] ont développé une technique de segmentation intra-tumorale pour caractériser les zones nécrotiques, péri-nécrotiques et viables des tumeurs U87 implantées chez des souris. Ils ont utilisé des données IRM quantitatives de coefficient de

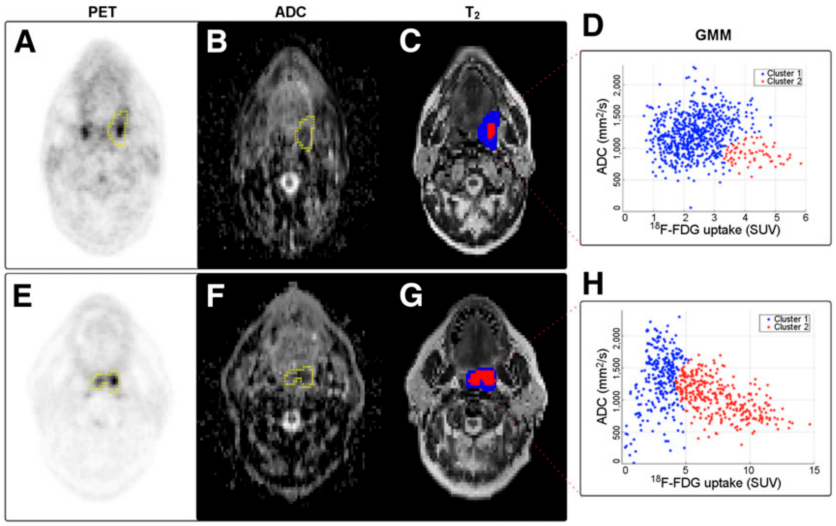


FIGURE 2.5 – (A) Image TEP (18F-FDG), (B) Carte de coefficient de diffusion et (C) image pondérée T2 d'un patient pour lequel l'IRM et la TEP ne correspondent pas bien. En jaune, le volume tumoral contouré sur l'image pondérée T2. (D) Nuage de points des coefficients de diffusion par rapport aux valeurs TEP mesurées, avec en couleur l'appartenance aux classes détectées. (E-H) Comme (A-D) mais pour un patient chez lequel le TEP et l'IRM produisent des contours tumoraux comparables. Figure extraite de Rasmussen et al 2017.

diffusion, de T2 avant et après injection de nanoparticules d'oxyde de fer, et de T2* avant et après l'injection de produit de contraste. À l'aide d'une approche de clustering spectral régularisé spatialement (SRSC), ils obtiennent des résultats très comparables à l'histologie (Figure 2.6). Cette approche repose par contre sur l'hypothèse que la tumeur contient toujours trois classes, ce qui limite sa généralisation.

La distribution de probabilité gaussienne multivariée est souvent utilisée pour modéliser des observations multidimensionnelles du fait de leur facilité d'utilisation, indépendamment de la dimension considérée. C'est le cas de Coquery et al. [16] qui utilisent 6 cartes IRM (Figure 2.7) afin d'identifier différents tissus en se basant sur l'analyse de groupes de voxels ayant des paramètres IRM quantitatifs similaires. Les données ont été acquises chez le rat et deux modèles de gliomes ont été testés (C6 chez le rat Wistar et F98 chez le rat Fischer, 26 rats en tout).

Dans un premier temps, trois ROI sont manuellement délimitées (tumeur, cortex sain, striatum sain), puis un modèle de mélange de lois gaussiennes en dimension 6 est ajusté sur l'ensemble des voxels de ces ROI. Le nombre de composantes, aussi appelées classes, du mélange est déterminé via le critère d'information bayé-

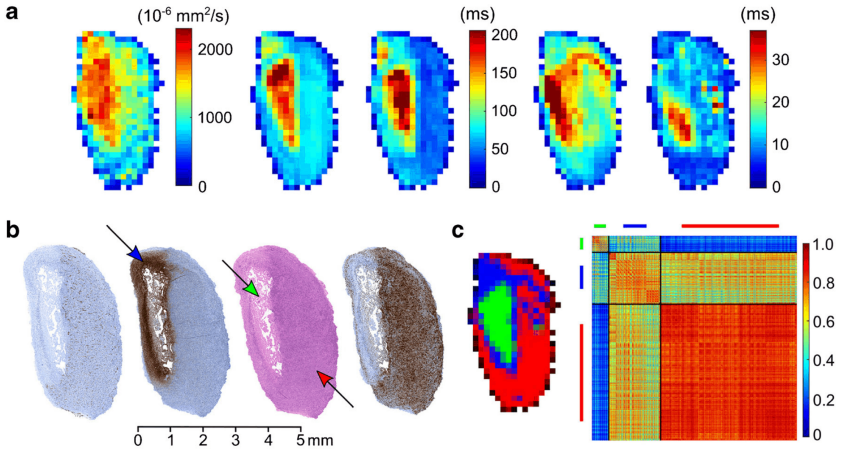


FIGURE 2.6 – a. Paramètres IRM acquis sur la tumeur : ADC (coefficient de diffusion), cartes T2 avant et après injection d'agent de contraste et cartes T2* avant et après injection d'agent de contraste. b. Coupes histologiques avec marquage CD31 (marquage de cellules endothéliales et donc des vaisseaux), GLUT-1 (transporteur du glucose, associé ici à l'hypoxie et donc à la zone périnécrotique), H&E (marqueur des noyaux cellulaires) et KI67 (marqueurs des cellules en prolifération). c. Carte de probabilité issue de l'analyse des cartes IRM où les couleurs verte, bleue et rouge correspondent aux régions nécrotiques, péri-nécrotiques et viable de la tumeur. La matrice d'affinité est obtenue à partir des vecteurs de caractéristiques obtenues pour chaque voxel. Ainsi les voxels étiquetés en vert sont peu similaire à ceux étiquetés en rouge. Les voxels étiquetés en rouge sont tous très similaires en eux, alors que c'est moins marqué pour les deux autres classes. Figure extraite de Katiyar et al 2016.

sien ("Bayesian Information Criterion" : BIC) (Figure 2.8). Cependant ce dernier ne présente pas de maximum net, et le nombre de composantes a donc été choisi au point d'inflexion de la courbe du BIC. Une étude histologique simple (inspection visuelle) est utilisée afin d'évaluer la pertinence de la classification obtenue. La Figure 2.9 montre une représentation en deux dimensions d'une décomposition en composantes principales des données. Même si les composantes principales n'ont pas été utilisées dans l'analyse (les valeurs des voxels ont été utilisées directement), cette représentation illustre la complexité de séparer des clusters en 6 dimensions.

La Figure 2.10 montre l'appartenance des voxels des ROIs aux classes pour tous les animaux de l'étude. Même si l'ensemble paraît globalement homogène, on perçoit que tous les animaux ne portent pas exactement la même tumeur et que le modèle F98 est plus variable que le modèle C6. Une approche de type leave-one-out a permis d'estimer que cette approche permet de prédire correctement à 84% le type de la tumeur d'un rat.

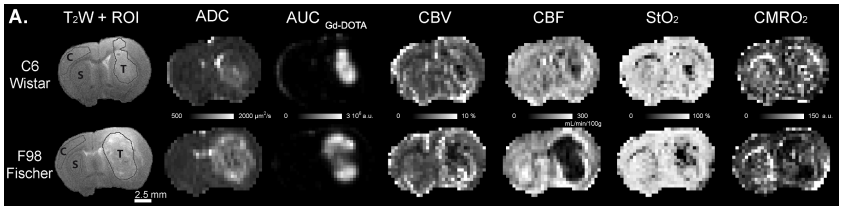


FIGURE 2.7 – IRM obtenues chez des rats Wistar porteurs de gliomes C6 (ligne du haut) et chez des rats Fischer porteurs de gliomes F98 (ligne du bas). Sur l'image pondérée T2, non quantitative, sont représentées les 3 ROI utilisées dans l'étude. Les 6 autres cartes sont quantitatives : coefficient de diffusion (ADC), perméabilité de paroi vasculaire (AUC-Gd-DOTA), volume sanguin (CBV), débit sanguin (CBF), saturation tissulaire en oxygène (StO2) et consommation d'oxygène (CMRO2). Figure extraite de Coquery et al 2014.

Boult et al. [17] se basent également sur une segmentation manuelle avant d'appliquer le modèle de k-means (cas particulier du modèle de mélange gaussien) sur 5 cartes IRM quantitatives afin d'identifier aussi différents types de tissus tumoraux (Figure 2.11). La validation repose également sur une étude histologique.

Ces différentes études présentent des résultats prometteurs mais reposent sur une délinéation manuelle de la tumeur, ou sur un réglage du nombre de classes imposé par l'utilisateur, voire les deux. De plus, le nombre de types de tumeur étudié est faible (1 ou 2). Ces différents éléments limitent fortement la portée des méthodes proposées.

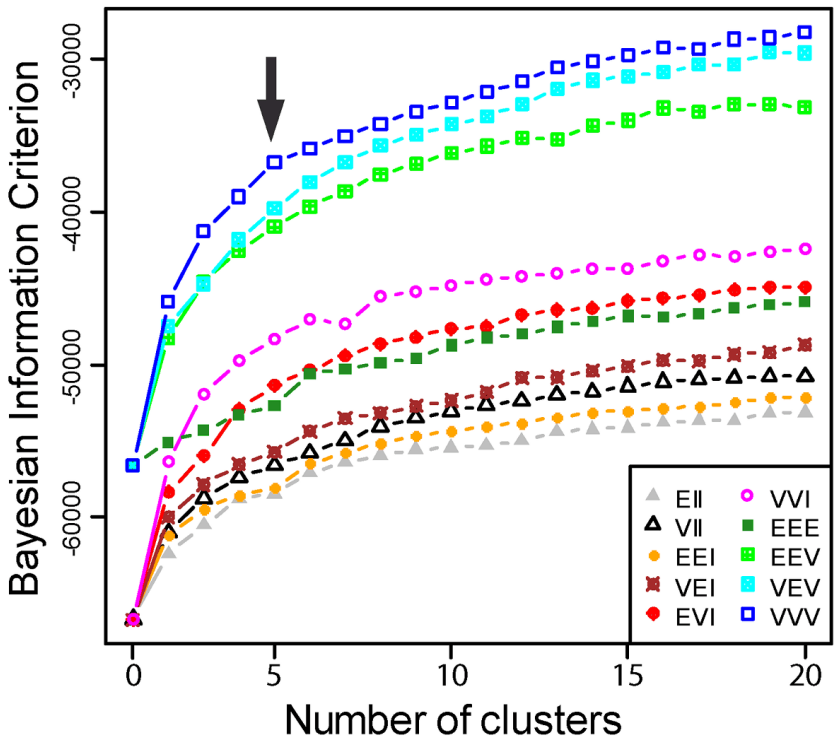


FIGURE 2.8 – Valeur du BIC pour les différents modèles de mélanges gaussiens évalué dans l'étude. La flèche noire correspond à l'inflexion de pente détectée par l'utilisateur et donc au nombre de classes retenues pour l'étude. Figure extraite de Coquery et al 2014, Supplément.

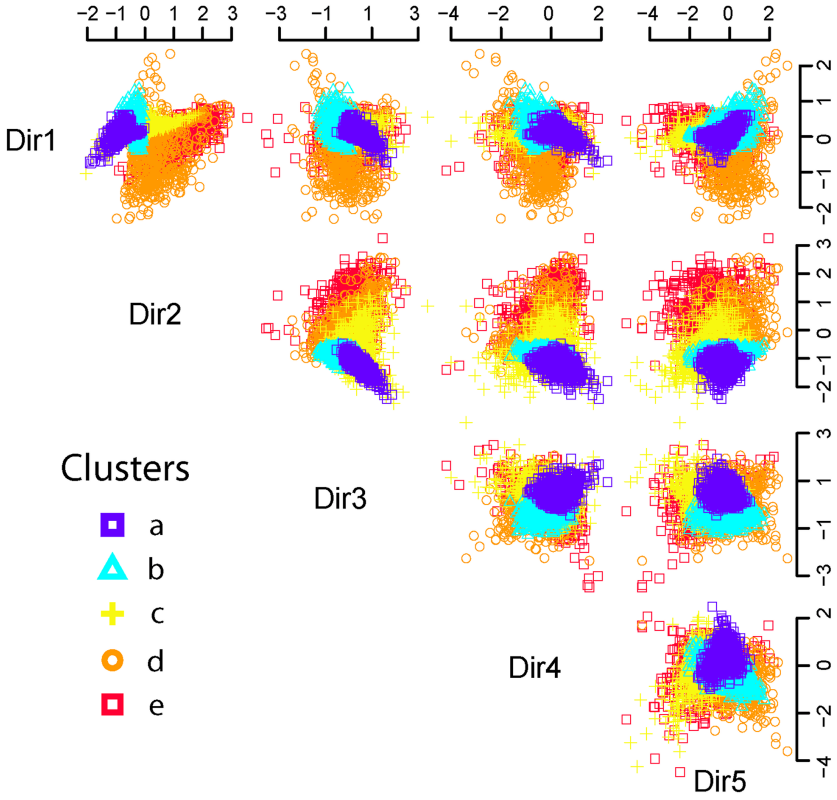


FIGURE 2.9 – Décomposition en composante principale des caractéristiques des voxels et représentation spatiale sous forme de nuages de points 2D des liens entre les cinq principales composantes. Figure extraite de Coquery et al 2014, Supplément.

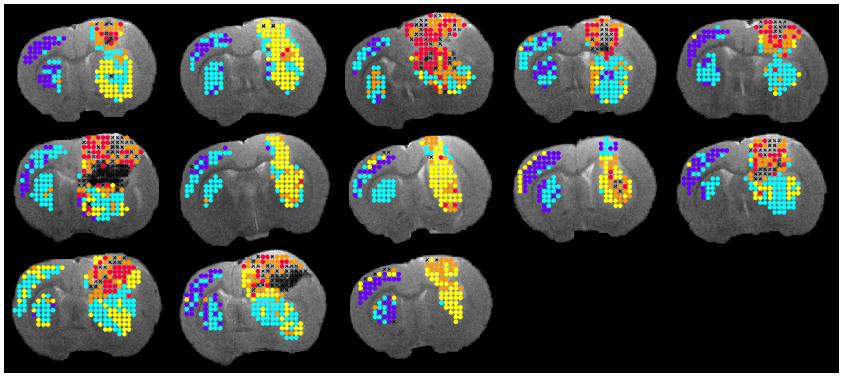


FIGURE 2.10 – Étiquetage des voxels dans les ROI : tissus sains à gauche (cortex et striatum) et tumeur à droite. Chaque couleur désigne l'appartenance à un cluster. Pour chaque rat de l'étude, la coupe centrale est représentée. Figure extraite de Coquery et al 2014.

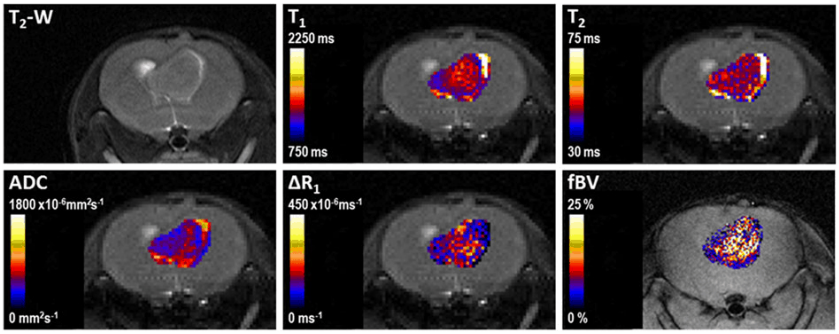


FIGURE 2.11 – Image pondérée T2 et superposition des valeurs quantitatives des 5 paramètres étudiées : T1, T2, ADC, variation de la vitesse de relaxation après agent de contraste et volume sanguin (fBV), pour une tumeur RG2 (une ligné tumorale de rat). Figure extraite de Boult et al 2016.

CHAPITRE 3

PROTOCOLE D'ANALYSE STATISTIQUE D'IRM MULTIPARAMÉTRIQUE BASÉ SUR DES MODÈLES DE MÉLANGE

Dans ce chapitre, nous présentons nos travaux sur l'utilisation des modèles de mélange de lois de Student à mélange d'échelles multiples (MMST) afin de réaliser une chaîne d'analyse statistique pour la localisation et la caractérisation automatiques d'anomalies au sein de données IRM quantitatives. Les sections 3.1, 3.2 et 3.3 résument les enjeux de telles données ainsi que le protocole d'analyse élaboré au cours de la thèse. Les détails sont présentés dans l'article publié dans le journal IEEE Transactions on Medical Imaging section 3.4.

3.1 Enjeux de l'analyse des tumeurs cérébrales

L'analyse de tumeurs cérébrales se décompose principalement en deux tâches liées : la localisation et la caractérisation de la pathologie. Celles-ci sont cruciales dans l'élaboration d'un diagnostic, ainsi que coûteuses en temps d'apprentissage par le radiologue. En effet, pour chaque ensemble de cartes IRM à étudier, ce dernier doit mémoriser comment les tumeurs considérées prennent le contraste sur les cartes IRM de façon à pouvoir en définir les contours et le type. Ce travail doit donc être refait lorsque le type de cartes IRM utilisées ou leur nombre changent. Comme nous l'avons vu au chapitre précédent, un certain nombre de solutions automatiques ont ainsi été développées au cours des dernières années, que ce soit via l'utilisation de méthodes de segmentation ou l'analyse des paramètres physiologiques. Cependant, celles-ci se concentrent sur l'automatisation d'une de ces tâches, la seconde étant réalisée manuellement ou sous le contrôle de l'utilisateur

[13]-[18].

Si les articles mentionnés précédemment montrent bien l'activité autour de l'analyse des données IRM de cancérologie, ils pointent également la difficulté à fournir une méthode entièrement automatisée et reposent en partie sur la connaissance extérieure apportée par un expert. La délimitation manuelle de la tumeur en est un exemple, mais les méthodes statistiques utilisées nécessitent également des choix tels que le nombre de classes pour un mélange, le type de lois de probabilité, ou encore la nature de la dépendance entre les paramètres IRM.

Une limitation dans les modélisations des études précédemment citées est l'absence de prise en compte de la dépendance, au sens de corrélation statistique, entre les paramètres IRM, ou l'utilisation du modèle de mélange gaussien. Coquery et al. [16] observent par exemple des distributions non gaussiennes pour certains groupes de voxels malgré l'utilisation de modèles gaussiens (Figure 2.10). C'est pourquoi il y a un vrai besoin de développer des outils d'analyse entièrement automatiques pour réaliser une étude des données IRM qui soit à la fois reproductible et en accord avec celle des experts.

3.2 Présentation du protocole d'analyse développé

L'approche développée dans cette thèse diffère de celles présentées précédemment par un objectif d'automatisation le plus complet possible et d'intégration de toute la chaîne de prise de décision, depuis les cartes IRM jusqu'à la prédiction du type de tumeur. Il s'agit donc d'unifier les deux tâches de localisation et de caractérisation des tumeurs sous la forme d'une procédure générique, automatique et adaptative aux données.

En partant de cartes IRM quantitatives, il s'agit de réaliser une délimitation automatique (approche non supervisée) de la pathologie vue comme anomalie par rapport à un modèle de référence appris, puis de caractériser cette anomalie, via l'extraction de marqueurs dans l'espace des paramètres IRM, de façon à réaliser un dictionnaire de signatures de tumeurs (approche supervisée). Les apprentissages non supervisés sont réalisés en utilisant le modèle statistique de mélange de lois de probabilité, tandis que les apprentissages supervisés sont basés sur des méthodes statistiques d'analyses discriminantes. La procédure est voulue avec le minimum d'information a priori, telles que des connaissances d'experts ou des données spatiales, cela afin d'évaluer uniquement la capacité de segmentation et de discrimination contenue dans les seuls paramètres IRM. C'est également pour cela que les tissus pathologiques ne sont pas modélisés directement mais détectés

comme des voxels anormaux par rapport à un modèle statistique décrivant les paramètres IRM de tissus sains.

Ce protocole nécessite ainsi l'utilisation de deux jeux de données d'apprentissage : un premier composé des voxels associés à des sujets sains, ou sujets de référence, et un second composé des voxels provenant de sujets pathologiques (par exemple atteints d'une tumeur) dont seul le type de la pathologie est connue. Le protocole se compose de 5 étapes successives :

1. un premier modèle de mélange sert à caractériser l'ensemble des voxels des sujets de référence afin de construire un modèle dit de référence qui décrit les tissus de référence ;
2. le modèle de référence est ensuite utilisé pour détecter les voxels anormaux, ou atypiques, à la fois parmi les voxels des sujets pathologiques et parmi ceux des sujets de référence (délimitation de la pathologie vue comme anomalie) ; cette détection se fait en sélectionnant les voxels dont le niveau d'anormalité est supérieur à un seuil qui est déterminé de manière automatique ;
3. un second modèle de mélange, dit modèle d'anomalie, est ajusté sur l'ensemble des voxels déclarés comme atypiques de façon à caractériser les tissus considérés comme anormaux ;
4. cet ajustement induit une classification des voxels en classe dont les proportions par sujet sert de signature individuelle afin de construire un modèle de signatures anormales (caractérisation des pathologies) ;
5. une étape supplémentaire de post-traitement spatial est possible, seul étape incorporant la nature spatiale des acquisitions IRM, de façon à éliminer des artefacts spatiaux et raffiner la construction des signatures anormales.

Ces 5 étapes correspondent à des parties élémentaires qui peuvent être modifiées ou remplacées par différents algorithmes issues de la littérature, cela en accord avec la volonté de présenter une procédure générique.

Lors de l'étape 2, il est intéressant de calculer le niveau d'anomalie des voxels issus de sujets sains pour deux raisons. Tout d'abord, cela permet de déterminer quels sont les niveaux d'anomalie présents dans des tissus sains : il s'agit des niveaux les plus faibles. À l'inverse, au sein des voxels issus de sujets pathologiques se trouvent les niveaux d'anomalies les plus élevés du fait des pathologies considérées. C'est cette division en deux groupes des niveaux d'anomalie qui permet de calculer le seuil utilisé pour la délimitation des pathologies.

La seconde raison à l'utilisation des voxels issus de sujets sains est que le modèle de signatures anormales de l'étape 4 contient ainsi une signature associée au

type sain, ce qui permet d'évaluer le taux de faux positifs, c'est-à-dire combien de sujets sains sont déclarés à tort comme pathologiques. Cette signature de tissus sains est également utilisée dans le post-traitement de l'étape 5 afin de supprimer une partie des zones saines déclarées à tort comme atypiques au sein de chaque sujet, ce qui permet d'améliorer le modèle de signatures anormales.

Les classifications non supervisées des étapes 1 et 3 sont réalisées en utilisant des modèles de mélange. Une comparaison des modèles basés sur des lois gaussiennes et ceux basés sur des lois de Student à mélange d'échelles multiples est réalisée de façon à déterminer l'impact du choix des lois de probabilités sur les classifications obtenues. Les lois de Student à mélange d'échelles multiples sont une généralisation des lois t de Student (Forbes et Wraith [7]) qui permettent d'étendre les modèles gaussiens pour incorporer de façon robuste des données atypiques, et également proposer une plus grande variété de formes pour les voxels vus comme observations de l'espace des paramètres IRM ; en particulier, il est possible d'obtenir des contours non elliptiques. Le modèle de mélange associé est détaillé dans la partie II.C du document supplémentaire de l'article section 3.4.

Enfin, comme il s'agit de mélange, il est nécessaire de déterminer le nombre de classes, ou composantes, à incorporer dans le modèle. Ce choix peut-être vu comme un problème de choix de modèles, auquel cas l'approche classique est de pénaliser la vraisemblance du modèle par la complexité du modèle. On se ramène ainsi à devoir choisir une pénalisation. Pour cela nous utilisons l'heuristique de pente dérivée de Baudry et al. [20]. Cette heuristique repose sur le fait que, sous certaines conditions de régularités, l'accroissement de la log-vraisemblance du modèle de mélange gaussien est linéaire en nombre de classes en cas de sur-ajustement. Ainsi, en ajustant différents modèles pour différents nombres de classes, il est possible de déterminer le nombre de classes à considérer pour décrire suffisant les données sans basculer dans le sur-ajustement.

3.3 Résultats sur des données d'IRM cérébrale chez le rat

Le protocole a été appliqué à des données d'IRM quantitatives provenant de 17 rats sains (6 pour l'apprentissage et 11 pour le test), et 36 rats porteurs de tumeurs cérébrales (9L, C6, F98 et RG2 - 26 pour l'apprentissage et 10 pour le test). Pour chaque rat, 5 cartes IRM ont été calculées : ADC, T1, T2, CBV, et AUC, et l'analyse des voxels a été faite à la fois avec le modèle de mélange de lois de Student à mélange d'échelles multiples (MMST) et le modèle de mélange

gaussien. La segmentation automatique du jeu d'apprentissage ainsi obtenue a été comparée avec celle manuelle de l'expert, et atteint après le post-traitement spatial un indice DICE de 0.704 dans le cas gaussien et de 0.721 dans le cas MMST. Similairement, l'Indice de Rand Ajusté [1] (ARI) est respectivement de 0.487 et de 0.581, ce qui montre à la fois une bonne concordance avec la segmentation experte et une performance accrue avec le modèle MMST. Ces résultats sont confirmés pour le jeu de tests où les scores DICE et ARI sont de 0.69 et 0.49 pour le modèle MMST.

La caractérisation des tumeurs se base quant à elle sur les proportions des classes obtenus lors de la classification des voxels déclarés atypiques, ainsi que sur la comparaison de trois algorithmes discriminants : une analyse linéaire discriminante (lda [21]), une modélisation par mélange gaussien (mclustda [22]) et une modélisation par mélange gaussien en grande dimension (hdda [23]). Le taux de bonne prédiction obtenu sur le jeu de test est de 0.857 en utilisant les proportions des classes issues du modèle de mélange gaussien et de 0.905 lorsque les classes sont issues du modèle MMST.

Les résultats ainsi obtenus présentent une très bonne cohérence spatiale alors que la structure spatiale des acquisitions n'a que peu été prise en compte dans les apprentissages statistiques (étape de nettoyage spatial). Il a également été obtenu une bonne concordance avec la segmentation manuelle, les écarts principaux observés étant dus : 1) à la connaissance spatiale a priori de l'expert, notamment en reconnaissant les ventricules comme des tissus non tumoraux, 2) à la difficulté de délimiter certaines tumeurs particulièrement diffuses. Enfin, la caractérisation des tumeurs a montré de bons résultats sur l'échantillon utilisé et doit maintenant être validée sur une étude plus large.

3.4 Article publié au journal IEEE Transactions on Medical Imaging

1. La définition des scores DICE et ARI est rappelée en annexe A, formules (A.1) et (A.2).

Fully Automatic Lesion Localization and Characterization: Application to Brain Tumors Using Multiparametric Quantitative MRI Data

Alexis Arnaud, Florence Forbes, Nicolas Coquery, Nora Collomb, Benjamin Lemasson, Emmanuel Barbier

Fully Automatic Lesion Localization and Characterization: Application to Brain Tumors using Multiparametric Quantitative MRI Data

Alexis Arnaud, Florence Forbes, Nicolas Coquery, Nora Collomb, Benjamin Lemasson, and Emmanuel L. Barbier

Abstract

When analyzing brain tumors, two tasks are intrinsically linked, spatial localization and physiological characterization of the lesioned tissues. Automated data-driven solutions exist, based on image segmentation techniques or physiological parameters analysis, but for each task separately, the other being performed manually or with user tuning operations. In this work, the availability of quantitative magnetic resonance (MR) parameters is combined with advanced multivariate statistical tools to design a fully automated method that jointly performs both localization and characterization. Non trivial interactions between relevant physiological parameters are captured thanks to recent generalized Student distributions that provide a larger variety of distributional shapes compared to the more standard Gaussian distributions. Probabilistic mixtures of the former distributions are then considered to account for the different tissue types and potential heterogeneity of lesions. Discriminative multivariate features are extracted from this mixture modelling and turned into individual lesion signatures. The signatures are subsequently pooled together to build a statistical fingerprint model of the different lesion types that captures lesion characteristics while accounting for inter-subject variability. The potential of this generic procedure is demonstrated on a data set of 53 rats, with 36 rats bearing 4 different brain tumors, for which 5 quantitative MR parameters were acquired.

Index Terms

Perfusion imaging, Magnetic resonance imaging (MRI), Animal models and imaging, Computer-aided detection and diagnosis, Probabilistic and statistical methods, Automatic segmentation, Automatic characterization, Brain tumor, Quantitative multiparametric MRI, Mixture model, Anomaly detection, Radiomics, Fingerprint model.

I. Introduction

MAGNETIC resonance imaging (MRI) is the recommended imaging modality for brain tumor analysis (De Angelis [1], Drevelegas and Papanikolaou [2], Wen et al [3]). Several sequences may be obtained from a single MRI exam. When diagnosis is required, the radiologist is then left with a potentially large number of information sources but relatively few analysis tools. In this work, we propose an automated data driven tumor identification procedure where identification includes both localisation (segmentation) and characterization (signature). Segmentation is an intermediate Region-Of-Interest (ROI) determination step to produce tumor signatures that provide representations of the observed lesions with respect to their tissue composition. These compositions being characterized by various physiological parameters in good accordance with the expected tissue types. The construction of such signatures is made possible by the availability of so-called quantitative MRI. Quantitative MRI refers to maps of meaningful physical or chemical variables that can be measured in physical units and compared between tissue regions and among subjects. The use of such quantitative data has emerged more recently (see *e.g.* [4] and the journal issue on quantitative Brain MRI [5]). Most clinical MRI acquisitions rely on so-called weighted images, whose contrast is determined by a combination of different factors, tissue or experiment dependent. To detect pathology, conventional intensity-based MRI (*e.g.* Menze et al [6]) relies on differences in signal intensities which are not specific to the underlying biological state. Nevertheless, the term quantitative is often used when numeric values of signal intensities are measured and used for tissue segmentation and classification. We consider here a more stringent definition of the term quantitative as described in [5]. An important and promising aspect of quantitative MRI is the possibility to perform a meaningful analysis beyond a few global features such as mass diameter, the occurrence of an edema or of contrast enhancement. In addition, as the number and relevance of images increase, the sources of variability, such as inter-operator difference or subjectivity, should be carefully minimized but most current clinical practice lacks quantitative and reproducible assessment (Hectors et al [7], Menze et al [6], Weltens et al [8]).

Several attempts, although less numerous than in conventional MRI analysis, have been made to analyze multiparametric quantitative MRI data to probe the information content of lesions. They usually consist of two steps, localization and characterization whose variability and accuracy can be controlled in two main ways respectively, through automated ROI selection

A. Arnaud and F. Forbes are with the Mistis team at Univ. Grenoble Alpes, INRIA, Laboratoire Jean Kuntzmann, Grenoble, France, E-mail: firstname.lastname@inria.fr. N. Coquery, N. Collomb, B. Lemasson, and E. Barbier are with Grenoble Institut des Neurosciences, Inserm U1216 & Univ. Grenoble Alpes, France, E-mail: firstname.lastname@univ-grenoble-alpes.fr.

and through quantitative feature extraction standardization. Most approaches focus on one or the other aspects: segmentation approaches are usually based on a few standard MRI maps, while more advanced feature extraction techniques commit to a preliminary manual ROI delineation. Coquery et al [9] propose to analyse 6 MR parameter maps and to identify different tissue types by looking for groups of voxels with similar parameter values. Voxels in manually segmented ROIs are clustered using a 6-dimensional Gaussian mixture model. The number of components (clusters) is chosen according to the Bayesian information criterion. Similarly, Boult et al [10] use manual segmentations followed by a k-means clustering of 3 MRI parameters maps to determine intra-tumoral tissue types whose relevance is assessed by comparison to histology data. In contrast, studies that focus on automated segmentation are generally faced with the issue of automatic lesion classes isolation. Unsupervised segmentation produces a partitioning into several classes with no clear semantic sense. Classes across different segmentations may not always represent the same tissue, complicating its biological interpretation. As an illustration, the intra-tumoral segmentation technique proposed by Katiyar et al [11] is restricted to a single tumor type localized and known in advance. These studies highlight the potential of conducting robust analysis from multiparametric quantitative data, but one common limitation is the number of tuning operations left to the user. Manual delineations have been already mentioned as an essential step, but most statistical inference procedures also rely on parameters that have to be set in advance such as the number of components in a mixture or the type of mixture distributions. Other limitations of the previous studies are that they do not account for dependencies between parameters or use multivariate Gaussian models because of their tractability in arbitrary dimensions and despite some observed parameter distributions have non Gaussian shapes [9]. In addition, the diagnosis ability of these approaches is not fully evaluated.

Therefore, there is a need for fully automated methods that can analyze multiple quantitative MR data in a reproducible way that correlates well with expert analyses.

A first challenge is the design of multivariate models that can capture non trivial interactions between physiological parameters while remaining tractable. The effort has to be put on the distributional modelling of the observed parameters whose deviation from standard Gaussian shapes may be of high significance. Similarly, extreme values of some of the parameters should be adequately modeled as important information may lie in the tails of the distributions rather than in their central part. For example, in human patients, glioblastoma are evaluated based on the high CBV (Cerebral Blood Volume) hotspot [12]. CBV values twice higher than that of normal appearing white matter are considered as originating from aggressive tumor tissue. To capture such non trivial interactions between multiple parameters, the usual multivariate Gaussian distributions but more generally the so-called elliptical distributions (Gaussian, multivariate Student, Laplace distributions, etc.) are limited by the type of elliptical shapes they allow. Observed physiological parameters seldom fit into such elliptical shapes (see Figure 1 below). Alternatives distributions, with a large variety of shapes, exist such as those using copula modelling [13]. Unfortunately copula models become rapidly intractable when more than 2 parameters have to be jointly modeled. When more than 2 parameters are available, it is important to design models that are both flexible in shapes and tractable in higher dimension. This is the case of the multiple scale t-distribution (MST) introduced in [14] which goes far beyond the standard Student distribution in terms of possible (not restricted to elliptical) shapes.

In addition to an accurate account of multiple parameters interactions, a second challenge is to perform accurate lesion localization. There are many ways to achieve lesion localization. The relevance of each of these ways depends on the available data. In this study we consider a weakly supervised case in which a moderate number of healthy (or control) subjects are available and identified as such. This automatically excludes deep learning methods that require a large amount of labelled voxels. Indeed, so far the most striking successes in deep learning have involved discriminative models and supervised classification tasks (see [15]–[17] and references therein). Models that have unsupervised learning capability with less requirement on ground truth labels are needed. Generative Adversarial Networks (GAN) [18] are unsupervised models that have been applied to natural images but have not been yet really assessed in a medical imaging context. Data augmentation and transfer learning, or the use of pre-trained networks, are promising directions of research but not quite mature yet. To our knowledge, there exists no pre-trained network or architecture for quantitative multi-modal medical images. Learning efficiently from limited data is still an important area of research. We believe more traditional unsupervised techniques are still useful to solve unsupervised tasks and to produce some proxy to the manual segmentations that are needed for neural networks. Among unsupervised segmentation methods, the vast majority of methods are clustering methods. But these are fully unsupervised and do not make use of the availability of a set of healthy subjects.

In our approach, we rather consider a novelty detection approach based on the identification of lesioned voxels as outliers with respect to a previously built reference model. Outlier detection can also be performed using a clustering approach (as in Van Leemput et al, 2001 [19], Gebru et al, 2016 [20] or Cuesta-Albertos et al, 2008 [21]) but most proposed methods are designed for only a few images per subject, usually less than 3. Detecting outliers in one dimensional data is less challenging than in multivariate data. In the multivariate case, the use of symmetric distributions may not be satisfying: the amount of outlying data has to be the same in each dimension, and this has no reason to occur in multi-contrast MRI data [22]. The proposed use of the MST distribution addresses specifically this problem. This is illustrated in Section III with statistical tests that reject Gaussianilty in both the healthy and pathological data cases. At last, one inconvenient of clustering approaches is that they usually require to use an atlas and to fix some sort of hyperparameters indicating for instance the expected number of outliers. In our novelty detection approach, we also have to set thresholds to distinguish intra-lesion classes but we propose

a data driven way based on model selection tools (for the determination of the number of classes) and extreme value theory (for the characterization of the thresholds as quantiles).

The approach proposed in the paper differs from existing work in various aspects. The use of flexible multivariate MST distributions allows to accumulate information from several (more than 3) physiologically meaningful parameters and therefore, successively 1) to build an accurate reference model from control subjects; 2) to perform automated lesion localization via novelty detection based on all imaged parameters but without the need of an atlas and hyperparameters tuning; and 3) to determine intra-lesion segmentations that can be turn into signatures and used to discriminate between different tumor types. The whole procedure is described in Section II and summarized in Figure 2. In Section III, its performance is illustrated on an independent evaluation data set.

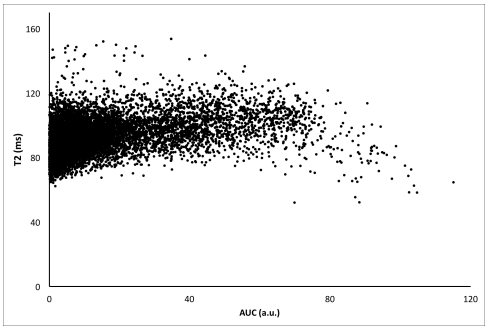


Fig. 1. Two observed physiological parameters: T2 vs AUC parameters for rats with C6 tumors. T2 and AUC values are plotted for each voxels in the tumors ROIs, illustrating skewness and different tail weights in the T2 and AUC directions. As expected, the permeability (*i.e.* AUC) measured in C6 tumors exhibits a large variability.

II. Proposed automatic and data-driven procedure

In the following developments, a generic and automatic data-driven procedure is described and its performance to segment and diagnose brain tumors without any high level expert or spatial knowledge is illustrated. The proposed procedure requires the availability of a dataset made of two sets of voxels: one from healthy subjects and one from pathological subjects for which the pathology (*e.g.* tumor) type is known. The considered features come from multiparametric quantitative MRI data that provides in each voxel a vector with several measures computed during the MRI session. Quantitative MR measures are considered, in particular, to be as independent as possible of the MRI scanner or the study center (Tofts [23]). In contrast to conventional MRI, image preprocessing issues such as bias field correction, resolution, etc. are less critical because parameters maps can be built so as to avoid most of these complications. For instance, in quantitative imaging, intensity bias is taken care of during the computation of parameter maps. In addition, the kind of data we consider here are acquired using the same geometry and at the same spatial resolution.

The proposed procedure consists then of five steps: i) a first mixture model is fitted to the healthy subjects voxels; ii) this reference model is used to detect voxels which exhibit abnormal MR features with respect to the reference model, in the healthy and pathological subjects; iii) a second mixture model is fitted to the detected abnormal voxels and yields a clustering of these voxels into several classes; iv) the proportions of these classes in each subject are used as a signature of the pathology and a discriminative (fingerprint) model is learned that can distinguish between different pathology types; v) an additional spatial post-processing can be carried out to remove some spatial artifacts and refine the pathology signatures.

A. Reference model

Starting from a set of reference, typically healthy subjects, the goal is to construct a statistical parametric model of the MR parameters associated to these subjects. Each reference subject is associated to a number M of co-localized MR parameter maps that provide for each voxel v a M-dimensional vector of parameters denoted by $\boldsymbol{y}_v$. All voxels from all subjects are gathered into a single set of voxels denoted by $\mathcal{V}_H$. The considered data set of M-dimensional vectors, pooling all vectors together, is denoted by $\boldsymbol{Y}_H = \{\boldsymbol{y}_v, v \in \mathcal{V}_H\}$. To characterize the distribution of these MR parameters, we consider a multivariate mixture model to account for the potential heterogeneity in the parameter values due to the presence of different tissue types. This corresponds to cluster the data $\boldsymbol{Y}_H$ into a number of groups (clusters) of similar parameter vectors and to model each group with a parametric distribution. In practice, in each dimension, the data are standardized to avoid scaling effects between MR parameters. This standardization is made at the whole dataset level and not for each subject individually.

In Coquery et al [9], a Gaussian mixture model is used assuming that each group is distributed according to a Gaussian distribution. However, the observed physiological parameters do not necessarily exhibit a Gaussian shape. Also Gaussian distributions are known to be sensitive to outliers whose occurrence may severely bias the estimation. As a more robust alternative, heavy tail distributions have the ability to accommodate potential outliers. In this paper, we consider such distributions and

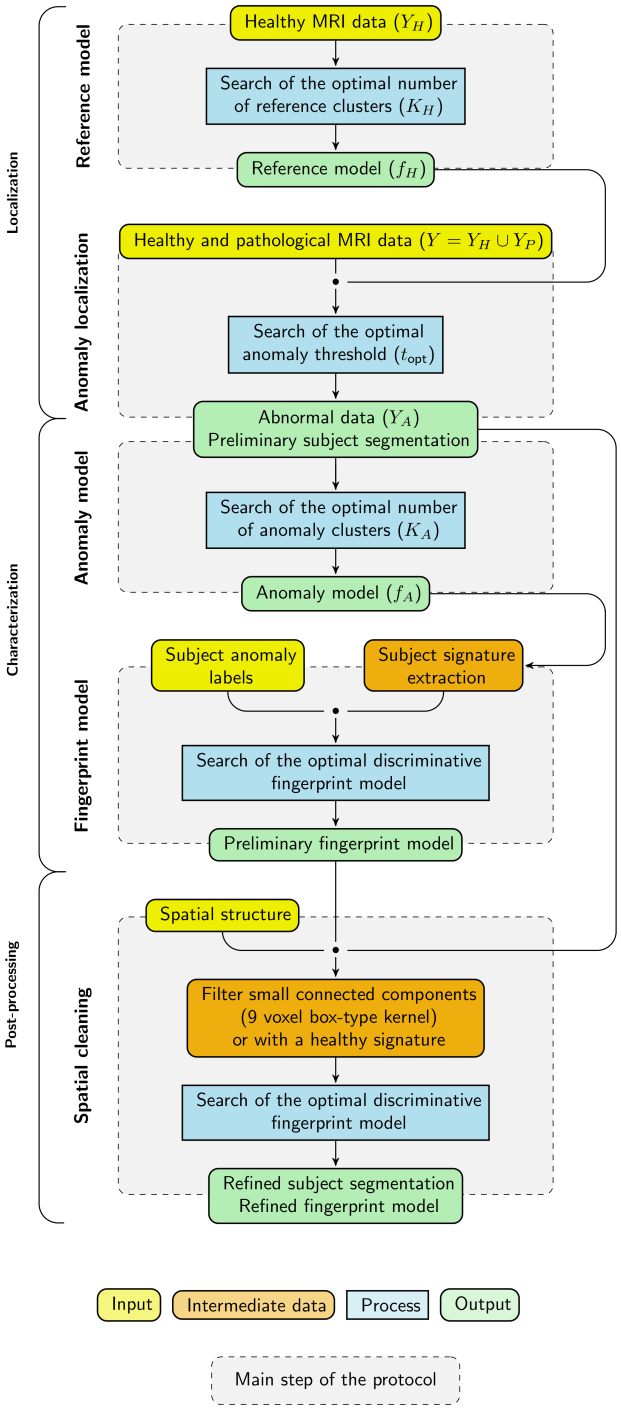


Fig. 2. Construction of a model to automatically localize and characterize lesions. Starting from subjects labeled as healthy or pathological, the procedure is made of 5 main steps.

in particular multiple scale t-distributions (MST) introduced in Forbes and Wraith [14] that generalize the standard Student t-distribution. MST distributions allow the assignment of outlying parameter values to clusters without degrading the location and scale of the clusters. One advantage of the MST distribution over the standard t-distribution is a varying amount of tail weight in each dimension resulting in a much greater variety of distributional shapes. The mixture of MST distributions (MMST) that best fits the observed data $\boldsymbol{Y}_H$ is estimated using an Expectation-Maximization (EM) algorithm as described in Forbes and Wraith [14]. This requires to set the number of groups in the mixture. This number, denoted by K_H, is chosen automatically from the data following a model selection approach. More specifically, we use the so-called slope heuristic (Baudry et al [24]) that generally provides a clear selection (see supplementary materials for more details). The clustering EM algorithm results in K_H clusters, each represented by a multivariate M-dimensional MST distribution. Each distribution summarizes the MR features of a group of voxels and can be seen as a *word* in a *dictionary* of healthy or reference clusters. A by-product of the

EM algorithm is that each voxel has a probability to be assigned to each cluster. The set of these membership probabilities can be seen as the signature of the voxel in the coding provided by the dictionary, the whole dictionary being itself summarized by the probability distribution function (pdf) of the estimated MMST model, namely

$$f_H(\boldsymbol{y}) \quad = \quad \sum_{k=1}^{K_H} \pi_k \, \mathcal{MST}(\boldsymbol{y}; \psi_k) \tag{1}$$

where $\mathcal{MST}(\boldsymbol{y}; \psi_k)$ denotes the MST pdf with parameter ψ_k and proportion π_k learned from $\boldsymbol{Y}_H$. The pdf in (1) is referred to as the reference model and will be used in the following anomaly detection step.

B. Anomaly localization

Considering a set of pathological subjects, the goal is to identify in each of them the lesioned voxels in order to provide a delineation of potential lesions. As lesions generally exhibit MR parameter values different from healthy tissues, lesion localization and delineation is recast into a novelty detection task. For a precise localization, mere visual inspection may not be enough and may be tedious when the number of available MR parameters increases. As before, the voxels of all pathological subjects are gathered into a set $\mathcal{V}_P$ of voxels and the corresponding set of MR vectors is denoted by $\boldsymbol{Y}_P = \{\boldsymbol{y}_v, v \in \mathcal{V}_P\}$. To deal with comparable values, the normalization applied to the reference observations $\boldsymbol{Y}_H$ is also applied to $\boldsymbol{Y}_P$. A voxel v is considered as abnormal with regards to the previously built reference model f_H in (1) if it corresponds to MR parameter values $\boldsymbol{y}_v$ with a low likelihood in the reference model. This likelihood is assessed via the reference pdf value at $\boldsymbol{y}_v$, *i.e.* $f_H(\boldsymbol{y}_v)$. Novelty detection can be used not only to detect the lesioned voxels from the normal ones but among them to identify various degrees of proximity to normality. The log-density score $\log f_H(\boldsymbol{y}_v)$ is considered as a measure of proximity of one voxel v (associated to value $\boldsymbol{y}_v$) to the reference healthy model (represented by f_H).

When considering the maps of all subjects, there are as many log-scores as there are voxels in the whole set of subjects. The idea is to look at the distribution of all these log-score values. As some of the voxels correspond to healthy tissues, there should be a group of voxels whose log-scores are high meaning a good adequacy with the reference model. More generally, we assume that there exist groups of voxels with similar log-scores and approximate the log-score values distribution as a MST mixture model. The number of groups L is determined automatically with the slope heuristic and a partition of the voxels into L groups is deduced. These groups can be ordered according to their mean log-score, from the farthest to the closest to the reference model. Then thresholds denoted by $\{t_1, \ldots, t_L\}$ are set to values that reflect the borders between two successive groups. Among these thresholds, the t_{opt} threshold that corresponds to the retained lesion segmentation is just one of them. The best 2-group partition of the log-scores is determined and the threshold that matches best with this partition is chosen as t_{opt}. For more explanations, see Appendix A. Figure 5 illustrates the obtained nested anomaly segmentations that reflect different abnormality levels and are indicative of different tissue types within lesions. The global threshold t_{opt} can also be used to separate the input data $\boldsymbol{Y} = \boldsymbol{Y}_H \cup \boldsymbol{Y}_P$ into a set of abnormal values $\boldsymbol{Y}_A$ and the rest:

$$\boldsymbol{Y}_A = \{\boldsymbol{y}_v, \; v \in \mathcal{V}_H \cup \mathcal{V}_P \quad s.t. \quad f_H(\boldsymbol{y}_v) < t_{opt}\} \; .$$

Note that, as the proposed thresholds are quantiles of the reference model pdf, healthy subjects may also exhibit a small fraction of abnormal voxels. Some of them are isolated voxels and can be easily removed using simple morphological operators (see Section II-E). However, these isolated voxels tend to be present in all subjects, for similar reasons (noise, skull stripping, artifacts, etc.), and their removal does not significantly affect the discriminative power of the fingerprint model. In contrast, other voxels may correspond to normal regions or structures (*e.g.* vessels, ventricles) whose physiological characteristics and then MR parameters are close to that of lesioned tissues. Individually, these voxels are therefore correctly detected as deviant from the reference model. However, their global signature is expected to be different from the signature of lesioned tissues and will then be learned by the fingerprint model.

C. Anomaly model

The goal of the two previous steps was essentially to perform automatic ROI localizations. The obtained ROIs provide a set of MR parameter vectors that are referred to as the abnormal data set $\boldsymbol{Y}_A$. An anomaly model is then constructed following the same procedure as for the reference model (Section II-A). The observations in $\boldsymbol{Y}_A$ are standardized in each dimension and then used to fit a MMST model with a number K_A of clusters selected with the slope heuristic. The fitted mixture is denoted by f_A,

$$f_A(\boldsymbol{y}) \quad = \quad \sum_{k=1}^{K_A} \eta_k \, \mathcal{MST}(\boldsymbol{y}; \phi_k) \tag{2}$$

where $\mathcal{MST}(\boldsymbol{y}; \phi_k)$ denotes the MST pdf with parameter ϕ_k and proportion η_k learned from $\boldsymbol{Y}_A$. This anomaly model is used in the next section to extract anomaly features from MR maps and construct a signature for each subject under consideration.

D. Fingerprint model

By fingerprint model we mean a model that can correctly characterize and classify a subject into one of a number of classes (*e.g.* different tumor types), based on MR parameters maps. Such a model is built in a supervised manner from pairs associating some chosen features to a class label. This requires the availability of a number of subjects for which the class label is known. The extracted features have then to be as informative as possible so as to allow the correct classification of unlabeled subjects. For each available subject in the learning data set, its anomaly class or label is known. These classes typically include the *healthy* label and different tumor types. For each subject S, features are extracted from a set $\mathcal{V}_S$ of n_S voxels corresponding to the voxels of S detected as abnormal in the previous step (Section II-B). As mentioned in Section II-B, healthy subjects also exhibit a small fraction of abnormal voxels. For each voxel $v \in \mathcal{V}_S$, the anomaly model (Section II-C) provides a probability ρ_k^v that voxel v belongs to cluster k among K_A clusters,

$$\rho_k^v = \frac{\eta_k \, \mathcal{MST}(\boldsymbol{y}_v; \phi_k)}{\sum_{l=1}^{K_A} \eta_l \, \mathcal{MST}(\boldsymbol{y}_v; \phi_l)}.$$

For features at the subject level, we compute for each $k = 1 : K_A$, the mean probability over voxels in $\mathcal{V}_S$,

$$\rho_k^S = \frac{\sum_{v \in \mathcal{V}_S} \rho_k^v}{n_S}.$$

The retained signature vector for subject S is then

$$\boldsymbol{\rho}^S = \{\rho_1^S, \ldots, \rho_{K_A-1}^S, n_S\} \tag{3}$$

where the last probability has been removed and replaced with the ROI size n_S to avoid co-linearity. Such a vector captures the expression level of each anomaly cluster in the ROI of subject S. Intuitively, it is expected to capture the proportions of the different tissue types in the ROI. The addition of the ROI size seems natural at the stage as we suspect size could be discriminant if large number of pathologies and lesion types are considered. In the experiments on tumors made in Section III, size appears to be useful while not being essential to discriminate between different tumor types, as indicated by the similar ratios of the between-group variance over the total variance for each feature.

As for the supervised learning part, we adjust different discriminant analysis models and compare their ability to correctly predict the label (*e.g.* lesion type) of each subject by a leave-one-out cross-validation procedure. The selected discriminant analysis model is the one providing the highest true positive rate. Further details are given in the application Section III-D.

E. Post-processing

The segmentations obtained in Section II-B can be further refined by removing connected components which are too small. This is done by applying an erosion-dilatation operator with a 9 voxels box-type kernel on each slice. In addition, the fingerprint model can improve the segmentations thanks to its ability to recognize healthy tissue. A fingerprint model is a signature definition as given in equation (3) (proportions of voxels in different groups) and a classifier able to distinguish between different signatures. A fingerprint model can then classify any ROI into a lesion type or as healthy. Therefore when a lesion is made of different connected components, these components can be seen as separated lesions for which a signature can be computed. Since signatures live all in the same space, whatever the size of the ROIs they are representing, the previously constructed classifier can be applied to classify each connected component separately. When a connected component is classified as healthy, we propose to remove it from the lesion. When doing this for all initial lesion segmentations, we obtain then (smaller) refined segmentations. Going back to the procedure described in Section II-D, the refined segmentations can in turn be used to relearn a refined classifier in two steps. First, since the removal of connected components may affect the different pixels proportions, signatures of the refined segmentations have to be recomputed. Then these new signatures can be used as input for the estimation of a new fingerprint model, referred to below as the refined fingerprint model. This post-processing enables the removal of groups of voxels that may individually exhibit MR parameter values close to lesioned tissues but show group anomaly proportions (signature) close to healthy components. After this stage, healthy tissues should correspond to a null signature (*i.e.* with no abnormal pixels) while the lesioned tissues signatures are cleaned from healthy tissues. An illustration is given in Figure 8.

F. Characterization and prediction

To predict the label of a new subject, its ROI is first determined using f_H the reference model (Section II-A) for anomaly detection (Section II-B). The obtained ROI is cleaned from potentially remaining healthy connected components that have not been removed by erosion-dilatation. The initial fingerprint model is used to identify these healthy connected components. The cleaned ROI (see Figure 6) is used to extract a $\boldsymbol{\rho}^S$ signature (see Figure 8) using the anomaly model (Section II-C) as explained in Section II-D. This signature is then given to the refined fingerprint model which provides an associated label (Table III).

III. Application to real multiparametric MRI data from rats with brain tumors

TABLE I

AVAILABLE DATA: A LEARNING SET ($\boldsymbol{Y} = \boldsymbol{Y}_H \cup \boldsymbol{Y}_P$) OF HEALTHY ($\boldsymbol{Y}_H$) AND PATHOLOGICAL ($\boldsymbol{Y}_P$) MR VALUES IS USED TO LEARNED A FINGERPRINT MODEL. ANOTHER TEST SET ($\boldsymbol{Y}^T$) OF BOTH HEALTHY ($\boldsymbol{Y}_H^T$) AND PATHOLOGICAL ($\boldsymbol{Y}_P^T$) MR VALUES IS USED FOR VALIDATION. THE SETS OF DETECTED ABNORMAL VOXELS ARE ALSO INDICATED IN BOTH CASES ($\boldsymbol{Y}_A$ AND $\boldsymbol{Y}_A^T$). THE OBSERVATIONS DIMENSION IS 5.

Learning set	voxels	subjects
$\boldsymbol{Y} = \boldsymbol{Y}_H \cup \boldsymbol{Y}_P$	260405	32
$\boldsymbol{Y}_H$	45051	6
$\boldsymbol{Y}_P$	215354	26
$\boldsymbol{Y}_A$	57547	32

Test set	voxels	subjects
$\boldsymbol{Y}^T = \boldsymbol{Y}_H^T \cup \boldsymbol{Y}_P^T$	150085	21
$\boldsymbol{Y}_H^T$	71340	11
$\boldsymbol{Y}_P^T$	78745	10
$\boldsymbol{Y}_A^T$	20105	21

OUR procedure is illustrated on a data set of 53 rats for which 5 quantitative MRI maps are available. Some of the rats were implanted with different tumor types. The study design was approved by the local institutional animal care and use committee (COMETHS). All animal procedures conformed to French government guidelines and were performed under permit 380820 and B3851610008 (for experimental and animal care facilities) from the French Ministry of Agriculture (Articles R214-117 to R214-127 published on 7 February 2013). This study is in compliance with the ARRIVE guidelines (Animal Research: Reporting in Vivo Experiments [25]).

A. Description of the MRI data

a) Rats and tumor types (Table I): The healthy subject group ($\boldsymbol{Y}_H$) is composed of 6 healthy Fisher rats. The pathological subject group ($\boldsymbol{Y}_P$) contains 26 subjects with 4 tumor types: 9L (6 Fisher rats), C6 (6 Wistar rats), F98 (7 Fisher rats) and RG2 (7 Fisher rats). The MR parameter maps of all these subjects form the data set $\boldsymbol{Y}$. For evaluation purpose, another group of subjects is available and contains 5 rats with 9L tumor, 5 rats with F98 tumor, and 11 healthy Fisher rats. The MR parameter maps associated to these subjects are kept as a test set and denoted by $\boldsymbol{Y}^T$ distinguishing the healthy sub-group $\boldsymbol{Y}_H^T$ from the pathological one $\boldsymbol{Y}_P^T$.

b) MRI parameters (Figure 3): The following 5 quantitative MR maps were acquired on 5 contiguous slices: apparent diffusion coefficient (ADC), T1, T2, CBV, and a vessel permeability map called area under the curve (AUC). All measures are naturally co-localized: all maps were acquired with the same geometry so that each voxel is described by the 5 parameters above. Section III-A in the supplementary materials provides further details about the MRI session. In addition, an anatomical T2-weighted image was acquired to allow automatic skull-striping. A manual delineation (superimposed red line in Figure 3 first row) of the tumor was performed using the anatomical image and the diffusion map. This manual segmentation is used as ground truth for the evaluation of the automatic tumor localization proposed in this study.

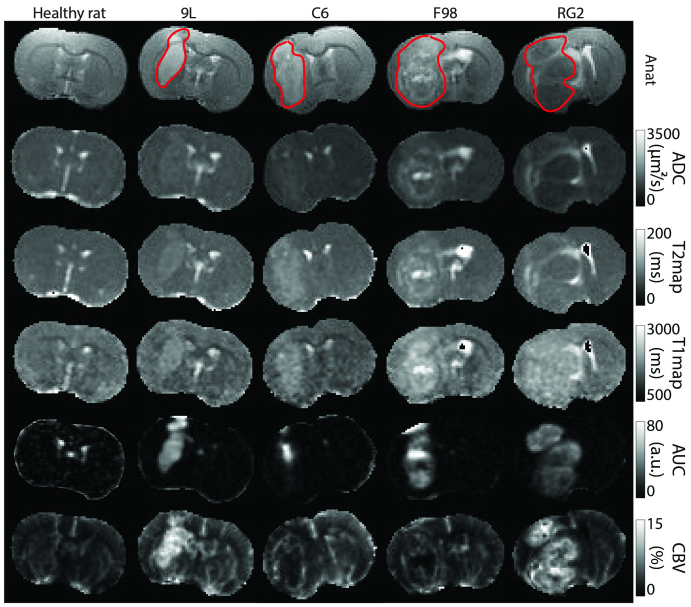


Fig. 3. MRI maps -ADC, T2, T1, AUC, CBV- associated to the central slice for one rat of each group (from the data set Y). The red line superimposed on the anatomical map corresponds to the manual segmentation.

B. A healthy subjects based reference model

The reference model f_H is built as a MMST model using an EM algorithm on the healthy subjects maps $\boldsymbol{Y}_H$. When defining this reference model, the slope heuristic approach selects $K_H = 10$ clusters. An example of this partitioning in $K_H = 10$ clusters is given in Figure 4 for the 5 slices of a healthy rat. In the whole set of healthy rats, 4 main clusters (red -1-, orange -2-, yellow -3-, light green -4-) gather 79.5% of the voxels in $\boldsymbol{Y}_H$. The 6 remaining clusters (green -5-, turquoise -6-, blue -7-, dark blue -8-, purple -9-, light purple -10-) correspond to less represented features such as vessels and ventricles (clusters 9 and 10) or interfaces between the main tissue types (interface between gray matter and cerebro-spinal fluid for clusters 6 and 7). Interestingly, although the model is adjusted without any spatial regularization, the resulting segmentations present spatially homogeneous regions and are rather visually consistent with brain anatomy: clusters 1 and 4 for the cortex and the corpus callosum, clusters 1 to 3 for the striatum, clusters 7 to 10 for the ventricles. The fitted MST distributions potentially include Gaussians that can be recovered by setting the degrees of freedom parameters to large values in the MST model [14]. To check that the fitted MST mixture does not reduce to a Gaussian mixture, the MST mixture component with the largest number of voxels was tested against Gaussianity using an Anderson-Darling test for each of the 5 parameters separately, (ADC, AUC, CBV, T1, T2). The respective p-values were (1.04e-06, 1.08e-01, 7.14e-24, 4.39e-15, 5.91e-19) indicating that only the AUC parameter could be considered as Gaussian, the probability of the data under the Gaussian distribution being less than 1e-06 for the other parameters. A similar conclusion was found with another statistical test referred to as the Shapiro-Wilk test.

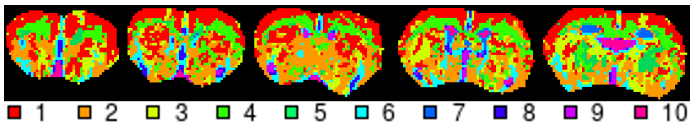


Fig. 4. Reference model clustering (with $K_H = 10$ clusters) on the 5 slices of one healthy rat. From left to right: slices in decreasing order along the vertical axis of the scanner.

C. Anomaly localization

As described in Section II-B, for each MR parameter vector $\boldsymbol{y}_v$ in $\boldsymbol{Y}$, the pdf $f_H(\boldsymbol{y}_v)$ is evaluated. We found that these values are best partitioned into 7 groups, according to the slope heuristic. Within each group, the voxels have similar likelihoods in the reference model, that is similar degrees of abnormality with respect to the reference model. This partition leads to 7 anomaly thresholds $\{t_1, \ldots, t_7\}$. The optimal threshold that best divides the voxels into 2 subsets, according to a 2-component mixture model (Section II-B), is the 4th threshold $t_{\text{opt}} = t_4$. The abnormal subset $\boldsymbol{Y}_A = \{\boldsymbol{y}_v, f_H(\boldsymbol{y}_v) \leq t_{\text{opt}}\}$ represents 22.1% of the full data set Y with a false positive rate of 1.4%, *i.e.* $p(f_H(\boldsymbol{Y}_v) \leq t_{\text{opt}}) = 0.014$ when the random variable $\boldsymbol{Y}_v$ is distributed according to the reference model. The voxels in $\boldsymbol{Y}_A$ form the subject ROIs.

An illustration is given in Figure 5-A last row. Colored voxels (thresholds t_1 -red-, t_2 -orange-, t_3 -yellow- and t_4 -green-) belong to $\boldsymbol{Y}_A$ while grey ones (thresholds t_5 -light grey-, t_6 -medium grey- and t_7 -dark grey-) are not tagged as abnormal. The concordance with manual segmentation (superimposed red line in Figure 5-A) is indicated in Figure 5-B right via the computation of the Adjusted Rand Index (ARI) for each threshold. The higher the ARI the better, the maximum value being 1 (Rand [26], Hubert and Arabie [27]). For a given threshold t_l, MR parameters $\boldsymbol{y}_v$ in $\boldsymbol{Y}$ such that $f_H(\boldsymbol{y}_v) \leq t_l$ are selected. The corresponding voxels form the segmentations linked to threshold t_l, on which the ARI is computed. More specifically, the ARI is computed at level t_l for each rat. The boxplots in Figure 5-B show the variations of these ARI's independently of the tumor type. Other scores such as the DICE [28] are shown in Supplementary-Table III. Regarding global tumor delineation, it appears that the selected 4th threshold in the MMST case, corresponds to the best all tumor average ARI (0.42). ARI values can also be averaged for rats with the same tumor type. It appears then that some tumors are easier to segment. For instance, in the MMST case, for threshold t_4, the average ARI's are respectively of 0.49, 0.40, 0.32, and 0.47 for 9L, C6, F98 and RG2 tumors. The corresponding boxplots are shown in Supplementary Figure 5a. Inside the tumors, the thresholds also yield some satisfying spatial coherence. The strongest abnormality (red -t_1- and orange -t_2- areas in Figure 5) is mainly located at the center of the tumor area. The delineated regions around (yellow -t_3- and green -t_4-) match with the border of the tumor area, and the highest thresholds ($\{t_5, t_6, t_7\}$ -gray levels-) are mainly for the healthy voxels. However, it also appears that some voxels are tagged as abnormal in healthy subjects (Figure 5-A 1st column) or in the contralateral part of pathological subjects (Figure 5-A 4 last columns). Based on their anatomical location, these voxels mainly correspond to ventricles and possibly blood vessels. As mentioned in Section II-E, most of these wrongly tagged voxels will be easily removed in Section III-D by using their signature after morphological operations. In terms of segmentation for the pathological rats, if we except the contralateral voxels, the 9L rat shows a high concordance with the manual segmentation, while the F98 segmentation is smaller and the RG2 segmentation is bigger. This is consistent with the intrinsic difficulty of delineating tumors of varying visual appearance across MR maps. As a matter of fact, on anatomical (Figure 5, 1st row) and diffusion images, 9L tumors are easier to delineate manually while F98 and RG2 tumors are more diffuse.

For comparison, our protocol is also applied with Gaussian mixture (GM) models. The anomaly localization results are shown in Figures 5-A, second row and B, left. The MMST model provides finer intra-tumoral descriptions with 4 abnormal

classes instead of 3 in the Gaussian case, and smoother segmentations. This is quantitatively confirmed by a MMST ARI of 0.42 which is 10.5% higher than the Gaussian ARI of 0.38. If healthy parts are considered instead, MMST and GM ARI are in the same order. Similar conclusions hold for the DICE values given in Supplementary-Table III. Regarding lesion segmentation, the main difference is in the contralateral areas. With GM models, the contralateral areas detected as abnormal are larger, and more of them are connected to the lesion areas, which leads to a less effective spatial post-processing. Indeed, only connected components not connected to the lesion area can be removed, because in case of contact the ROI signature does not correspond to the healthy one. As shown in Figure 6 for 9L and F98 tumors, the GM case requires more user interpretation than the MMST model to differentiate the lesion from the contralateral area.

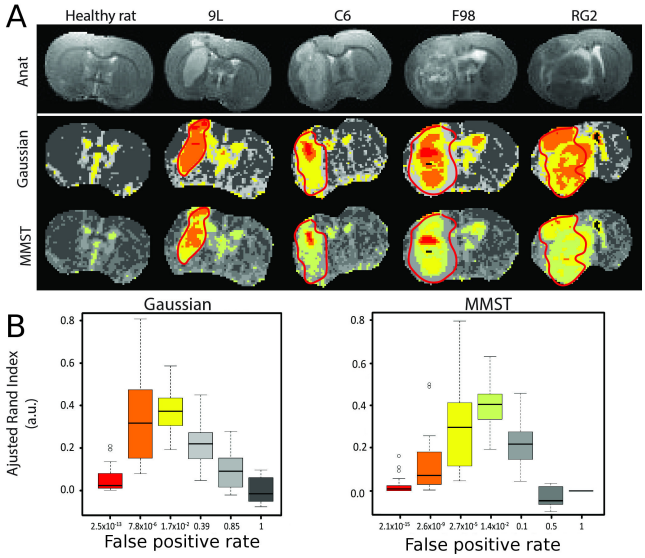


Fig. 5. Nested anomaly segmentations for the different anomaly thresholds (part A), with the associated false positive rates and respective adjusted rand indices (part B) for the overlap with the manual segmentation. The thresholds are ordered from lowest (most abnormal in red) to highest (less abnormal in dark gray). The colored ones are those defining the abnormal data set $\boldsymbol{Y}_A$. The gray ones correspond to normal groups. The ARI value for a threshold is obtained by concatenating the segmentations associated to the thresholds lower to this threshold. The results are given for both Gaussian mixture and MMST models.

D. Anomaly and fingerprint models

A MMST model f_A is adjusted on the abnormal $\boldsymbol{Y}_A$ data set with a number of clusters $K_A = 10$ selected using the slope heuristic. As in Section III-B, Anderson-Darling tests confirmed that the data did not exhibit Gaussian distributions with p-values of (3.70e-24, 3.70e-24, 3.75e-14, 4.62e-04, 3.58e-11) respectively for parameters (ADC, AUC, CBV, T1, T2). For each rat S in the learning set, we then extract a signature $\boldsymbol{\rho}^S$ as explained in Section II-D representing the proportions of each cluster in the ROI. The signatures are associated to the known subject labels in order to build a discriminative fingerprint model. Three discriminant analysis are compared based on their true positive rate with a leave-one-out procedure: linear discriminant analysis (lda: 90.6%, R package MASS, Venables and Ripley [29]), high dimensional discriminant analysis (hdda: 93.7%, R package HDclassif, Bergé et al [30]), and a discriminant analysis based on a Gaussian finite mixture modeling (mclustda: 90.6%, R package mclust, Fraley et al [31]). The hdda analysis is retained due to its higher score to built a first fingerprint model. This fingerprint model is used in turn to refine the segmentations made in Section III-C. Too small connected components are removed with a 9 voxel box-type kernel and the fingerprint model is used to remove other connected components classified as healthy.

The obtained cleaned segmentations are illustrated for some rats in Figure 6. The ARI values for the 26 pathological rats in the training set are summarized in Figure 7 (boxplots), while mean DICE values are shown in Table II. Similar boxplots for DICE can be seen in Supplementary Figure 12. A signature for each subject is then computed using the cleaned ROI possibly still made of several connected components. The signatures are shown in Figure 8 for the 4 pathological groups and the healthy group. As a result of the post-processing, the healthy subjects present empty signatures. Although the four tumor models used in this study are aggressive glioma with a median survival of 3 to 4 weeks, the proportions of clusters differ between tumors. The 9L tumor is mainly composed of cluster 6 (turquoise), and to a lesser extent of clusters 1 (red), 3 (yellow), 9 (purple), and 5 (green). Cluster 6 is characterized by a high CBV and a mild edema as compared to normal tissue. This correspond to the high vascular density previously reported in 9L tumors [32]. The C6 tumor is mainly composed of clusters 2 (orange) and 3, and to a lesser extent of clusters 5, 1, 6 and 9. Clusters 2 and 3 are characterized by a high permeability and CBV values similar to that of normal tissue. This corresponds to the combination of low vascular density and large vascular diameter reported in [33] The F98 tumor is mainly composed of cluster 2, but also of clusters 3, 5, and 6.

Cluster 2 exhibits ADC, permeability, T1, and T2 values above that of cluster 3, and CBV below that of cluster 3, in line with [9]. Finally, tumor RG2 is mainly composed of clusters 1 and 4 (light green), and to a lesser extent of clusters 2, 3, 5, 6, and 7 (blue). Cluster 1 is characterized by its normal ADC and its high permeability and CBV as compared to normal values. This signature characterizes an angiogenesis without edema and corresponds to previous reports [34]. These results suggest that all tumors are heterogeneous and that different tissue types can lead to an aggressive tumor.

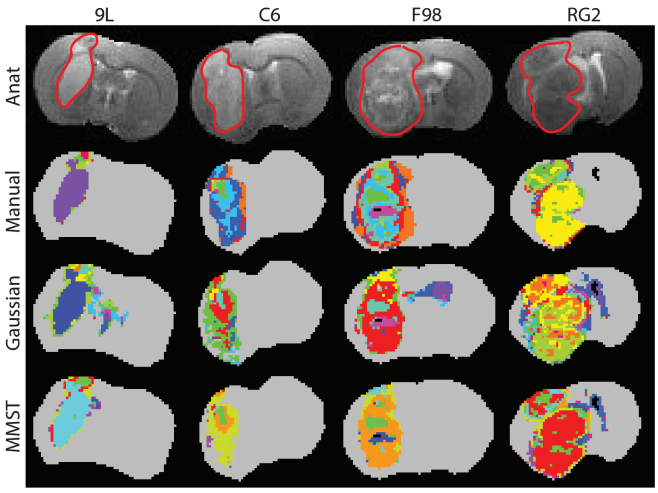


Fig. 6. Intra-tumoral segmentations for some of the pathological rats in the training set. First two rows: manual delineations and Gaussian clustering ($K_A = 13$) as described in [9]. Last two rows: Automatic segmentations and clustering using $K_A = 13$ and $K_A = 10$ with a Gaussian (3rd row) and MST (4th row) mixture model. The ROIs correspond to the refined segmentation (i.e. after spatial post-processing).

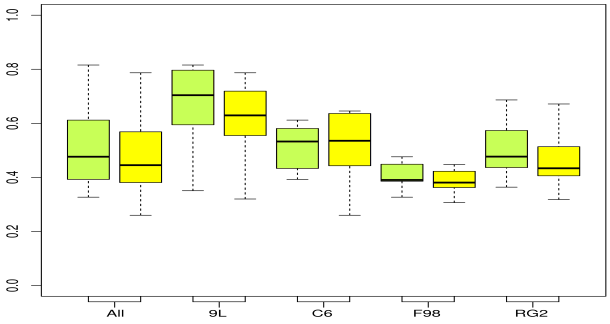


Fig. 7. Pathological rats in the training set: Adjusted rand index values per tumor type for the refined segmentations (i.e. after spatial post-processing), using the MMST (green) or Gaussian mixture (yellow) models.

TABLE II
PATHOLOGICAL RATS IN THE TRAINING SET: MEAN COVERING SCORES FOR LESION SEGMENTATIONS AFTER POST-PROCESSING USING GAUSSIAN MIXTURE AND MMST MODELS.

	Gaussian mixture model					MMST model				
	9L	C6	F98	RG2	All	9L	C6	F98	RG2	All
DICE index	0.677	0.636	0.665	0.823	0.704	0.718	0.644	0.677	0.833	0.721
ARI	0.607	0.509	0.386	0.466	0.487	0.661	0.514	0.409	0.507	0.518

E. Validation on a test data set

To evaluate the potential of our procedure, a test set different from the learning data set is used. It is composed of 9L rats ($n = 5$), F98 rats ($n = 5$), and healthy rats ($n = 11$). All voxels of these rats are gathered into data set $\boldsymbol{Y}^T$. After normalization with the normalization values used for the reference model, the pdf $f_H(\boldsymbol{y}_v)$ for each voxel v is computed and values lower than t_{opt} define the abnormal voxels and subset $\boldsymbol{Y}_A^T$ of $\boldsymbol{Y}^T$. Each vector of parameters in $\boldsymbol{Y}_A^T$ is normalized using the normalization values computed for the anomaly model f_A. Individual subject signatures $\boldsymbol{\rho}^S$ are then first extracted as described in Section II-D, eq. (3), for each rat in the test set. Spatial post-processing is then applied as explained in Section II-E to produce individual refined signatures on which are based the final predictions with the refined fingerprint model. All 9L rats are correctly predicted, but one F98 rat and one healthy rat are misclassified (Table III). The misclassified healthy rat, visible in Figure 8-right part, has only a few number of voxels tagged as abnormal (size, first line). Small differences on these

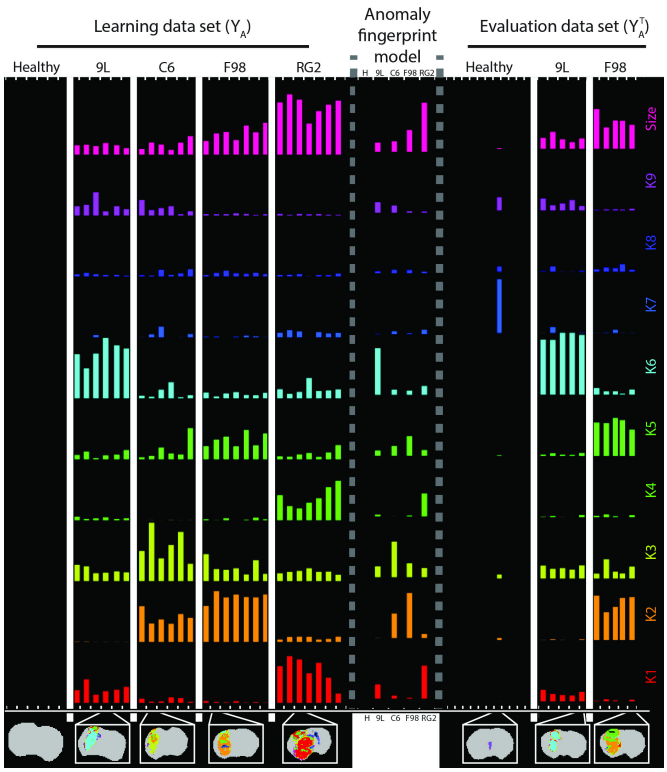


Fig. 8. Refined anomaly signatures by subject. Each signature is represented by a column with 10 rows (bars) indicated on the right hand side by K1 - K9 and Size. In each signature and each row, the bar corresponds to the proportion of voxels in the cluster corresponding to the row for the animal corresponding to the column. Left: individual subject refined signatures associated to the learning data set $\boldsymbol{Y}$ and grouped by tumor type. Center: mean signatures for each tumor type as provided by the refined fingerprint model. Right: individual refined signatures for each subject in the evaluation data set $\boldsymbol{Y}^T$. For each tumor type (9L, C6, F98, RG2) and healthy animals (H), an illustration of the segmentation is presented at the bottom.

few voxels may lead to very different signatures potentially closer to pathological ones. The misclassification of the F98 rat as a C6 rat may be explained by the C6 and F98 fingerprint models similarity: they share the same clusters but with different proportions. The Gaussian model faces the same difficulties to differentiate the C6 and F98 rats, with 3 miss-classifications. In contrast, all the healthy rats are well recovered (Table III). When manual segmentations are used like in [9], the confusion between F98 and C6 tumors is worse with all F98 rats classified as C6 rats both for the Gaussian mixture and MMST models.

The refined segmentations for the evaluation data set test are presented in the Supplementary-Figure 13, with the associated values of ARI (Supplementary-Figure 14a) and DICE (Supplementary-Figure 14b). The average ARI and DICE scores are respectively 0.52 and 0.72.

TABLE III

TUMOR TYPE PREDICTION USING THE REFINED FINGERPRINT MODEL BUILT WITH THE TRAINING DATA SET ($\boldsymbol{Y}$) AND EVALUATED ON AN INDEPENDENT TEST DATA SET ($\boldsymbol{Y}_T$) FOR THE GAUSSIAN MIXTURE AND MMST MODELS BASED ON MANUAL OR AUTOMATIC SEGMENTATIONS.

Manual segmentation & Gaussian mixture or MMST model

	9L	C6	F98	RG2	Healthy
Blind 9L (n=5)	5	.	.	.	.
Blind F98 (n=5)	.	5	.	.	.

Automatic segmentation & Gaussian mixture model

	9L	C6	F98	RG2	Healthy
Blind 9L (n=5)	5	.	.	.	.
Blind F98 (n=5)	.	3	2	.	.
Blind Healthy (n=11)	.	.	.	.	11

Automatic segmentation & MMST model

	9L	C6	F98	RG2	Healthy
Blind 9L (n=5)	5	.	.	.	.
Blind F98 (n=5)	.	1	4	.	.
Blind Healthy (n=11)	.	1	.	.	10

IV. DISCUSSION

QUANTITATIVE MRI allows the comparison of measurements between subjects and with normative values acquired in a healthy population. The monitoring of subtle changes is then possible and represents an important component to design imaging biomarker candidates. In this study, a modular, fully automatic and data-driven procedure to detect and characterize abnormality within medical images was proposed. This procedure was tested on MRI data collected in rats bearing a brain tumor (4 tumor types). The lesions were localized within the brain as anomaly with respect to a reference probabilistic mixture model built using a dataset collected on healthy animals. Abnormal voxels were then clustered in groups with similar MR parameter values. The proportions of each group in each animal were used to construct a signature of each animal. For each tumor type, the signatures for the animals with this tumor type were used to build a fingerprint model of each tumor type. The specificity and sensitivity of the obtained fingerprints were eventually illustrated on a diagnostic task performed successfully without user interaction on an additional test data set. This first application of a procedure whose purpose is more general (any lesion visible by a radiologist, any type of clinical imaging modalities) relied on data from 4 tumor types and characterized by 5 quantitative parameter maps each. Six healthy rats and 26 pathological rats were used to learn the reference and pathology models and 21 additional rats were used for testing. While this dataset is large compared to other preclinical studies [9]–[11] and was sufficient to prove our concept (predictive rate greater than 90% in this particular application), it remains small compare to the volume of data collected daily in patients. Further evaluations are thus required on larger, pre-clinical and clinical data sets to confirm the robustness of the proposed method.

One technical point of interest is that MST distributions performed better than Gaussian distributions: they yielded a better spatial agreement with the manual delineation (6.4% higher ARI on the learning set) and a better predictive rate (5.6% higher on the test set). While the quantitative MRI dataset obtained for tumor models and used in this study is probably not the most challenging one to compare the performance of the two distributions, the MST appears promising for its greater ability to accommodate outliers while maintaining a good separation between clusters [14].

The main evaluation criterion of our procedure was the final diagnostic. Further evaluations of the anomaly detection and of the anomaly model would also be useful to refine the procedure, including the number and type of MR parameter maps used to perform the automated diagnosis. To improve robustness, the results of the intermediate steps could be compared to that of histology. For our specific application (tumors), one could thereby evaluate the ability of MRI to detect abnormal cell densities and differentiate tissue types based on the standard pathologist diagnostic (Louis et al [35]). In addition to improving the quality of the procedure, the results of the intermediate steps may be of interest *per se*. The anomaly detection step, which progressively separates the tumor from the healthy tissue, could help neurosurgeons in planning a tumor resection (Barone, Lawrie and Hart [36]). The anomaly model, which discriminates tissue types, with specific parameter values (*c.f.* Supplementary Table IV), within the tumor, could be of interest to pathologists who cannot evaluate functional parameters such as blood flow. Clustering does not alter the types of information used as input. Thereby, radiologists or pathologists may directly interpret the output tissue characteristics, while such an interpretation is reportedly more difficult with deep learning techniques. Interestingly, the cluster maps show an excellent spatial coherence although no spatial information was used during the clustering steps of the procedure: all voxels were considered independent from each other. Moreover, at the lesion detection step, the ARI score reached 0.52 and the DICE 0.72 on the learning set, and respectively 0.49 and 0.69 on the test set (after post-processing in both cases). These scores were obtained from the comparison between a manual delineation performed on two images (anatomical and diffusion-weighted images as prescribed in standard clinical evaluation) and our procedure which used 5 parameter maps to delineate the tumor. As the images used to perform the delineations differed, it was not surprising to find different lesions between the automatic and the manual delineations. Histology could be used as a more reliable ground truth of high cell proliferation areas than manual delineation (De Angelis [1]). As in a clinical setting, it is generally not possible to obtain histological ground truth data on the entire tumor, this validation step has to be performed at a preclinical level.

The procedure proposed in this study is limited to the detection of tissues whose signal intensity (or parameter values) differs from that of normal tissue. However, a lesion may also appear as a tissue structure with a different volume (*e.g.* Alzheimer patients exhibit a reduction in the cortical thickness compared to age-matched, healthy, subjects [37]). While a large change in structure volume might be detected with our procedure as a change in the proportions of clusters in normal tissue, it would be of interest to add a spatial detection module based on a priori knowledge (*e.g.* atlas [22]). Moreover, a spatial regularization criterion could be added to the proposed procedure to exploit the fact that spatially close pixels have a higher probability to belong to the same tissue type [11], [38]. Exploiting complementarity between spatial proximity and parametric proximity should strongly reinforce our procedure. Once each subject of the training dataset has been labeled (healthy/pathology, lesion type) and the type of MR parameter maps chosen, the proposed procedure requires no human intervention. The determination of the optimal number of clusters, the anomaly threshold, and the final fingerprint model are data-driven to maximize the contrast between the reference data and the lesions and to best discriminate the tumor fingerprints.

In this respect, the proposed procedure differs from that in [9] and is a first attempt to combine both detection and characterization in an all-in-one procedure. It stands as an alternative to texture analysis in the context of radiomics [39]. In the context of multicentre studies, the mixture models at the heart of the procedure could be trained and controlled per center, thereby accounting for inter-center variability.

Finally, as each procedure step is based on a statistical model, quality control tests may readily be introduced (*e.g.* data

homogeneity, outliers) to obtain robust training datasets or to check the data set quality prior to performing an automated diagnostic. The proposed extensions would help meet the challenge of a human application in which the volume of data and number of tissue types represent an exciting challenge for the proposed computer-aided diagnosis procedure. Quantitative MR data are already available for humans [4], [5] but not standard however.

A follow up of this work would be to prove the feasibility and utility on human data with the hope then that quantitative images would be acquired on a more standard basis. To apply our method to more standard non quantitative MRI, one would need to decide on some normalization using for instance ideas from intensity normalization. The algorithm of Nyul et al, 1999 [40] is one of the most popular normalization techniques. Other recent approaches that would require further investigation are described in [41], [42]. Intensity bias correction would be more critically required and more attention should be put on inter-scanner variability [41].

APPENDIX

A. Nested anomaly segmentations

A voxel v is considered as abnormal with regards to the reference model f_H in (1), if it corresponds to parameter values $\boldsymbol{y}_v$ with a low likelihood $f_H(\boldsymbol{y}_v)$. Since pdf values cannot be interpreted as probabilities, the following key step is to decide on a threshold t_{opt} below which voxels will be declared as abnormal, when $f_H(\boldsymbol{y}_v) \leq t_{opt}$. This threshold can be fixed to control the false positive error, *i.e.* when $\boldsymbol{Y}_v$ is distributed according to f_H (healthy tissue), we seek for t_{opt} so that $p(f_H(\boldsymbol{Y}_v) < t_{opt}) = \alpha$ with a small value of α. However, the α value generally chosen (5%) is arbitrary and is likely not to coincide with lesions present in the data set. A threshold specific to the data under consideration is then preferable and can be computed as follows. Likelihood scores are computed for all voxels in $\mathcal{V}_P$, *i.e.* $f_H(\boldsymbol{y}_v)$, for all $v \in \mathcal{V}_P$, but also for the voxels that were used to construct f_H, *i.e.* $f_H(\boldsymbol{y}_v)$ for all $v \in \mathcal{V}_H$. Intuitively, high $f_H(\boldsymbol{y}_v)$ scores corresponding to parameters close to the reference model should separate from the others. To fix the separation in a data driven way, we fit a MMST model to the log-score data set $\{\log f_H(\boldsymbol{y}_v), v \in \mathcal{V}_H \cup \mathcal{V}_P\}$. The slope heuristic is used to set the number L of components in the mixture. Due to the heterogeneity of lesions, L is generally greater than 2 indicating the presence of abnormal tissues with different anomaly levels in the pathological data. Clusters can then be ordered according to their respective mean log-score. The lowest mean corresponds to a group whose departure from the reference model is the highest, while the highest mean should correspond to healthy voxels. The voxels are then partitioned into L groups of successive anomaly levels. Anomaly thresholds are set to the highest likelihood score in each group and denoted by $\{t_1, \ldots, t_L\}$. They are used to provide nested anomaly segmentations that reflect the structure of the lesions (*e.g.* Figure 5). For a global lesion segmentation, another MMST model is fitted to the log-scores with the number of groups set to 2. The two groups are ordered according to their means. For consistency with the previously computed thresholds, we retain as the global threshold t_{opt} the value in the series $\{t_1, \ldots, t_L\}$ which is the closest to the highest score in the first group of the 2-component mixture. The thresholds that correspond to abnormality levels are the ones lower or equal to t_{opt}, that is $\{t_1, \ldots, t_{\text{opt}}\}$. They are associated to colored voxels in Figure 5. For each of them, we compute the false positive probability $\alpha_l = p(f_H(\boldsymbol{Y}_v) < t_l)$ when $\boldsymbol{Y}_v$ is distributed according to f_H. Unfortunately the distribution of $f_H(\boldsymbol{Y}_v)$ is usually not known and the α_l's need to be computed using simulations. However, the thresholds t_l correspond in general to extreme quantiles so that standard empirical estimation would lead to $\alpha_l = 0$ in most cases. For a more precise estimation, we propose to use extreme value theory that enables a more accurate modelling of the distribution tail (see *e.g.* [43]). Details are given in the following Appendix B.

B. Extreme quantile estimation

The computation of $\alpha = p(f_H(\boldsymbol{Y}) \leq t)$ when $\boldsymbol{Y}$ follows f_H is problematic because the pdf of $f_H(\boldsymbol{Y})$ and its theoretical quantiles are not available in general. Empirical estimates are easy to obtain via simulation of *i.i.d.* realizations of $f_H(\boldsymbol{Y})$. However, the t-values of interest are in general extreme quantiles and very few simulated values of $f_H(\boldsymbol{Y})$ will be smaller than t in practice so that α will be estimated to 0. This is a standard issue in extreme quantile estimation that can be addressed via extreme value theory (EVT). EVT focuses on distributions maxima and minima. Considering a set of *i.i.d.* variables $\{Z_{m,n}, m = 1..M, n = 1..N\}$, a set $\{Z_1^*, \ldots, Z_M^*\}$ of M maxima can be defined as $Z_m^* = \max(Z_{m,1}, \ldots, Z_{m,N})$. EVT provides an estimation of the pdf denoted by f of the Z_m^*s via the estimation the generalized extreme value (GEV) distribution. In particular, most EVT procedures provide good estimations of $p(Z_m^* \leq \eta)$ when η is an extreme value in the upper tail of f. Our task can then be recast as follows,

$$\begin{aligned} p(f_H(\boldsymbol{Y}) \leq t) &= 1 - p(f_H(\boldsymbol{Y}) > t) \\ \text{with} \quad p(f_H(\boldsymbol{Y}) > t) &= p(\min(f_H(\boldsymbol{Y}_1), \ldots, f_H(\boldsymbol{Y}_N)) > t)^{1/N} \end{aligned}$$

where $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$ are *i.i.d.* with distribution f_H. Then

$$\begin{aligned} p(\min(f_H(\boldsymbol{Y}_1) \ldots f_H(\boldsymbol{Y}_N)) > t) &= p(\max(-f_H(\boldsymbol{Y}_1) \ldots - f_H(\boldsymbol{Y}_N)) \leq -t) \\ &= p(\max(-\log f_H(\boldsymbol{Y}_1) \ldots - \log f_H(\boldsymbol{Y}_N)) \leq -\log t) \ . \end{aligned}$$

This latter quantity is computed setting $\eta = -\log t$ and $Z_i = -\log f_H(\boldsymbol{Y}_i)$. The log turns the bounded upper tail of $-f_H(\boldsymbol{Y})$ into an unbounded upper tail in $-\log f_H(\boldsymbol{Y})$. This provides more stable estimation of the GEV parameters obtained here with the R package [44].

ACKNOWLEDGEMENTS

The authors acknowledge the excellent technical support of the MRI Facility of Grenoble (UMS IRMaGe). IRMaGe is partly funded by the French program *Investissement d'Avenir* run by the French National Research Agency, grant *Infrastructure d'avenir en Biologie Santé* [ANR-11-INBS-0006].

REFERENCES

[1] L. M. De Angelis, "Brain Tumors," *New England Journal of Medicine*, vol. 344, no. 2, pp. 114–123, January 2001.
[2] A. Drevelegas and N. Papanikolaou, *Imaging of Brain Tumors with Histological Correlations*. Springer Berlin Heidelberg, 2011, ch. Imaging Modalities in Brain Tumors, pp. 13–33.
[3] P. Y. Wen, D. R. Macdonald, D. A. Reardon, T. F. Cloughesy, A. G. Sorensen, E. Galanis, J. DeGroot, W. Wick, M. R. Gilbert, A. B. Lassman, C. Tsien, T. Mikkelsen, E. T. Wong, M. C. Chamberlain, R. Stupp, K. R. Lamborn, M. A. Vogelbaum, M. J. van den Bent, and S. M. Chang, "Updated Response Assessment Criteria for High-Grade Gliomas: Response Assessment in Neuro-Oncology Working Group," *Journal of Clinical Oncology*, vol. 28, no. 11, pp. 1963–1972, 2010.
[4] O. Wu, R. M. Dijkhuizen, and A. G. Sorensen, "Multiparametric magnetic resonance imaging of brain disorders," *Topics in Magnetic Resonance Imaging*, vol. 21, no. 2, pp. 129–138, 2010.
[5] C. Pierpaoli, "Quantitative Brain MRI," *Topics in Magnetic Resonance Imaging*, vol. 21, no. 2, p. 63, 2010.
[6] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, October 2015.
[7] S. J. C. G. Hectors, I. Jacobs, G. J. Strijkers, and K. Nicolay, "Automatic segmentation of subcutaneous mouse tumors by multiparametric MR analysis based on endogenous contrast," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 28, no. 4, pp. 363–375, 2015.
[8] C. Weltens, J. Menten, M. Feron, E. Bellon, P. Demaerel, F. Maes, W. van den Bogaert, and E. van der Schueren, "Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging," *Radiotherapy and Oncology*, vol. 60, no. 1, pp. 49–59, 2001.
[9] N. Coquery, O. François, B. Lemasson, C. Debacker, R. Farion, C. Rémy, and E. L. Barbier, "Microvascular MRI and unsupervised clustering yields histology-resembling images in two rat models of glioma," *Journal of Cerebral Blood Flow & Metabolism*, vol. 34, no. 8, pp. 1354–1362, May 2014.
[10] J. K. Boult, M. Borri, A. Jury, S. Popov, G. Box, L. Perryman, S. A. Eccles, C. Jones, and S. P. Robinson, "Investigating intracranial tumour growth patterns with multiparametric MRI incorporating Gd-DTPA and USPIO-enhanced imaging," *NMR in Biomedicine*, vol. 29, no. 11, pp. 1608–1617, 2016.
[11] P. Katiyar, M. R. Divine, U. Kohlhofer, L. Quintanilla Martinez, B. Schölkopf, B. J. Pichler, and J. A. Disselhorst, "A Novel Unsupervised Segmentation Approach Quantifies Tumor Tissue Populations Using Multiparametric MRI: First Results with Histological Validation," *Molecular Imaging and Biology*, vol. 19, no. 3, pp. 391–397, 2017.
[12] M. Law, S. Yang, J. Babb, E. Knopp, J. Golfinos, D. Zagzag, and G. Johnson, "Comparison of cerebral blood volume and vascular permeability from dynamic susceptibility contrast-enhanced perfusion MR imaging with glioma grade," *American Journal of Neuroradiology*, vol. 25, no. 5, pp. 746–755, 2004.
[13] I. Kosmidis and D. Karlis, "Model-based clustering using copulas with applications," *Statistics and Computing*, vol. 26, no. 5, pp. 1079–1099, 2016.
[14] F. Forbes and D. Wraith, "A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering," *Statistics and Computing*, vol. 24, no. 6, pp. 971–984, 2014.
[15] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions." *Journal of Digital Imaging*, vol. 30, no. 4, pp. 449–459, 2017.
[16] G. Litjens, T. Kooi, B. E. Bejnordi, A. Arindra-Adiyoso-Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
[17] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.
[18] I. J. Goodfellow, J.Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio., "Generative adversarial nets." in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, Montreal, Quebec, Canada, 2014, pp. 2672–2680.
[19] K. van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Transactions on Medical Imaging*, vol. 20, no. 8, pp. 677–688, 2001.
[20] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "EM algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2402–2415, 2016.
[21] J. A. Cuesta-Albertos, C. Matran, and A. Mayo-Iscar, "Robust estimation in the normal mixture model based on robust clustering," *Journal of the Royal Statistical Society Series B*, vol. 70, no. 4, pp. 779–802, 2008.
[22] F. Forbes, S. Doyle, D. Garcia Lorenzo, C. Barillot, and M. Dojat, "A Weighted Multi-Sequence Markov Model For Brain Lesion Segmentation," in *13th International Conference on Artificial Intelligence and Statistics, AISTATS 2010, May, 2012*, ser. JMLR Workshop and Conference Proceedings, Neil Lawrence, Ed., vol. 9, Sardinia, Italie, 2010, pp. 225–232.
[23] P. S. Tofts, *Quantitative MRI of the Brain*. John Wiley & Sons, Ltd, July 2004, ch. Concepts: Measurement and MR, pp. 1–15.
[24] J.-P. Baudry, C. Maugis, and B. Michel, "Slope heuristics: overview and implementation," *Statistics and Computing*, vol. 22, no. 2, pp. 455–470, 2012.
[25] C. Kilkenny, W. J. Browne, I. C. Cuthill, M. Emerson, and D. G. Altman, "Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research," *PLOS Biology*, vol. 8, no. 6, pp. 1–5, June 2010.
[26] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, December 1971.
[27] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
[28] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
[29] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. Springer-Verlag New York, 2002.
[30] L. Bergé, C. Bouveyron, and S. Girard, "HDclassif : An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data," *Journal of Statistical Software*, vol. 46, no. 1, pp. 1–29, 2012.
[31] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca, "mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation," Department of Statistics, University of Washington, Box 354322 Seattle, WA 98195-4322 USA, Tech. Rep. 597, 2012.
[32] B. Lemasson, N. Pannetier, N. Coquery, L. S. Boisserand, N. Collomb, N. Schuff, M. Moseley, G. Zaharchuk, E. L. Barbier, and T. Christen, "MR Vascular Fingerprinting in Stroke and Brain Tumors Models," *Scientific Reports*, vol. 6, p. 37071, Nov 2016.

[33] S. Valable, B. Lemasson, R. Farion, M. Beaumont, C. Segebarth, C. Remy, and E. L. Barbier, "Assessment of blood volume, vessel size, and the expression of angiogenic factors in two rat glioma models: a longitudinal in vivo and ex vivo study," *NMR in Biomedecine*, vol. 21, no. 10, pp. 1043–1056, Nov 2008.
[34] M. Beaumont, B. Lemasson, R. Farion, C. Segebarth, C. Remy, and E. L. Barbier, "Characterization of tumor angiogenesis in rat brain using iron-based vessel size index MRI in combination with gadolinium-based dynamic contrast-enhanced MRI," *Journal of Cerebral Blood Flow & Metabolism*, vol. 29, no. 10, pp. 1714–1726, Oct 2009.
[35] D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella Branger, W. K. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison, "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary," *Acta Neuropathologica*, vol. 131, no. 6, pp. 803–820, 2016.
[36] D. G. Barone, T. A. Lawrie, and M. G. Hart, "Image guided surgery for the resection of brain tumours," *Cochrane Database of Systematic Reviews*, no. 1, 2014.
[37] A.-T. Du, N. Schuff, J. H. Kramer, H. J. Rosen, M. L. Gorno Tempini, K. Rankin, B. L. Miller, and M. W. Weiner, "Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia," *Brain*, vol. 130, no. 4, pp. 1159–1166, March 2007.
[38] J. Juan Albarracín, E. Fuster Garcia, J. V. Manjón, M. Robles, F. Aparici, L. Martí Bonmatí, and J. M. García Gómez, "Automated Glioblastoma Segmentation Based on a Multiparametric Structured Unsupervised Classification," *PLoS ONE*, vol. 10, no. 5, pp. 1–20, May 2015.
[39] R. Gillies, P. Kinahan, and Hedvig Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.
[40] L. G. Nyul and J. Udupa, "On standardizing the MR image intensity scale," *Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1072–1081, 1999.
[41] J. Fortin, E. Sweeney, J. Muschelli, C. Crainiceanu, and R. Shinohara, "Alzheimer's disease neuroimaging initiative. removing inter-subject technical variability in magnetic resonance imaging studies," *Neuroimage*, vol. 132, pp. 198–212, May 2016.
[42] R. Ghassemi, R. Brown, S. Narayanan, B. Banwell, K. Nakamura, and D. Arnold, "Normalization of white matter intensity on T1-weighted images of patients with acquired central nervous system demyelination." *Journal of Neuroimaging*, vol. 25, no. 2, pp. 184–190, 2015.
[43] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events: for Insurance and Finance*, 1st ed., ser. Stochastic Modelling and Applied Probability. Springer-Verlag Berlin Heidelberg, 1997, vol. 33.
[44] A. G. Stephenson, "evd: Extreme Value Distributions," *R News*, vol. 2, no. 2, pp. 31–32, June 2002.

Supplementary Materials for Fully automatic Lesion Localization and Characterization: Application to Brain Tumors Using Multiparametric MRI Data

by Alexis Arnaud, Florence Forbes, Nicolas Coquery, Nora Collomb, Benjamin Lemasson, and Emmanuel L. Barbier

I. Introduction

When analyzing brain tumors, two tasks are intrinsically linked, spatial localization and physiological characterization of the lesioned tissues. Automated data-driven solutions exist, based on image segmentation techniques or physiological parameters analysis, but for each task separately, the other being performed manually or with user tuning operations. In this work, the availability of quantitative magnetic resonance (MR) parameters is combined with advanced multivariate statistical tools to design a fully automated method that jointly performs both localization and characterization.

The statistical approach used is presented in Section II. After a presentation of mixture models, Section II-A, and the associated question of the cluster number choice, Section II-B, probabilistic mixture models of generalized Student distributions are then considered and described Section II-C. These generalized distributions provide a larger variety of distributional shapes compared to the more standard Gaussian distributions (e.g. non-elliptical shapes), and can capture non trivial interactions between relevant physiological parameters to account for the different tissue types, to distinguish anomaly (i.e. lesion), and to explore potential heterogeneity of lesions. The proposed procedure is then summed up in Section II-D.

The potential of this generic procedure is demonstrated on a data set of 32 rats, with 26 rats bearing 4 different brain tumors, for which the 5 acquired quantitative MR are presented Section III-A. The localization of the abnormal voxels and a comparison with Gaussian mixture model results are shown in Sections III-B and III-C. The characterization of the abnormal voxels is provided in Sections III-D and III-E, also followed by a comparison with Gaussian mixture model results. Eventually, the localization and characterization on an independent data set of 21 rats is presented in Section III-F.

II. Clustering with a parametric mixture model

A popular approach for data clustering is the use of a parametric finite mixture model: it relies on the assumption that the data come from several components, or clusters, each one controlled by a parametric distribution. From an observed sample, a parametric finite mixture that best gathers similar observations in clusters is estimated.

A. Parametric mixture model

Let $\boldsymbol{Y} \in \mathcal{R}^{\mathrm{M}}$ be a real random variable in dimension M, $\mathrm{M} \in \mathcal{N}^*$. A parametric mixture model is a statistical model based on a convex combination of $\mathrm{K} \in \mathcal{N}^*$ parametric probability distributions $\{f_1, \ldots, f_{\mathrm{K}}\}$ described by their respective parameters $\theta = \{\theta_1, \ldots, \theta_{\mathrm{K}}\}$. The probability density function p of the mixture is defined by:

$$\forall \boldsymbol{y} \in \mathcal{R}^{\mathrm{M}}, \quad p\left(\boldsymbol{y} \,;\, \theta, \pi\right) = \sum_{k=1}^{\mathrm{K}} \pi_k \, f_k\left(\boldsymbol{y} \,;\, \theta_k\right) \tag{1}$$

where $\pi = \{\pi_1, \ldots, \pi_{\mathrm{K}}\}$ are the mixture proportions: $\forall k \in [\![1; \mathrm{K}]\!]\,,\ 0 < \pi_k < 1$ and $\sum_{k=1}^{\mathrm{K}} \pi_k = 1$.

An equivalent and more convenient representation of a mixture model used an hidden random variable Z which links the random variable $\boldsymbol{Y}$ to one of the distributions $\{f_1, \ldots, f_{\mathrm{K}}\}$ with the following hierarchical representation:

$$\begin{aligned} \left(\boldsymbol{Y} \,|\, Z = k \,;\, \theta_k\right) &\sim f_k \quad \text{for } k \in [\![1 \,;\, \mathrm{K}]\!] \\ \left(Z \,;\, \pi\right) &\sim \mathcal{M}_1\left(\pi_1, \ldots, \pi_{\mathrm{K}}\right) \end{aligned} \tag{2}$$

where $\mathcal{M}_1\left(\pi_1, \ldots, \pi_{\mathrm{K}}\right)$ is the multinomial distribution for one trial, with probability mass function:
$p_{\mathcal{M}_1}\left(Z = k \,;\, \pi_1, \ldots, \pi_{\mathrm{K}}\right) = \pi_k\,,\ k \in [\![1 \,;\, \mathrm{K}]\!]$.

Given a sample of independent realizations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{\mathrm{N}}\}$ from the mixture model, the likelihood of the sample is defined by:

$$p\left(\boldsymbol{y_1}, \ldots, \boldsymbol{y}_{\mathrm{N}} ; \theta, \pi\right) = \prod_{n=1}^{\mathrm{N}} \sum_{k=1}^{\mathrm{K}} \pi_k f_k\left(\boldsymbol{y_n} ; \theta_k\right) \tag{3}$$

The estimation of a mixture model from a sample is generally done with an Expectation-Maximization algorithm (EM). Under some standard assumptions, the EM algorithm jointly estimates the mixture proportions π and parameters θ by reaching a (local) maximum of the log-likelihood:

$$\left(\widehat{\theta}, \widehat{\pi}\right) = \arg\max_{(\theta,\pi)} \log p\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{\mathrm{N}} ; \theta, \pi\right) \tag{4}$$

B. Choice of the cluster number using the slope heuristic

The estimation of a mixture model requires the preliminary choice of the number of clusters K. A too small K may not allow a good fit of the data while a too large K may lead to overfitting issues. We treat this choice as a model selection problem by minimizing a penalized log-likelihood criterion to penalize the complexity of the model. The main issue in this standard approach is the need to tune the penalization term. We use the so-called slope-heuristic (Baudry, Maugis and Michel [1]) to choose this penalization and thus the cluster number in a data-driven way (using the data-driven slope estimation method). Baudry, Maugis and Michel list the validated frameworks for this heuristic, as well as promising empirical results for a wide range of application fields. For a collection of mixture models identified by their cluster numbers between 1 and $K_{\max}$, we extract a subset $S_K = \{K, \ldots, K_{\max}\}$ with $K \in [\![1, \ldots, K_{\max} - 2]\!]$. Each model of S_K is described by its log-likelihood γ_K and its number of free parameters D_K. On this subset, we compute a robust linear regression on the negative log-likelihoods $\{\gamma_K, \ldots, \gamma_{K_{\max}}\}$ with $\{D_K, \ldots, D_{K_{\max}}\}$ as regressors, and estimate the regression slope C_K. We then determine the model $k \in [\![K, \ldots, K_{\max}]\!]$ which minimizes the associated penalized log-likelihood:

$$2C_K D_k - \gamma_k$$

For each subset S_K we get an optimal model k_{opt,S_K}, and we choose the optimal model K_{opt} for the whole set $1, \ldots, K_{\max}$ as the first value such as $\forall K \geq K_{\text{opt}}\ k_{\text{opt},S_K} = K$. This corresponds to the first model from which the log-likelihood is in a linear regime with respect to the number of free parameters. It is also the last model before overfitting.

C. Mixture of multiple scale t-distributions

A standard parametric family for mixture models is the family of Gaussian distributions. In our study, the use of Gaussian distributions is not optimal, because the physiological parameters of interest are observed with large spreads in their statistical distributions [2]. The usual parametric family that can account for heavy tails is the Student distribution family. However, the standard multivariate Student distribution does not allow to adjust different tail thickness to the different dimensions of the data. In this study, we use a mixture of multiple scaled t-distributions (MST), which are a generalization of the multivariate t-distributions [3] that can account for the presence of extreme values in some dimensions. This generalization is based on the expression of the usual t-distribution as an infinite mixture of scaled Gaussian distributions.

There exist quite a few forms of the multivariate t-distribution. Among all the possible multivariate presentations, the most common form is the scale mixture of Gaussians. This leads to the density denoted by $t_M(\boldsymbol{y}; \mu, \boldsymbol{\Sigma}, \nu)$ of the M-dimensional t-distribution with parameters μ (real location vector), $\boldsymbol{\Sigma}$ ($M \times M$ real positive definite scale matrix) and ν (positive real degrees of freedom parameter) given by

$$\begin{aligned} t_M(\boldsymbol{y}; \mu, \boldsymbol{\Sigma}, \nu) &= \int_0^{\infty} \mathcal{N}_M(\boldsymbol{y}; \mu, \boldsymbol{\Sigma}/w)\ \mathcal{G}(w; \nu/2, \nu/2)\ \mathrm{d}w \\ &= \frac{\Gamma((\nu+M)/2)}{|\boldsymbol{\Sigma}|^{1/2}\ \Gamma(\nu/2)\ (\pi\nu)^{M/2}}\ [1 + \delta(\boldsymbol{y}, \mu, \boldsymbol{\Sigma})/\nu]^{-(\nu+M)/2} \end{aligned} \tag{5}$$

where $\delta(\boldsymbol{y}, \mu, \boldsymbol{\Sigma}) = (\boldsymbol{y} - \mu)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \mu)$ is the Mahalanobis distance between $\boldsymbol{y}$ and μ (^{T} means transpose) and Γ is the Gamma function: $\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t)\, dt\ ,\ x \in \mathcal{R}_+^*$. Note that μ is the mean when $\nu > 1$ but $\boldsymbol{\Sigma}$ is not strictly speaking the covariance matrix of the t-distribution which is $\nu/(\nu-2)\boldsymbol{\Sigma}$ when $\nu > 2$.

In most of the work on multivariate scale mixture of Gaussians, the weight variable W has been considered as univariate which results in tails with the same heaviness in all dimensions. The extension we propose consists of introducing the parametrization of the scale matrix parameter into $\boldsymbol{\Sigma} = \boldsymbol{DAD}^T$, where $\boldsymbol{D}$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ is a diagonal matrix with the corresponding eigenvalues of $\boldsymbol{\Sigma}$. The matrix $\boldsymbol{D}$ determines the orientation of the Gaussian and $\boldsymbol{A}$ its shape. The scaled Gaussian part is then set to $\mathcal{N}_M(\boldsymbol{y}; \mu, \boldsymbol{D\Delta_w AD}^T)$, where $\boldsymbol{\Delta_w} = \mathrm{diag}(w_1^{-1}, \ldots, w_M^{-1})$ is the $M \times M$ diagonal matrix whose

diagonal components are the inverse weights $\{w_1^{-1}, \ldots, w_M^{-1}\}$. When the weights are all one, a standard multivariate Gaussian case is recovered. The generalization introduced in [3] is therefore to define:

$$p(\boldsymbol{y}; \mu, \boldsymbol{D}, \boldsymbol{A}, \theta) \quad = \quad \int_0^\infty \ldots \int_0^\infty \mathcal{N}_M(\boldsymbol{y}; \mu, \boldsymbol{D}\boldsymbol{\Delta}_{\boldsymbol{w}}\boldsymbol{A}\boldsymbol{D}^T)\; f_{\boldsymbol{w}}(w_1 \ldots w_M; \theta)\; \mathrm{d}w_1 \ldots dw_M \tag{6}$$

where $f_{\boldsymbol{w}}$ is now a M-variate density function. The weights are independent so that, *i.e.* with $\theta = \{\theta_1, \ldots, \theta_M\}$, $f_{\boldsymbol{w}}(w_1 \ldots w_M; \theta) = f_{W_1}(w_1; \theta_1) \ldots f_{W_M}(w_M; \theta_M)$. We can use then the expression below

$$\mathcal{N}_M(\boldsymbol{y}; \mu, \boldsymbol{D}\boldsymbol{\Delta}_{\boldsymbol{w}}\boldsymbol{A}\boldsymbol{D}^T) \quad = \quad \prod_{m=1}^{M} \mathcal{N}_1([\boldsymbol{D}^T(\boldsymbol{y} - \mu)]_m; 0, A_m w_m^{-1}) \tag{7}$$

where $[\boldsymbol{D}^T(\boldsymbol{y} - \mu)]_m$ denotes the mth component of vector $\boldsymbol{D}^T(\boldsymbol{y} - \mu)$ and A_m the mth diagonal element of the diagonal matrix $\boldsymbol{A}$ (or equivalently the mth eigenvalue of $\boldsymbol{\Sigma}$). Using (7), it follows that

$$p(\boldsymbol{y}; \mu, \boldsymbol{D}, \boldsymbol{A}, \theta) \quad = \quad \prod_{m=1}^{M} \int_0^\infty \mathcal{N}_1([\boldsymbol{D}^T(\boldsymbol{y} - \mu)]_m; 0, A_m w_m^{-1})\; f_{W_m}(w_m; \theta_m)\; \mathrm{d}w_m\ . \tag{8}$$

The terms in the product reduce then to standard univariate scale mixtures and in particular, when setting $f_{W_m}(w_m; \theta_m)$ to a Gamma distribution $\mathcal{G}(w_m; \nu_m/2, \nu_m/2)$, it follows a generalization of the multivariate t-distribution. We can use (8) to express easily the density denoted by $p_{\text{MST}}(\boldsymbol{y}; \mu, \boldsymbol{\Sigma}, \nu)$ with $\nu = \{\nu_1, \ldots, \nu_M\}$:

$$p_{\text{MST}}(\boldsymbol{y}; \mu, \boldsymbol{D}, \boldsymbol{A}, \nu) = \prod_{m=1}^{M} \frac{\Gamma((\nu_m + 1)/2)}{\Gamma(\nu_m/2)(A_m \nu_m \pi)^{1/2}} \left(1 + \frac{[\boldsymbol{D}^T(\boldsymbol{y} - \mu)]_m^2}{A_m \nu_m}\right)^{-(\nu_m+1)/2} \tag{9}$$

We can then rewrite equation (1) as follows to have the density of a mixture of MST (MMST model):

$$\begin{aligned} \forall \boldsymbol{y} \in \mathcal{R}^{\mathrm{M}},\ p_{\text{MMST}}\left(\boldsymbol{y}\ ;\ \psi\right) \quad &= \quad \sum_{k=1}^{\mathrm{K}} \pi_k\, p_{\text{MST}}\left(\boldsymbol{y}\ ;\ \mu_k, \boldsymbol{D}_k, \boldsymbol{A}_k, \nu_k\right) \\ &= \quad \sum_{k=1}^{\mathrm{K}} \pi_k \prod_{m=1}^{\mathrm{M}} \frac{\Gamma\left(\frac{\nu_{k,m}+1}{2}\right)}{\Gamma\left(\frac{\nu_{k,m}}{2}\right)\left([\boldsymbol{A}_k]_{m,m}\,\nu_{k,m}\pi\right)^{\frac{1}{2}}} \left(1 + \frac{[\boldsymbol{D}_k\left(\boldsymbol{y} - \mu_k\right)]_m^2}{[\boldsymbol{A}_k]_{m,m}\,\nu_{k,m}}\right)^{-\frac{\nu_{k,m}+1}{2}} \end{aligned} \tag{10}$$

with: $\psi = \{\psi_1, \ldots, \psi_{\mathrm{K}}\}$ where $\psi_k = (\pi_k, \mu_k, \boldsymbol{D}_k, \boldsymbol{A}_k, \nu_k)$ for $k \in [\![1\ ;\ \mathrm{K}]\!]$.

D. Proposed procedure

The approach presented in this paper is a generic and automatic data-driven procedure to detect and characterize an abnormality within data of interest in comparison to reference data. These two main tasks are performed by a mixture of MST distributions, and will be illustrated on quantitative multiparametric MRI data with the goal to segment and diagnose rat brain tumors. The data have to be continuous quantitative measures, and the learning data have to be made of two sets of voxels: one from healthy subjects and one from pathological subjects for which the pathology (here the tumor) type is known. The procedure consists then of five steps: i) a first mixture model f_H is fitted to the healthy subject voxels; ii) this reference model is used to detect abnormal voxels in the healthy and pathological subjects; iii) a second mixture model f_A is fitted to the detected abnormal voxels and yields a clustering of these voxels into several classes; iv) the proportions of these classes in each subject are used as a signature of the pathology and a discriminative (fingerprint) model is learned that can distinguish between different pathology types; v) an additional spatial post-processing is carried out to remove some spatial artifacts and refine the pathology signatures.

The whole procedure is summarized in Figure 1.

III. Application to brain tumors using multiparametric MRI data

A. Description of the MRI data

a) Data acquisition: The proposed procedure is illustrated on a data set of 53 rats for which 5 quantitative MRI maps are available. MRI was conducted with a horizontal bore 4.7 T Biospec animal imager (Bruker Biospin, Ettlingen, Germany) with an actively decoupled cross-coil setup (body coil for radiofrequency transmission and quadrature surface coil for signal reception) and using Paravision 5.0.1. After second-order shimming, the following MRI protocol was performed. The image acquisition positions were identical for all MRI sequences. 5 slices were acquired with a voxel size of $234 \times 234 \times 800\ \mu m$, unless mentioned otherwise:

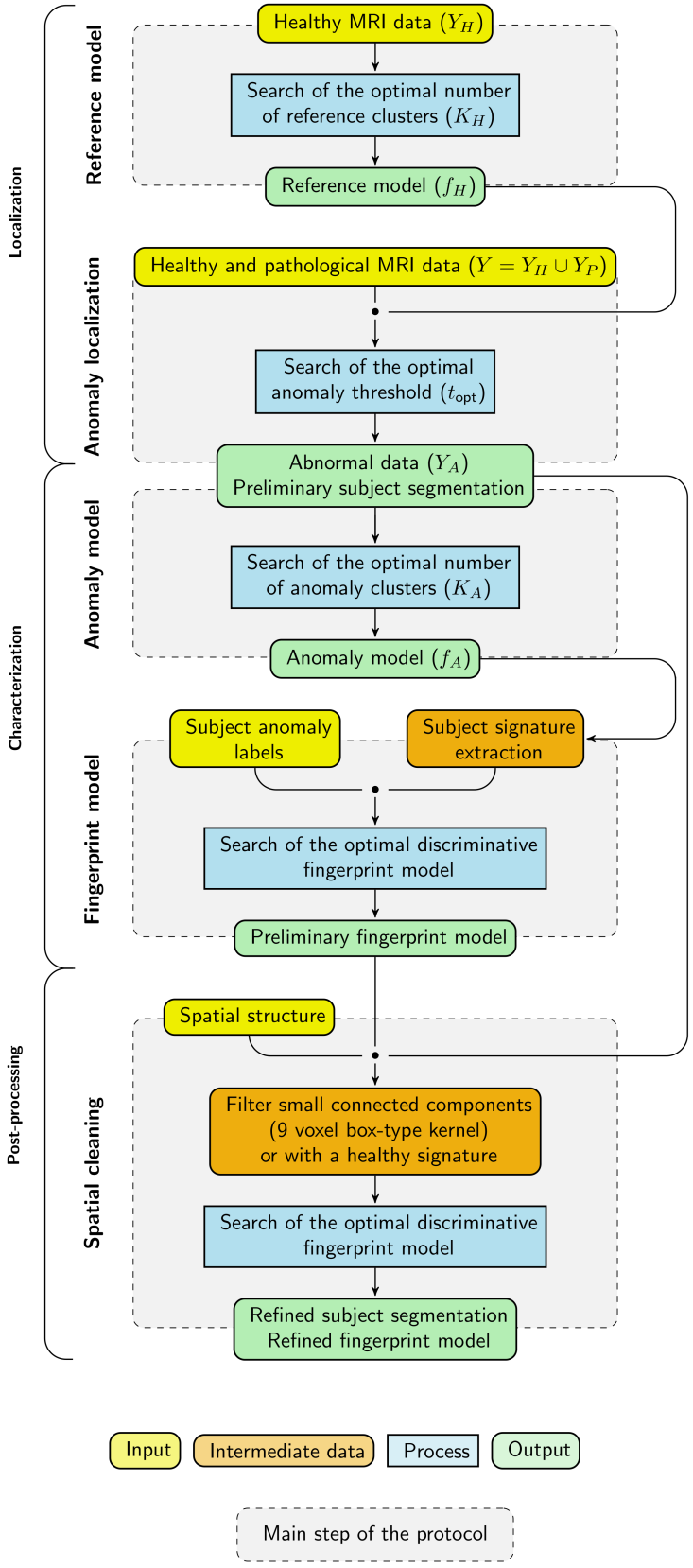


Fig. 1. Construction of a model to automatically localize and characterize lesions. Starting from subjects labeled as healthy or pathological, the procedure is made of 5 main steps.

1) Anatomical T2-weighted (T2w) images were acquired using a spin-echo MRI sequence (repetition time (TR)/effective

echo-time (TE) = 4000/33 ms, acceleration factor = 4, NEX = 2, 31 slices with a field of view (FOV) = $30 \times 30\ mm^2$, acquisition matrix = 256×256 and voxel size = $117 \times 117 \times 800\ \mu m^3$).

2) Cerebral blood flow (CBF) was determined using pseudo-continuous arterial spin labeling (pCASL; spin-echo echo-planar imaging (EPI), TR/TE = 4500/17.2 ms, labeling duration = 4 s, postlabeling delay = 0.2 s, 20 label/control pairs).
3) T1 of brain tissue was determined using an inversion recovery sequence (spin-echo EPI, TR/TE = 10000/20 ms, 25 inversion times: TI = 100-3700 ms).
4) Apparent Diffusion Coefficient (ADC) was mapped using a diffusion-weighted, spin-echo, EPI (TR/TE = 3000/28.6 ms, 8 averages). This sequence was applied 6 times; three without diffusion weighting and three times with diffusion weighting (b = 800 s/mm^2) in three orthogonal directions.
5) T2 (resp. T2*) relaxometric maps were acquired using multiple spin (resp. gradient) echo sequences. Multi spin-echo 2D (MSME) (TR = 2000 ms, 26 spin-echoes, ΔTE = 12 ms). Multi gradient echo 2D (MGE2D) (TR = 2000 ms, 8 gradient-echoes, ΔTE = 4.5 ms).
6) Ultrasmall superparamagnetic iron oxide (USPIO) particles were injected via the tail vein in about 20 sec (200 $\mu mol\ Fe/kg$ body weight; P904®, Guerbet, Roissy, France). 3 min after the injection of USPIOs, a post-contrast T2* relaxometric map was acquired.
7) The vascular wall integrity was assessed using a dynamic contrast-enhanced MRI approach as previously described in [4]. Briefly, 60 T1-weighted, spin-echo, images (TR/TE = 800/4.2 ms, 15.6 s per image) were acquired. After acquisition of 10 baseline images, a bolus of Gd-DOTA (200 $\mu mol/kg$; Guerbet SA, France) was administered through the tail vein and flushed with 250 μL of saline.

b) Data processing for MRI parameter maps: All processing were computed on a voxel basis using custom code developed in the Matlab environment (The MathWorks Inc., Natick, Ma, USA):

- ADC maps were computed as the mean of the ADCs observed in each of three orthogonal directions.
- T1, T2*, and T2 maps were derived using a non-linear fitting algorithm and from the inversion recovery, gradient echoes, and spin-echoes data, respectively.
- The CBF computation was based on the equations described in [5] and assuming a transit delay much shorter than the longitudinal relaxation time of blood.
- Cerebral blood volume (CBV) maps were estimated using the steady-state approach described by [6] and [7]. Changes in transverse relaxation rates due to USPIO (ΔR_2^*) were obtained from the T2* maps collected before and after injection of USPIO. CBV was computed using:

$$\text{CBV} = \frac{3}{4\pi\gamma B_0 \Delta_\chi} \Delta R_2^* \tag{11}$$

  where $\gamma = 2.67502 \cdot 10^8 rad/s/T$, and $B_0 = 4.7T$. Δ_χ, the susceptibility difference between blood in the presence and in the absence of USPIO, was set to 3.5 ppm (SI units).
- The vascular wall integrity was estimated by the area under the signal enhancement curve (AUC) following the Gd-DOTA injection, after baseline removal ([4]).

Voxels were included in the analysis according to the following criteria: $0 <$ ADC $< 4000 \mu m^2/s$, $-2 \cdot 10^5 <$ AUC $< 1.5 \cdot 10^6$a.u., $-50 <$ CBV $< 50\%$, $0 <$ T1 $< 5000ms$ and $0 <$ T2 $< 400ms$. Meeting a single exclusion criterion led to an exclusion of the voxel from all maps. Using these criteria, 0.058% of the voxels were excluded.

B. Healthy subject based reference model

The first steps of our procedure are dedicated to localize all abnormal voxels within the voxel set formed by the pathological subjects (26 rats). This localization is done as an anomaly detection in comparison of a reference model f_H that we fit using a MST mixture on the voxel set from the healthy subjects (6 rats). The slope heuristic selects a number $K_H = 10$ of clusters for the reference model f_H. Figure 2 presents the obtained clustering associated to f_H on a healthy rat, with one color per cluster. Each cluster corresponds to a MST distribution and is described by the distribution parameters $\{\mu_k, \boldsymbol{D}_k, \boldsymbol{A}_k, \nu_k\}$ and the proportion parameter π_k (for the k-th cluster). Figure 3 gives a summary of these clusters through the mean parameter μ_k of each distribution, and the cluster size to illustrate the proportion parameter π_k in terms of voxel number. Each hexagonal web-diagram represents a cluster with, at the center, the lowest MR values of these mean parameters: ADC$= 775 \mu m^2/s$, AUC$= 710.4 a.u.$, CBV$= 1.93\%$, T1$= 1345ms$, T2$= 68.3ms$ and size$= 411$ voxels; and at the border there are the maximal MR values of these mean parameters: ADC$= 1474 \mu m^2/s$, AUC$= 98588 a.u.$, CBV$= 5.34\%$, T1$= 2058ms$, T2$= 126ms$ and size$= 11779$ voxels. The respective parameter values of each cluster are indicated Table I.

Cluster maps appear spatially structured, with a symmetry between the two hemispheres. One can recognize some main anatomical features: the cortex appears in red, the corpus callosum in green, ventricles in purple, blue and dark blue, the striatum is a mixture of orange (main color), red, and light green. Vascular structures, which exhibit a large AUC, appear in blue (cluster 7).

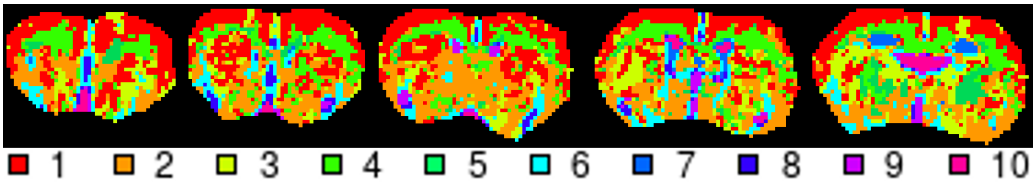
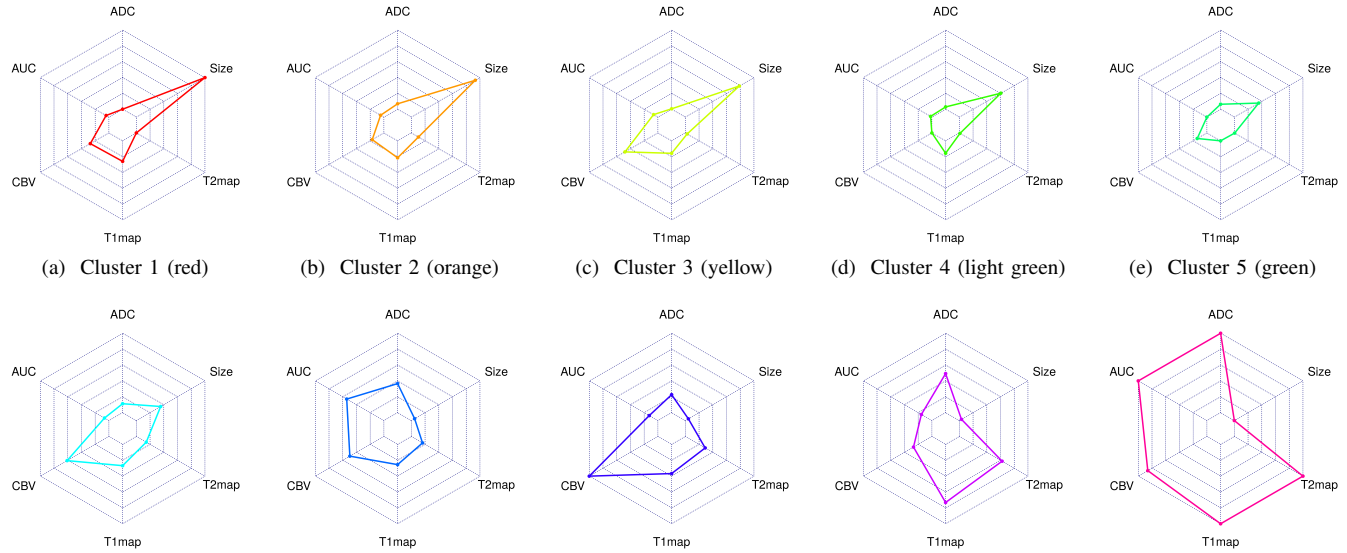


Fig. 2. Clustering using the reference model f_h (MMST model with $K_H = 10$ clusters) on the 5 slices of one healthy rat. From left to right: slices in decreasing order along the vertical axis of the scanner. Each color corrresponds to a cluster.



(a) Cluster 1 (red) (b) Cluster 2 (orange) (c) Cluster 3 (yellow) (d) Cluster 4 (light green) (e) Cluster 5 (green)

(f) Cluster 6 (turquoise) (g) Cluster 7 (blue) (h) Cluster 8 (dark blue) (i) Cluster 9 (purple) (j) Cluster 10 (light purple)

Fig. 3. MR parameter means and size of the reference model clusters (f_H). Range of these parameters between the center and the border of each diagram: ADC$\in [775, 1474]\,\mu m^2/s$, AUC$\in [710.4, 98588]\,a.u.$, CBV$\in [1.93, 5.34]\,\%$, T1$\in [1345, 2058]\,ms$, T2$\in [68.3, 126]\,ms$ and Size$\in [411, 11779]$ voxels.

TABLE I
MEAN AND SIZE PARAMETERS OF THE REFERENCE MODEL CLUSTERS (MMST MODEL f_h WITH $K_H = 10$ CLUSTERS).

Cluster	1	2	3	4	5	6	7	8	9	10
ADC ($\mu m^2/s$)	775	823	779	795	819	854	1031	934	1120	1474
AUC ($a.u.$)	4713	5387	6905	2540	710.4	7011	53751	13040	15723	98588
CBV (%)	2.85	2.51	3.56	1.93	2.41	4.01	3.62	5.34	2.84	4.87
T1 (ms)	1530	1499	1458	1457	1345	1536	1527	1607	1867	2058
T2 (ms)	68.3	74.3	69.8	68.9	68.6	76.7	77.8	85.1	104	126
size (voxels)	11779	10025	7155	6863	3185	3422	816	665	730	411

C. Anomaly localization

The reference model f_H is used to detect the voxels which are abnormal compared to f_H in the sense that they are not well statistically explained by this model. The MR parameter values of the voxels from the pathological and healthy subjects form the data set Y on which we want to detect anomaly by computing the log-likelihood of f_H: $\{\log f_H(\boldsymbol{y}), \boldsymbol{y} \in Y\}$. As described in the main article, Section III-C, we adjust an other mixture model on this data set to identify the voxel subsets sharing similar log-likelihood values, that is the same abnormality with respect to the reference model f_H. This clustering provides L successive abnormal thresholds $\{t_1, \ldots, t_L\}$ and defines nested anomaly segmentations. One of these thresholds is also determined for its ability to best split the data set Y into a healthy part and an abnormal part Y_A. It is called the global threshold t_{opt}.

Figure 4 illustrates these segmentations: the colored voxels are selected as abnormal (most abnormal in red), and the grey ones are considered as healthy (less abnormal in dark grey). A manual delineation (superimposed red line) of the tumor was performed on the anatomical image and the diffusion map. We compare the obtained segmentations using the manual segmentation as ground truth through the associated contingency table (Table II), with the following counts:

- TN (True Negative): the healthy voxels correctly declared as healthy
- FN (False Negative): the tumoral voxels wrongly declared as healthy
- FP (False Positive): the healthy voxels wrongly declared as abnormal

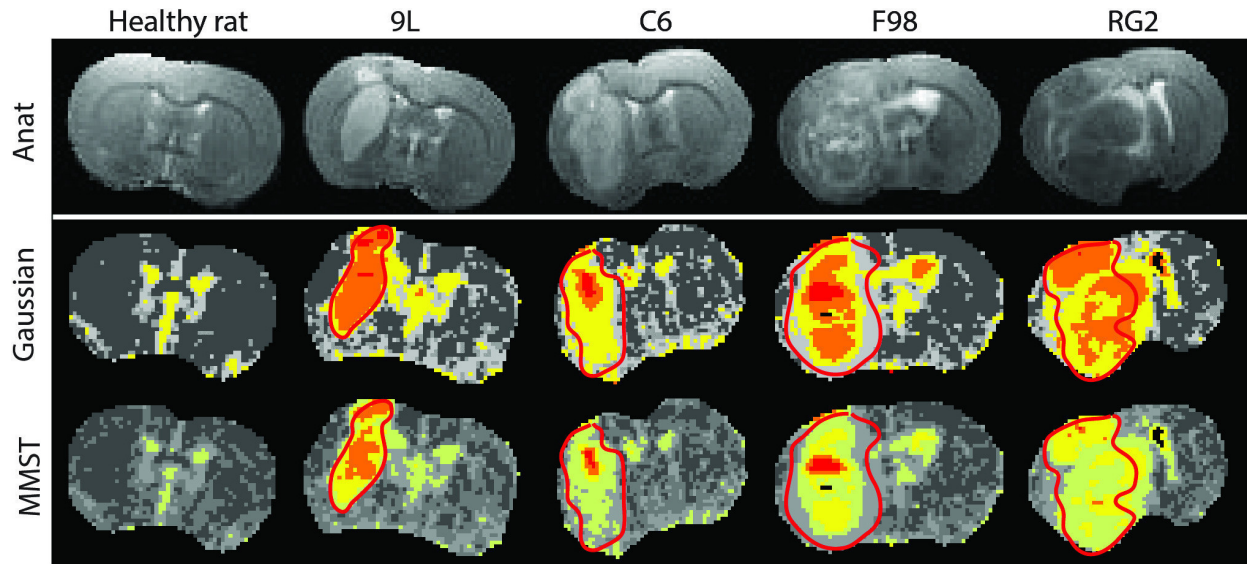


Fig. 4. Nested anomaly segmentations for the different anomaly thresholds with the Gaussian model (center row) and the MMST model (bottom row). The anatomical map (upper row) is displayed as a reference for a visual comparison. The thresholds are ordered from lowest log-likelihood (most abnormal in red) to highest log-likelihood (less abnormal in dark gray) as displayed in the following figures about covering indexes. The colored voxels are those defining the abnormal data set Y_A (i.e. below the global threshold). The gray ones correspond to normal groups (i.e. above the global threshold). The red line superimposed corresponds to the manual segmentation. These segmentations are obtained prior to the spatial post-processing.

- TP (True Positive): the tumoral voxels correctly declared as abnormal

TABLE II
CONTINGENCY TABLE BETWEEN THE AUTOMATIC AND MANUAL SEGMENTATIONS.

		Automatic segmentation		
		Normal	Abnormal	*Total*
Manual segmentation	Healhty	TN	FP	n_H
	Tumor	FN	TP	n_T
	Total	n_{pH}	n_{pT}	n

We use several covering indexes to evaluate the performance of the automatic segmentation:

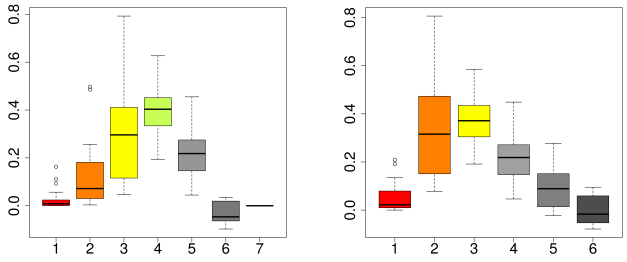
a) Adjusted Rand Index (ARI): is a covering index defined by Equation (12) [8] which measures the agreement between the manual and automatic segmentations for the 2 subsets: healthy and tumor. These index is derived from the rand index [9] which computes the agreement as the number of voxel pairs which are in the same subset in the manual and in the automatic segmentation, plus the number of voxel pairs which are in different subsets in the manual and in the automatic segmentation. The rand index takes values between 0 and 1, the higher the better: at 0 the segmentations agree on no voxel pairs, at 1 they agree on all pairs (perfect concordance). The ARI is a normalization of the rand index such that the ARI value is still lesser than 1 (perfect concordance) but can take negative values when the segmentation is worse than a random clustering under hypergeometric assumptions. The ARI formula is:

$$\text{ARI} = \frac{\binom{\text{TN}}{2} + \binom{\text{FN}}{2} + \binom{\text{FP}}{2} + \binom{\text{TP}}{2} - \left[\binom{n_H}{2} + \binom{n_T}{2}\right]\left[\binom{n_{pH}}{2} + \binom{n_{pT}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\binom{n_H}{2} + \binom{n_T}{2} + \binom{n_{pH}}{2} + \binom{n_{pT}}{2}\right] - \left[\binom{n_H}{2} + \binom{n_T}{2}\right]\left[\binom{n_{pH}}{2} + \binom{n_{pT}}{2}\right] / \binom{n}{2}} \leq 1 \tag{12}$$

Figure 5 shows the ARI values for the obtained nested segmentations. The MMST model detects 7 thresholds (Figure 5a, the boxplots are sorted by abnormality threshold: the higher in red, the lower in dark gray) with a global threshold associated to the 4th threshold. We can observe that the global threshold is also the one with the highest median for the ARI values. If we focus on this threshold, the distribution of the ARI values differs per tumor type (Figure 11a): the 9L and RG2 rats present the best agreement with the manual segmentation, and the F98 rats seem to be the most difficult.

b) DICE index: is a covering index defined by Equation (13) below [10], also known as the F1 score, which measures the similarity between the manual and automatic segmentations of the tumor area. It is a weighted average of the precision and sensitivity for the recovery of the tumor area. The DICE index takes values between 0 and 1, the higher the better: 0 means that all the tumoral voxels are not correctly predicted as healthy; 1 means that the tumoral voxels and only the tumoral voxels are declared as abnormal (perfect concordance).

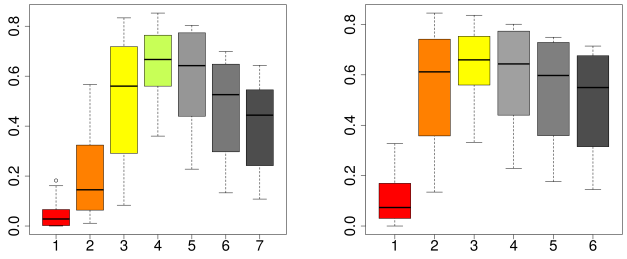
$$\text{DICE} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} = \frac{2\text{TP}}{n_{pT} + n_T} \in [0, 1] \tag{13}$$



(a) MMST: 7 thresholds. (b) GM: 6 thresholds.

Fig. 5. Adjusted rand index (ARI) of the abnormal voxels associated to the successive abnormal thresholds for the MMST model (a) and the GM model (b), without spatial post-processing. For the MMST model (a), the color code is the same as that of Figure 4.

Figure 6 shows the DICE values for the nested segmentations. As for the ARI, the DICE index is also the highest for the 4th of the 7 detected threshold in terms of median values (Figure 6a). Focusing on this threshold, the distribution per tumor type of the DICE values differs from the one observed with the ARI (Figure 12a): the RG2 rats have the best DICE values. We may remark that the ARI and DICE values give different orders for the tumor types. It can be explained by the fact that the DICE index focuses on the true condition (here the tumor area), e.g. the number of true negatives does not matter, whereas the ARI takes into account all the classes of the segmentations. One common point between ARI and DICE values is the variability of the 9L segmentations, which is high in comparison with the others types, while the one of the C6 rats indices variability is low.



(a) MMST: 7 thresholds. (b) GM: 6 thresholds.

Fig. 6. DICE index of the abnormal voxels associated to the successive abnormal thresholds for the MMST model (a) and the GM model (b), without spatial post-processing. For the MMST model (a), the color code is the same as that of Figure 4.

Table III details additional covering indices. The sensitivity ($\frac{TP}{n_T}$) indicates how many tumoral voxels are correctly predicted as abnormal. The specificity ($\frac{TN}{n_H}$) indicates how many healthy voxels are correctly predicted as healthy. The ROC (Receiver operating characteristic) curve is defined by the pairs $\left(\frac{TP}{n_T}, \frac{FP}{n_H}\right)$ (false positive rate against sensitivity), and the area under the ROC curve is equivalent to the probability that the procedure gives a higher abnormal level to a tumoral voxel than to a healthy voxel. It takes values between 0 and 1; a value of 0.5 is associated to a random allocation.

Table III shows that the ability to detect abnormality is higher for MMST than GM: the sensitivity is higher (more abnormal voxels are correctly declared as abnormal) with a smaller specificity (more healthy voxels are wrongly declared as abnormal). But globally, the increase of true positive compensates the false positive: the area under the ROC curve is equivalent, the DICE index and the ARI are slightly better. We can notice that the MMST model detects 7 thresholds instead of 6 for the GM model. It thus allows a finer segmentation of the abnormal voxels: the variability of the 2nd GM threshold is equal to the variability of the combined thresholds 2 and 3 of the MMST model (Figures 5b and 6b). For the similarities, the chosen global threshold is the one with the highest median values of ARI and DICE for both the MMST and GM models. When focusing on this threshold, the order of tumor type in terms of ARI and DICE median values is the same for the MMST and GM models, except for the tumor type order based on the DICE values, for which the ranks of 9L and C6 rats are switched (Figures 11c and 12c).

D. Anomaly model

The last steps of our procedure are dedicated to characterize the pathologies using the voxels declared as abnormal. We first adjust a MMST model on the abnormal voxels Y_A, in order to best decribe the data using a mixture model before using it to compute the subject fingerprints. The anomaly model f_A is composed of $K_A = 10$ clusters, according to the slope heuristic.

TABLE III
PATHOLOGICAL RATS IN THE TRAINING SET: MEAN COVERING SCORES FOR THE LESION AUTOMATIC SEGMENTATIONS FOR THE GAUSSIAN MIXTURE AND MMST MODELS, WITHOUT SPATIAL POST-PROCESSING.

	Gaussian model					MMST model				
	9L	C6	F98	RG2	All	9L	C6	F98	RG2	All
sensitivity	0.830	0.560	0.553	0.853	0.699	0.862	0.641	0.636	0.880	0.755
specificity	0.922	0.926	0.912	0.829	0.895	0.884	0.884	0.871	0.779	0.852
area under ROC curve	0.928	0.855	0.867	0.894	0.883	0.928	0.852	0.865	0.897	0.883
DICE index	0.510	0.582	0.671	0.801	0.648	0.581	0.582	0.639	0.819	0.661
ARI	0.400	0.372	0.332	0.421	0.381	0.489	0.402	0.322	0.467	0.418

Figure 7 presents the clustering associated to f_A on the pathological rats from the training set, with one color per cluster. Note that the clusters now differ from that obtained from the Y_H data set.

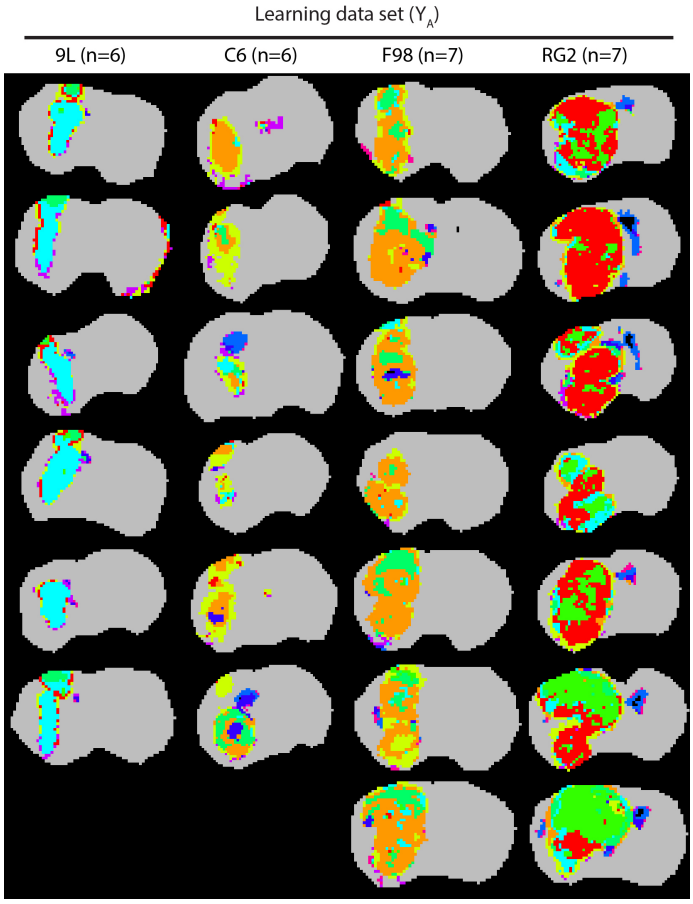
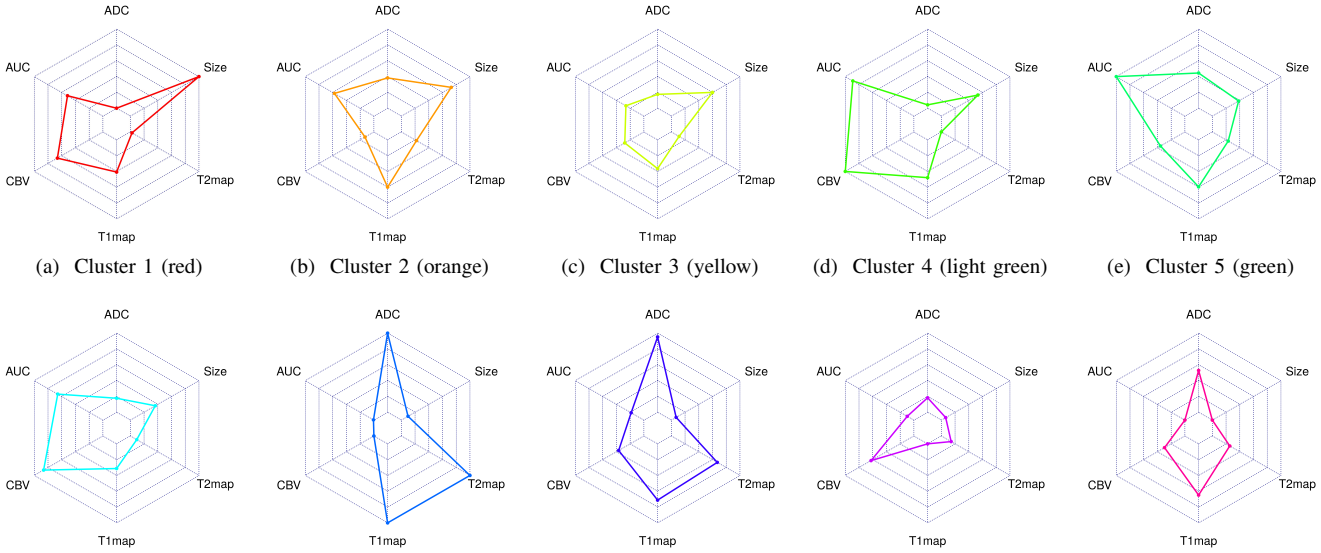


Fig. 7. Anomaly clustering of the pathological rats from the training set Y_A using the anomaly model f_A (MST mixture with $K_A = 10$ clusters) and including spatial post-processing. Each column gathers the rats bearing the same pathology (tumor 9L, C6, F98, or RG2); the central slice is the only one displayed per rat and contains the biggest tumoral area. Note that the color code is different from that of Figure 2. The clusters associated to Y_A are described Figure 8 and Table IV.

As for the reference model, each cluster corresponds to a MST distribution and is described by the distribution parameters $\{\mu_l, \boldsymbol{D}_l, \boldsymbol{A}_l, \nu_l\}$ and the proportion parameter π_l (for the l-th cluster). Figure 8 gives a summary of these clusters through the mean parameter μ_l of each distribution, and the cluster size to illustrate the proportion parameter π_l in terms of voxel numbers. Each hexagonal web-diagram represents a cluster with, at the center, the lowest MR values of these mean parameters: ADC$= 744\mu m^2/s$, AUC$= 12188 a.u.$, CBV$= 1.31\%$, T1$= 1484ms$, T2$= 74.3ms$ and size$= 2814$ voxels; and at the border there are the maximal MR values of these parameters: ADC$= 1668\mu m^2/s$, AUC$= 474460 a.u.$, CBV$= 9.21\%$, T1$= 2438ms$, T2$= 179ms$ and size$= 11612$ voxels. The mean parameters associated to each cluster are indicated Table IV.

Clusters represent different types of tissues. For example, clusters 7 and 8 show high ADC with limited permeability and could correspond to ventricles/plexus choroid. Cluster 1 shows a normal ADC, a CBV 50% above normal brain values and a high permeability. The highest permeability (cluster 5, green) is associated with an ADC between that of normal tissue and that of ventricles and with a normal CBV.



(a) Cluster 1 (red) (b) Cluster 2 (orange) (c) Cluster 3 (yellow) (d) Cluster 4 (light green) (e) Cluster 5 (green)

(f) Cluster 6 (turquoise) (g) Cluster 7 (blue) (h) Cluster 8 (dark blue) (i) Cluster 9 (purple) (j) Cluster 10 (light purpple)

Fig. 8. MR parameter means and size of the anomaly model clusters. Range of these parameters: ADC$\in [744, 1668]\,\mu m^2/s$, AUC$\in [12188, 474460]\,a.u.$, CBV$\in [1.31, 9.21]\,\%$, T1$\in [1484, 2438]\,ms$, T2$\in [74, 179]\,ms$ and Size$\in [11284, 49732]$ voxels. The color code is similar to that of Figure 7.

TABLE IV
MEAN AND SIZE PARAMETERS OF THE ANOMALY MODEL CLUSTERS (MMST MODEL f_A WITH $K_A = 10$ CLUSTERS).

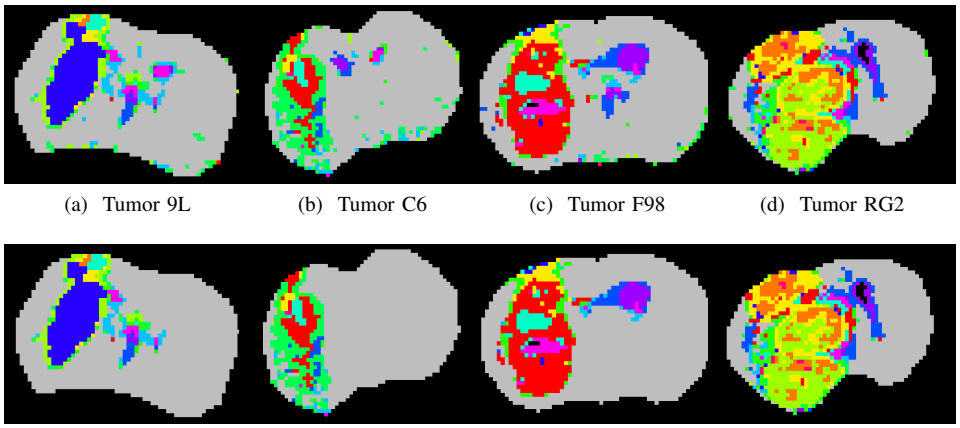
	1	2	3	4	5	6	7	8	9	10
ADC ($\mu m^2/s$)	744	1098	906	783	1156	908	1668	1624	913	1232
AUC ($a.u.$)	250599	277915	133332	423702	474460	316320	15018	98009	56680	12188
CBV (%)	6.54	2.33	3.51	9.21	4.12	8.13	1.31	4.21	6.24	3.65
T1 (ms)	1872	2055	1832	1941	2051	1782	2438	2163	1484	2103
T2 (ms)	76.8	97.4	86.2	74.3	98.6	84.1	179	145	89.3	101
size (voxels)	11612	9015	8474	6724	4617	5159	3274	2756	3102	2814

E. Post-processing

We present here the impact of the spatial post-processing on the tumor localization.

Figure 9 presents the anomaly clustering using a Gaussian mixture model based on an automatic localization of the tumor (a variant of the proposed procedure in terms of probability distributions). For this Gaussian mixture also, the slope heuristic selected $K_A = 13$ clusters. As described in Section III-C, the automatic segmentation wrongly declares some voxels from the contralateral part as abnormal. The spatial cleaning allows to remove the isolated voxels and the ones at the border of the skull, but the abnormal voxels in the ventricles often form a subset touching the tumor area. These voxels cannot be removed by our spatial cleaning, which leads to an overestimation of the abnormal part.

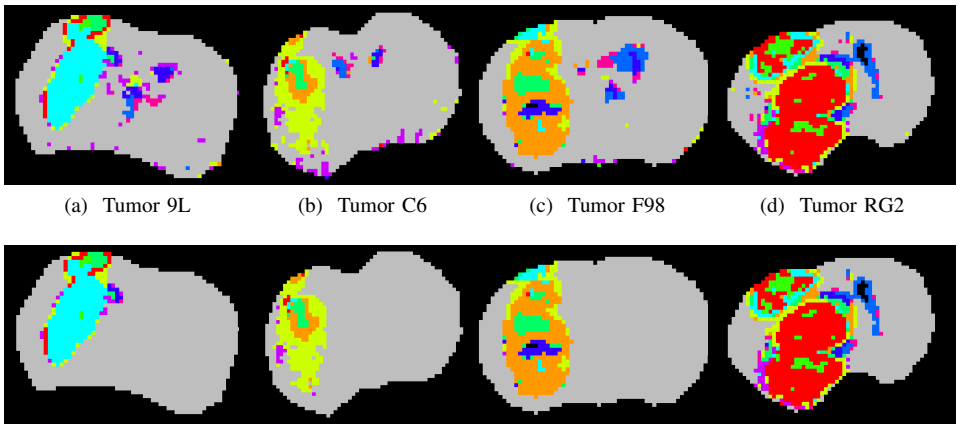
Figure 10 presents the anomaly clustering using a MMST model based on an automatic localization of the tumor (the proposed procedure). For the MMST model, the slope heuristic selected $K_A = 10$ clusters. With this model, the abnormal contralateral area is smaller than the one with the Gaussian mixture, which leads to less contacts with the tumor area, and then a better spatial cleaning, as we can observe for the 9L and F98 tumor types.



(a) Tumor 9L (b) Tumor C6 (c) Tumor F98 (d) Tumor RG2

(e) Tumor 9L (f) Tumor C6 (g) Tumor F98 (h) Tumor RG2

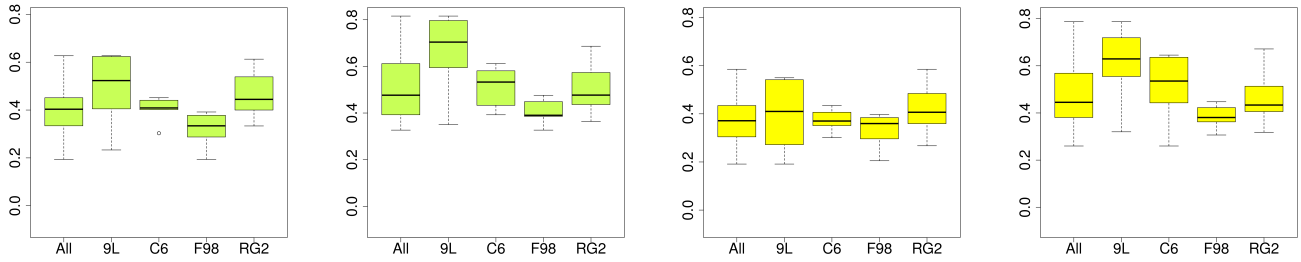
Fig. 9. Gaussian anomaly clustering ($K_A = 13$ clusters): before (upper row) and after (bottom row) the spatial cleaning by removing too small components or components with a healthy fingerprint.



(a) Tumor 9L (b) Tumor C6 (c) Tumor F98 (d) Tumor RG2

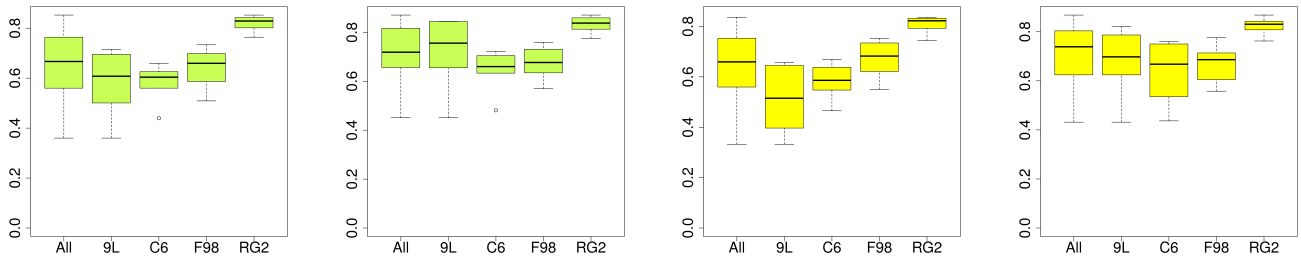
(e) Tumor 9L (f) Tumor C6 (g) Tumor F98 (h) Tumor RG2

Fig. 10. MMST anomaly clustering ($K_A = 10$ clusters): before (upper row) and after (bottom row) the spatial cleaning by removing too small components or components with a healthy fingerprint.



(a) MMST - before (b) MMST - after (c) GM - before (d) GM - after

Fig. 11. ARI per tumor type for the segmentation associated to the global threshold, before (MMST: a, GM: c) and after (MMST: b, GM: d) the spatial post-processing.



(a) MMST - before (b) MMST - after (c) GM - before (d) GM - after

Fig. 12. DICE index per tumor type for the segmentation associated to the global threshold, before (MMST: a, GM: c) and after (MMST: b, GM: d) the spatial post-processing.

TABLE V

PATHOLOGICAL RATS IN THE TRAINING SET: MEAN COVERING SCORES FOR THE LESION AUTOMATIC SEGMENTATIONS FOR THE GAUSSIAN MIXTURE AND MMST MODELS, AFTER SPATIAL POST-PROCESSING.

	Gaussian model					MMST model				
	9L	C6	F98	RG2	All	9L	C6	F98	RG2	All
sensitivity	0.856	0.530	0.541	0.879	0.702	0.827	0.519	0.538	0.851	0.685
specificity	0.949	0.980	0.960	0.809	0.921	0.964	0.984	0.977	0.856	0.943
area under ROC curve	0.916	0.794	0.787	0.882	0.840	0.935	0.867	0.885	0.902	0.894
DICE index	0.677	0.636	0.665	0.823	0.704	0.718	0.644	0.677	0.833	0.721
ARI	0.607	0.509	0.386	0.466	0.487	0.661	0.514	0.409	0.507	0.518

F. Evaluation on an independent test data set

We test our procedure on an independent data set composed of 9L rats ($n = 5$), F98 rats ($n = 5$), and healthy rats ($n = 11$), as described in the manuscript Section III-E. All the results concerning the fingerprint model are presented in the manuscript, Section III-D. We only present here the automatic segmentation after spatial post-processing, Figure 13, for the pathological rats. The healthy test rats present as desired a segmentation with no abnormal voxels, except for the one already shown in the manuscript, Section III-E. For these rats, average ARI and DICE indexe are 0.49 and 0.69 after spatial post-processing, respectively, values comparable to that obtained during training. These data demonstrate the ability of the proposed procedure to automatically segment tumors from an unseen data set.

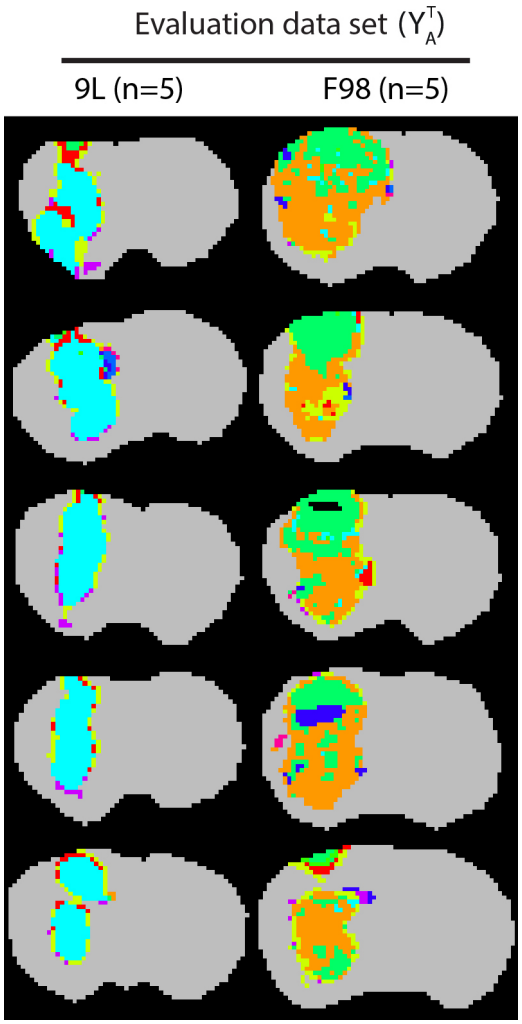
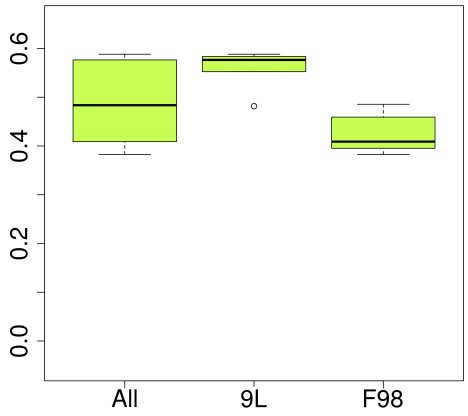
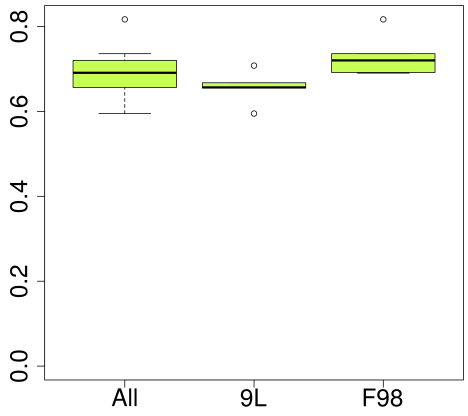


Fig. 13. Anomaly clustering of the pathological rats from the test set Y_A^T using the anomaly model f_A (MST mixture with $K_A = 10$ clusters) and the refined ROI. Each column gathers the rats bearing the same pathology (tumor 9L, or F98); the central slice is the only one displayed per rat and contains the biggest tumor area. The color code is similar to that of Figures 7 and 8.



(a) Adjusted rand index.



(b) DICE index.

Fig. 14. Adjusted rand index (a) and DICE index (b) of the refined segmentation for the test data set Y_A^T, using the reference model and the refined fingerprint model. The results are obtained under the MMST model.

REFERENCES

[1] J.-P. Baudry, C. Maugis, and B. Michel, "Slope heuristics: overview and implementation," *Statistics and Computing*, vol. 22, no. 2, pp. 455–470, 2012.
[2] N. Coquery, O. François, B. Lemasson, C. Debacker, R. Farion, C. Rémy, and E. L. Barbier, "Microvascular MRI and unsupervised clustering yields histology-resembling images in two rat models of glioma," *Journal of Cerebral Blood Flow & Metabolism*, vol. 34, no. 8, pp. 1354–1362, May 2014.
[3] F. Forbes and D. Wraith, "A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering," *Statistics and Computing*, vol. 24, no. 6, pp. 971–984, 2014.
[4] B. Lemasson, R. Serduc, C. Maisin, A. Bouchet, N. Coquery, P. Robert, G. Le Duc, I. Troprès, C. Rémy, and E. L. Barbier, "Monitoring Blood-Brain Barrier Status in a Rat Model of Glioma Receiving Therapy: Dual Injection of Low-Molecular-Weight and Macromolecular MR Contrast Media," *Radiology*, vol. 257, no. 2, pp. 342–352, 2010.
[5] C. S. Debacker, A. Daoust, S. Köhler, J. Voiron, J. M. Warnking, and E. L. Barbier, "Impact of tissue T1 on perfusion measurement with arterial spin labeling," *Magnetic Resonance in Medicine*, vol. 77, no. 4, pp. 1656–1664, 2017.

[6] I. Troprès, S. Grimault, A. Vaeth, E. Grillon, C. Julien, J.-F. Payen, L. Lamalle, and M. Décorps, "Vessel size imaging," *Magnetic Resonance in Medicine*, vol. 45, no. 3, pp. 397–408, 2001.
[7] S. Valable, B. Lemasson, R. Farion, M. Beaumont, C. Segebarth, C. Rémy, and E. L. Barbier, "Assessment of blood volume, vessel size, and the expression of angiogenic factors in two rat glioma models: a longitudinal in vivo and ex vivo study," *NMR in Biomedicine*, vol. 21, no. 10, pp. 1043–1056, 2008.
[8] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
[9] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, December 1971.
[10] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

CHAPITRE 4

SÉLECTION DE MODÈLES DE MÉLANGE DE LOIS DE PEARSON TYPE VII À MÉLANGE D'ÉCHELLES MULTIPLES DANS UN CADRE BAYÉSIEN

Dans ce chapitre, nous présentons nos travaux sur la sélection du nombre de classes pour les modèles de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP). Les sections 4.1 et 4.2 résument les principales approches possibles ainsi que l'heuristique développée au cours de cette thèse. Les détails sont fournis dans un article qui sera soumis sous peu section 4.3.

4.1 Sélection de modèle dans le cadre de modèles de mélange

Au cœur du protocole présenté au chapitre 3 se trouve les modèles de mélange : que ce soit pour l'apprentissage des caractéristiques des paramètres IRM pour les patients sains ou pathologiques, la détection des voxels anormaux ou encore l'élaboration d'une signature pathologique. Il s'agit d'un modèle d'apprentissage statistique largement utilisé en classification qui permet une bonne adaptation aux données tout en ayant une interprétation simple. Toutefois, ces modèles reposent sur deux choix : i) les lois de probabilité utilisées et ii) le nombre de classes au sein du mélange.

Classiquement, les lois gaussiennes sont utilisées pour leur facilité de maniement dans les calculs d'estimation. Toutefois, ces lois sont particulièrement sensibles aux valeurs atypiques, ce qui peut entraîner une détérioration de la classification obtenue par l'augmentation du nombre de classes à considérer pour décrire cor-

rectement les données, ou bien par une erreur d'estimation des paramètres. Les lois de Student sont au contraire adaptées à la modélisation de données ayant des valeurs atypiques car permettant précisément de prendre en compte le niveau d'atypie de chaque observation pour l'estimation des paramètres du modèle. La généralisation considérée (Forbes et Wraith [7]) permet en outre de considérer cette atypie pour chaque dimension dans le cas de données multidimensionnelles, tout en conservant des formules explicites pour l'estimation du mélange. Cette propriété est ici particulièrement intéressante du fait que dans le cas de lésions, telles que les tumeurs, il est souhaitable de pouvoir détecter une anormalité n'apparaissant que suivant un des paramètres IRM acquis, ce qui correspond à l'approche des radiologues pour déterminer si une zone est ou non anormale.

Le deuxième point clé des modèles de mélange est le nombre de classes au sein du mélange. Le choix de ce nombre est d'ailleurs un problème intrinsèque aux modèles de mélange. La vraisemblance du modèle augmentant avec le nombre de classes, sans contrainte le meilleur modèle est celui où chaque observation forme une unique classe. Une manière classique de résoudre ce problème est de considérer le choix du nombre de classes comme un problème de choix de modèles. De nombreuses solutions à ce problème ont été proposées en utilisant une pénalisation de la vraisemblance du modèle (pénalisation telle que AIC, BIC etc.), de façon à choisir le modèle le plus parcimonieux au sens de celui maximisant ou minimisant un critère pénalisé. Ainsi, lorsque le nombre de classes choisi est trop grand, la pénalisation l'emporte sur l'information apportée par les données, ce qui permet d'éviter de tomber dans le sur-ajustement. Cette approche présente toutefois un inconvénient très limitant en pratique : afin de comparer plusieurs modèles, il est nécessaire de tous les ajuster, avant de pouvoir en déduire le nombre de classes à garder. Ces apprentissages sont particulièrement coûteux en temps de calcul lorsque le modèle de mélange est difficile à estimer, ce qui est le cas pour les lois de Student à mélange d'échelles multiples.

Un des objectifs principaux de cette thèse a donc été de considérer des méthodes d'estimation conjointe du nombre de classes et du mélange. Dans cette veine, il est possible de distinguer principalement deux catégories de méthodes : les méthodes avec un nombre infini de classes, et celles avec un nombre fini. Le premier cas regroupe par exemple les modèles de mélange infini et les processus de Dirichlet, mais ne sont pas étudiés ici car il faudrait pouvoir interpréter d'un point de vue physiologique l'ensemble infini des classes associées à ces modèles. Le second cas contient les approches entièrement bayésiennes et celles semi-bayésiennes. Les modèles entièrement bayésiens correspondent aux modèles pour lesquels tous les paramètres du modèle deviennent des variables aléatoires associés à une loi a priori, y compris le nombre de classes du mélange. De ce fait, ces méthodes sont ici également écartées car elles ne visent pas à estimer le nombre de classes mais

une loi sous-jacente, sauf dans le cas des méthodes de Monte-Carlo par chaînes de Markov à saut réversible, mais ces méthodes sont aussi très coûteuses en temps de calcul. Concernant les méthodes semi-bayésiennes, uniquement une partie des paramètres du modèle deviennent des variables aléatoires, et en particulier le nombre de classes reste un paramètre à estimer. Dans ce dernier cas, si les hypothèses de Rousseau et Mengersen ([24]) sont vérifiées, alors lorsque l'algorithme EM utilisé pour l'ajustement du mélange contient un plus grand nombre de classes que le vrai nombre de classes, l'algorithme EM vide les classes surnuméraires à la convergence. En pratique, il est possible d'utiliser cette propriété pour supprimer des classes au fur et à mesure plutôt qu'à la convergence de l'algorithme. Ainsi, une classe est enlevée lorsqu'un certain critère est suffisamment faible : ce critère pouvant être la proportion de la classe (Corduneanu et Bishop [25]) ou un critère de longueur de message (Law, Figueiredo and Jain [26]). C'est ce que nous pouvons observer Figure 4.1 où le nombre de classes réduit au cours des itérations de l'algorithme. C'est dans cette catégorie des méthodes semi-bayésiennes que se situent nos travaux de sélection de modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP).

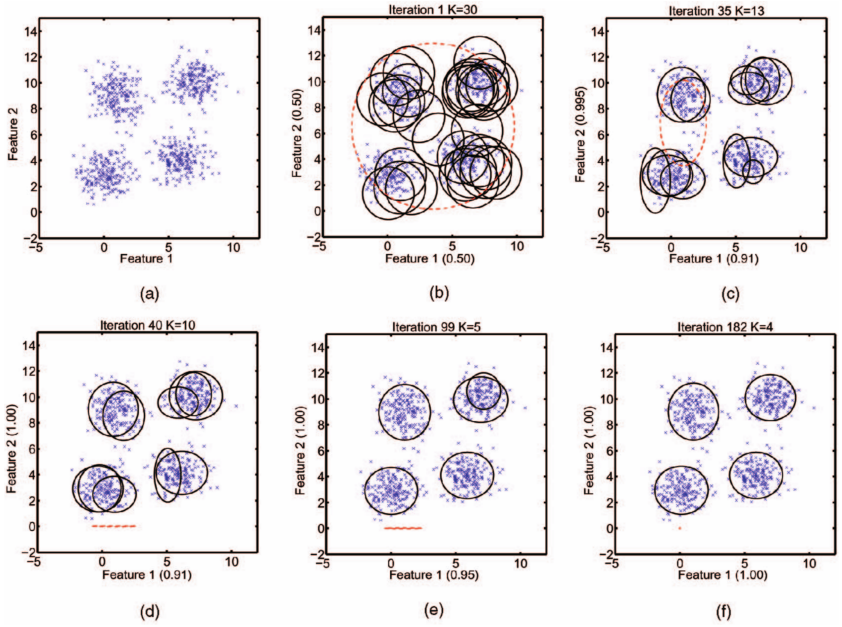


FIGURE 4.1 – Suppression des classes surnuméraires au cours des itérations de l'algorithme EM. Figure extraite de Law, Figueiredo and Jain 2004.

4.2 Heuristique pour la sélection de modèles bayésiens de mélange de lois de Pearson type VII à mélange d'échelles multiples

Nous avons réalisé deux formulations bayésiennes du modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP - deuxième partie de l'article section 4.3). La première ne contient pas de loi a priori sur les proportions de classes, ce qui revient à réaliser une maximisation de type II de la vraisemblance (Corduneanu et Bishop [25]). La seconde suppose une loi a priori de type Dirichlet sur les proportions du mélange, avec des paramètres de faibles valeurs afin d'avoir une représentation parcimonieuse du mélange (Malsiner-Walli et al. [27]). Les paramètres du modèle qui ne présentent pas de lois a priori conjuguée simples dans ce contexte restent en tant que paramètres à estimer. C'est le cas des degrés de liberté et des matrices orthogonales intervenant dans les lois de Pearson type VII à mélange d'échelles multiples. Ainsi, les seuls variables bénéficiant d'un a priori bayésien sont les vecteurs de positions (loi Normale), et les valeurs propres de la matrice d'échelle (loi Gamma). Les lois a priori considérées ici sont des lois classiquement utilisées en modèle de mélange. Le modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples est détaillé dans la première partie de l'article section 4.3.

L'estimation du modèle n'est plus directement faisable avec l'algorithme EM usuel, à cause du calcul des lois a posteriori. Il serait possible d'utiliser une méthode de type Monte-Carlo par chaînes de Markov, mais la convergence d'un tel algorithme est difficile à contrôler et le temps d'exécution souvent long. Au lieu de simuler la loi a posteriori, nous approchons donc plutôt cette dernière par une approximation variationnelle qui permet d'utiliser l'algorithme EM pour maximiser directement une borne inférieure de la vraisemblance, l'énergie libre du modèle (troisième partie de l'article section 4.3).

En pratique la sélection de modèle se fait en supprimant les classes dont la proportion devient faible, par exemple $\frac{1}{N}$ avec N la taille de l'échantillon. Cependant ce critère n'est atteint que lorsque l'algorithme a suffisamment convergé et dans le cas d'un mélange de lois complexes, ces étapes avant la suppression d'une classe peuvent représenter un temps non négligeable. Pour accélérer la convergence de l'algorithme, nous proposons également une heuristique sous la forme d'un test basé sur l'énergie libre du modèle (quatrième partie de l'article section 4.3). Ainsi à chaque itération de l'algorithme EM, en plus d'estimer le modèle avec toutes les classes, nous testons aussi si la suppression d'une classe permet d'accroître l'énergie libre du modèle, et finalement nous retenons le modèle de plus grande énergie

libre. Une manière d'interpréter cette heuristique est de voir la suppression d'une classe comme un changement de l'initialisation de EM ; ce qui met en évidence la nécessité de démarrer l'algorithme avec suffisamment de classes pour couvrir tout l'espace présentant des observations.

Résultats

Nous avons testé l'ensemble des heuristiques développées, à la fois sur données simulées issues d'un mélange de lois de Pearson type VII à mélange d'échelles multiples (équivalent à un cas traité par [25]) et sur données réelles (Old Faithful, un jeu de données couramment utilisé en sélection de modèle). Les premiers résultats montrent de meilleurs performances que la sélection de modèle avec le BIC (cinquième partie de l'article section 4.3). Le nombre de classes sélectionné est plus robuste, et les temps de calcul sont réduits d'un facteur 3 à 5 suivant le jeu de données.

4.3 Article à soumettre

Noname manuscript No.
(will be inserted by the editor)

Bayesian mixtures of multiple scale distributions

Alexis Arnaud · Florence Forbes · Russel Steele · Benjamin Lemasson · Emmanuel Barbier

Received: date / Accepted: date

Abstract Multiple scale distributions are multivariate distributions that exhibit a variety of shapes not necessarily elliptical while remaining analytical and tractable. In this work we consider mixtures of such distributions for their ability to handle non standard typically non-gaussian clustering tasks. We propose a Bayesian formulation of the mixtures and a tractable inference procedure based on variational approximation. The interest of such a Bayesian formulation is illustrated on an important mixture model selection task, which is the issue of selecting automatically the number of components. We derive promising procedures that can be carried out in a single-run, in contrast to the more costly comparison of information criteria. Preliminary results on simulated and real data show promising performance in terms of clustering and computation time.

Keywords Covariance matrix decomposition · EM algorithm · Gaussian scale mixture · Bayesian analysis · Bayesian model selection · Variational approximation

1 Introduction

Multiple scale distributions refer to a recent generalization of scale mixtures of Gaussians in a multivariate setting [Forbes and Wraith, 2014]. This new family of distributions has the ability to generate a number of flexible distributional forms with closed-form densities and interesting properties. It nests in particular several

F. Forbes, A. Arnaud
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
E-mail: firstname.lastname@inria.fr

R. Steele
McGill, Montreal, Canada
E-mail: steele@math.mcgill.ca

B. Lemasson, E. Barbier
Grenoble Institut des Neurosciences, Inserm U1216, Univ. Grenoble Alpes, France
E-mail: firstname.lastname@univ-grenoble-alpes.fr

symmetric multiple scale heavy tailed distributions (such as generalized multivariate Student distributions [Forbes and Wraith, 2014]) and asymmetric multiple scale generalized hyperbolic distributions [Wraith and Forbes, 2015]. The multiple scale framework has also been used by [Franczak et al., 2015] for multiple scale shifted asymmetric Laplace distributions. The multiple scale framework has the advantage of allowing different tail and skewness behaviours in each dimension of the variable space with arbitrary correlation between dimensions. This is interesting when targeting clustering applications using mixtures of such distributions (see [Forbes and Wraith, 2014, Wraith and Forbes, 2015] for illustration). In this work we consider mixtures of multiple scale distributions in a Bayesian formulation for the many advantages that the Bayesian framework offers in the mixture models context, including natural procedures to automatically select the number of mixture components.

A standard M-variate scale mixture of Gaussians is a distribution of the form:

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\theta}) = \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/w)\ f_W(w;\boldsymbol{\theta})\ \mathrm{d}w \tag{1}$$

where $\mathcal{N}_M(\ .\ ;\boldsymbol{\mu},\boldsymbol{\Sigma}/w)$ denotes the M-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}/w$ and f_W is the probability distribution of a univariate positive variable W referred to hereafter as the weight variable. A common form is obtained when f_W is a Gamma distribution $\mathcal{G}(\nu/2,\nu/2)$ where ν denotes the degrees of freedom (we shall denote the Gamma distribution when the variable is X by $\mathcal{G}(x;\alpha,\gamma) = x^{\alpha-1}\Gamma(\alpha)^{-1}\exp(-\gamma x)\gamma^\alpha$ where Γ denotes the Gamma function). For this form, (1) is the standard density denoted by $t_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ of the M-dimensional Student t-distribution with parameters $\boldsymbol{\mu}$ (real location vector), $\boldsymbol{\Sigma}$ ($M\times M$ real positive definite scale matrix) and ν (positive real degrees of freedom parameter). Most of the work on multivariate scale mixture of Gaussians has focused on studying different choices for the weight distribution f_W (see *e.g.* Eltoft et al. [2006]) but the dimension of the weight variable W in most cases has been considered as univariate.

The extension proposed by Forbes and Wraith [2014] consists of introducing a multidimensional weight. To do so, the scale matrix is decomposed into $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^T$, where $\boldsymbol{D}$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ is a diagonal matrix with the corresponding eigenvalues of $\boldsymbol{\Sigma}$. This spectral decomposition is classically used in Gaussian model-based clustering [Banfield and Raftery, 1993, Celeux and Govaert, 1995]. The matrix $\boldsymbol{D}$ determines the orientation of the Gaussian and $\boldsymbol{A}$ its shape. Using this parameterization of $\boldsymbol{\Sigma}$, the scale Gaussian part in (1) is set to $\mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T)$, where $\boldsymbol{\Delta}_{\mathbf{w}} = \mathrm{diag}(w_1^{-1},\ldots,w_M^{-1})$ is the $M\times M$ diagonal matrix whose diagonal components are the inverse weights $\{w_1^{-1},\ldots,w_M^{-1}\}$. When the weights are all one, a standard multivariate Gaussian case is recovered. The multiple scale generalization consists therefore of:

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\theta}) = \int_0^\infty \ldots \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T)\ f_{\mathbf{w}}(w_1\ldots w_M;\boldsymbol{\theta})\ \mathrm{d}w_1\ldots\mathrm{d}w_M \tag{2}$$

where $f_{\mathbf{w}}$ is now a M-variate density function to be further specified. In the following developments, we will consider only independent weights, *i.e.* $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}$ with $f_{\mathbf{w}}(w_1\ldots w_M;\boldsymbol{\theta}) = f_{W_1}(w_1;\boldsymbol{\theta}_1)\ldots f_{W_M}(w_M;\boldsymbol{\theta}_M)$.

As an example, setting $f_{W_m}(w_m;\theta_m)$ to a Gamma distribution $\mathcal{G}(w_m;\alpha_m,\gamma_m)$ results in a multivariate generalization of a Pearson type VII distribution (see *e.g.*

Johnson et al. [1994] vol.2 chap. 28 for a definition of the Pearson type VII distribution) while setting $f_{W_m}(w_m)$ to $\mathcal{G}(w_m; \nu_m/2, \nu_m/2)$ leads to a generalization of the multivariate t-distribution. In both cases, we can express easily the respective densities denoted by $\mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ and $\mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$ with $\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_M\}$, $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_M\}$ and $\boldsymbol{\gamma} = \{\gamma_1 \ldots \gamma_M\}$:

$$\mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \prod_{m=1}^{M} \frac{\Gamma(\alpha_m + 1/2)}{\Gamma(\alpha_m)(2A_m\gamma_m\pi)^{1/2}} \left(1 + \frac{[\boldsymbol{D}^T(\mathbf{y} - \boldsymbol{\mu})]_m^2}{2A_m\gamma_m}\right)^{-(\alpha_m+1/2)} \tag{3}$$

Similarly,

$$\mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}) = \prod_{m=1}^{M} \frac{\Gamma((\nu_m + 1)/2)}{\Gamma(\nu_m/2)(A_m\nu_m\pi)^{1/2}} \left(1 + \frac{[\boldsymbol{D}^T(\mathbf{y} - \boldsymbol{\mu})]_m^2}{A_m\nu_m}\right)^{-(\nu_m+1)/2} \tag{4}$$

However for identifiability, model (3) needs to be further specified by fixing all γ_m parameters, for instance to 1. Despite this additional constraint, the decomposition of $\boldsymbol{\Sigma}$ still induces another identifiability issue. Both (3) and (4) are invariant to a same permutation of the column of $\boldsymbol{D}, \boldsymbol{A}$ and elements of $\boldsymbol{\alpha}$ or $\boldsymbol{\nu}$. In a frequentist setting this can be solved by imposing a decreasing order for the eigenvalues in $\boldsymbol{A}$. In a Bayesian setting one way to solve the problem is to impose on $\boldsymbol{A}$ a non symmetric prior (see Section 2.2) since an appropriate prior on $\boldsymbol{D}$ would be more difficult to set.

As shown in [Forbes and Wraith, 2014, Wraith and Forbes, 2015], this generalization to a multiple scale representation allows for a greater variety of shapes and in particular contours that are not necessarily elliptical. It is possible to account for very different tail behaviors across dimensions, such as a Gaussian (infinite ν_m) tail in one dimension and a Cauchy ($\nu_m = 1$) tail in an other dimension.

In previous work [Forbes and Wraith, 2014, Wraith and Forbes, 2015], inference has been carried out based on maximum likelihood principle and using the EM algorithm. In this work, we consider a Bayesian treatment of these models for the many advantages that the Bayesian framework offers in the mixture model context. Mainly, it avoids the ill-posed nature of maximum likelihood due to the presence of singularities in the likelihood function. A mixture component may collapse by becoming centered at a single data vector sending its covariance to 0 and the model likelihood to infinity. A Bayesian treatment protects the algorithm from this problem occurring in ordinary EM. Also, Bayesian model comparison embodies the principle that states that simple models should be preferred. Typically, maximum likelihood does not provide any guidance on the choice of the model order as more complex models can always fit the data better. For standard scale mixture of Gaussians, the usual Normal-Wishart prior can be used for the Gaussian parameters. In contrast, for multiple scale distributions, the decomposition of the covariance in (2) requires appropriate separated priors on the eigenvectors and eigenvalues of the scale matrix. Such priors do not derive easily from a standard conjugate choice. We propose a simple solution for which we can derive an inference scheme based on variational approximation and show how to apply it to select automatically an appropriate number of components in the mixtures. Following common practice that is to start from deliberately overfitting mixtures (*e.g.* Malsiner-Walli et al. [2016], Corduneanu and Bishop [2001], McGrory and Titterington [2007], Attias [1999]), we investigate the component-elimination property

of the Bayesian setting to select this number from a single run. We consider three different single-run strategies that avoid the repetitive inference and comparison of all possible models. They are all based on a Bayesian formulation specified in Section 2 and on a variational approximation inference detailed in Section 3. The three strategies are described in Section 4 and compared on simulated and real data in Section 5.

2 Bayesian mixture of multiple scale distributions

In this section, we outline the Bayesian model for a mixture of multiple scale Pearson VII distributions. In a Bayesian setting, it is more convenient to use the precision matrix $\boldsymbol{T}$ decomposed into $\boldsymbol{T} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^T$, which is the inverse of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{D}^T$, in the parameterization. Note that in (3) and in previous work, $\boldsymbol{A}$ is then replaced by $\boldsymbol{A}^{-1}$. Moreover, we consider for identifiability that all γ_m are set to 1. In a K-component mixture, the distributions we consider are therefore of the form, for each $k = 1 \dots K$,

$$\mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k, \boldsymbol{\alpha}_k) = \prod_{m=1}^{M} \frac{\Gamma(\alpha_{km} + 1/2) A_m}{\Gamma(\alpha_{km})(2\pi)^{1/2}} \left(1 + \frac{A_m [\boldsymbol{D}_k^T (\mathbf{y} - \boldsymbol{\mu}_k)]_m^2}{2}\right)^{-(\alpha_{km}+1/2)} \quad (5)$$

2.1 A hierarchical Bayesian model

Let us consider an *i.i.d* sample $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ from a K-component mixture of multiple scale distributions as defined in (5). With the usual notation for the mixing proportions $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$ and $\boldsymbol{\psi}_k = \{\boldsymbol{\mu}_k, \boldsymbol{T}_k, \boldsymbol{\alpha}_k\}$ for $k = 1 \dots K$, we consider,

$$p(\mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k \mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{T}_k, \boldsymbol{\alpha}_k)$$

where $\boldsymbol{\phi} = \{\boldsymbol{\pi}, \boldsymbol{\psi}\}$ with $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \dots \boldsymbol{\psi}_K\}$ denotes the mixture parameters. An additional variable Z can be introduced to identify the class labels: $\{Z_1, \dots, Z_N\}$ define respectively the components of origin of $\{\mathbf{y}_1, \dots, \mathbf{y}_N\}$. An equivalent modelling is therefore:

$$\forall i \in \{1 \dots N\}, \quad \mathbf{Y}_i | \mathbf{W}_i = \mathbf{w}_i, Z_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{D}_k \boldsymbol{\Delta}_{\mathbf{w}_i} \boldsymbol{A}_k^{-1} \boldsymbol{D}_k^T) \,,$$

$$\mathbf{W}_i | Z_i = k \sim \mathcal{G}(\alpha_{k1}, 1) \otimes \dots \otimes \mathcal{G}(\alpha_{kM}, 1) \,,$$

$$\text{and} \quad Z_i \sim \mathcal{M}(1, \pi_1, \dots, \pi_k) \,,$$

where $\boldsymbol{\Delta}_{\mathbf{w}_i} = \text{diag}(w_{i1}^{-1}, \dots, w_{iM}^{-1})$, symbol $\otimes$ means that the components of $\mathbf{W}_i$ are independent and $\mathcal{M}(1, \pi_1, \dots, \pi_k)$ denotes the Multinomial distribution.

2.2 Priors on component-specific parameters

To complete the Bayesian formulation, we assign priors on the parameters. However, it is common (see *e.g.* Archambeau and Verleysen [2007]) not to impose priors on the parameters $\boldsymbol{\alpha}_k$ since no convenient conjugate prior exist for these parameters. Then the scale matrix decomposition imposes that we set priors on $\boldsymbol{\mu}_k$ and $\boldsymbol{D}_k, \boldsymbol{A}_k$. For the means $\boldsymbol{\mu}_k$, the standard Gaussian prior can be used:

$$\boldsymbol{\mu}_k \mid \boldsymbol{A}_k, \boldsymbol{D}_k \sim \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{D}_k \boldsymbol{\Lambda}_k^{-1} \boldsymbol{A}_k^{-1} \boldsymbol{D}_k^T) \,, \tag{6}$$

where $\boldsymbol{m}_k$ (vector) and $\boldsymbol{\Lambda}_k$ (diagonal matrix) are hyperparameters. For $\boldsymbol{A}_k$ and $\boldsymbol{D}_k$ a natural solution would be to use the distributions induced by the standard Wishart prior on $\boldsymbol{T}_k$ but this appears not to be tractable in inference scheme even in a variational framework. The difficulty lies in considering an appropriate and tractable prior for $\boldsymbol{D}_k$. There exists a number of priors on the Stiefel manifold among which a good candidate could be the Bingham prior and extensions investigated by Hoff [2009]. However, it is not straightforward to derive from it a tractable E-$\boldsymbol{\Phi}^1$ step (see below) that could provide a variational posterior distribution. Nevertheless, this kind of priors could be added in the M-$\boldsymbol{D}$-step. The simpler solution adopted in the present work consists of considering $\boldsymbol{D}_k$ as an unknow fixed parameter and imposing a prior only on $\boldsymbol{A}_k$, which is a diagonal matrix containing the positive eigenvalues of $\boldsymbol{T}_k$. It is natural to choose:

$$\boldsymbol{A}_k \sim \otimes_{m=1}^{M} \mathcal{G}(\lambda_{km}, \delta_{km}) \,, \tag{7}$$

where $\boldsymbol{\lambda}_k = \{\lambda_{km}, m = 1 \dots M\}$ and $\boldsymbol{\delta}_k = \{\delta_{km}, m = 1 \dots M\}$ are hyperparameters. It follows the joint prior

$$p(\boldsymbol{\mu}, \boldsymbol{A}; \boldsymbol{D}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k | \boldsymbol{A}_k; \boldsymbol{D}_k)\, p(\boldsymbol{A}_k) \tag{8}$$

where the first term is given by (6) and the second term by (7).

2.3 Priors on mixing weights

We examine two cases. First, following Corduneanu and Bishop [2001], no prior is imposed on $\boldsymbol{\pi}$ (Section 3.1). Then a standard Dirichlet prior $\mathcal{D}(\tau_1, \dots, \tau_K)$ is used in a second case (Section 3.2).

For the complete model, the whole set of parameters is denoted by $\boldsymbol{\Phi}$. In our first setting, $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}^1, \boldsymbol{\Phi}^2\}$ is decomposed into a set $\boldsymbol{\Phi}^1 = \{\Phi_1^1, \dots \Phi_K^1\}$ with $\boldsymbol{\Phi}_k^1 = \{\boldsymbol{\mu}_k, \boldsymbol{A}_k\}$ of parameters for which we have priors and a set $\boldsymbol{\Phi}^2 = \{\boldsymbol{\pi}, \boldsymbol{D}, \boldsymbol{\alpha}\}$ of unknown parameters considered as fixed. In addition, hyperparameters are denoted by $\boldsymbol{\Phi}^3 = \{\Phi_1^3, \dots \Phi_K^3\}$ with $\boldsymbol{\Phi}_k^3 = \{\boldsymbol{m}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\lambda}_k, \boldsymbol{\delta}_k\}$. When a Dirichlet prior is used for $\boldsymbol{\pi}$, the parameters definitions change to $\boldsymbol{\Phi}_k^1 = \{\boldsymbol{\mu}_k, \boldsymbol{A}_k, \pi_k\}$, $\boldsymbol{\Phi}^2 = \{\boldsymbol{D}, \boldsymbol{\alpha}\}$ and $\boldsymbol{\Phi}_k^3 = \{\tau_k, \boldsymbol{m}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\lambda}_k, \boldsymbol{\delta}_k\}$.

3 Inference by variational Expectation-Maximization

The main task in Bayesian inference is to compute the posterior probability of the latent variables $\boldsymbol{X} = \{\boldsymbol{W}, \boldsymbol{Z}\}$ and the parameter $\boldsymbol{\Phi}$ for which only the $\boldsymbol{\Phi}^1$ part is considered as random. We are therefore interested in computing the posterior $p(\boldsymbol{X}, \boldsymbol{\Phi}^1 \mid \boldsymbol{y}, \boldsymbol{\Phi}^2)$. This posterior is intractable and approximated here using a variational approximation $q(\boldsymbol{X}, \boldsymbol{\Phi}^1)$ with a factorized form $q(\boldsymbol{X}, \boldsymbol{\Phi}^1) = q_X(\boldsymbol{X})\ q_{\Phi^1}(\boldsymbol{\Phi}^1)$. We propose to use an Expectation-Maximization (EM) framework to compute q. At iteration (r), the current fixed parameters values are denoted by $\boldsymbol{\Phi}^{2(r-1)}$. The EM alternating procedure proceeds as follows, with $\mathcal{D}$ the set of probability distributions that factorize as $q(\boldsymbol{X}, \boldsymbol{\Phi}^1) = q_X(\boldsymbol{X})\ q_{\Phi^1}(\boldsymbol{\Phi}^1)$,

$$\textbf{E-step:} \quad q^{(r)}(\boldsymbol{X}, \boldsymbol{\Phi}^1) = \arg\max_{q \in \mathcal{D}} \mathcal{F}(q, \boldsymbol{\Phi}^{2(r-1)})$$

$$\textbf{M-step:} \quad \boldsymbol{\Phi}^{2(r)} = \arg\max_{\boldsymbol{\Phi}^2} \mathcal{F}(q^{(r)}, \boldsymbol{\Phi}^2)\ ,$$

where $\mathcal{F}$ is the usual free energy

$$\mathcal{F}(q, \boldsymbol{\Phi}^2) = E_q[\log p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2)] - E_q[\log q(\boldsymbol{X}, \boldsymbol{\Phi}^1)]. \tag{9}$$

The full expression of the free energy is not necessary to maximize it and to derive the variational EM algorithm. However, computing the free energy is useful first because it allows a stopping criterion and a safety check for the variational implementation as the free energy should increase at each iteration. Then it can be used as specified in section 4.2 as a replacement of the likelihood to provide a model selection procedure. The detailed expression is then given in Appendix A and B, respectively for the case without and with prior on the weights.

The E-step above divides into two steps. At iteration (r), denoting in addition by $q_X^{(r-1)}$ the current variational distribution for $\boldsymbol{X}$:

$$\textbf{E-}\boldsymbol{\Phi}^1\textbf{-step:} \quad q_{\Phi^1}^{(r)}(\boldsymbol{\Phi}^1) \propto \exp(E_{q_X^{(r-1)}}[\log p(\boldsymbol{\Phi}^1 | \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\Phi}^{2(r-1)})])$$

$$\textbf{E-X-step:} \quad q_X^{(r)}(\boldsymbol{X}) \propto \exp(E_{q_{\phi^1}^{(r)}}[\log p(\boldsymbol{X} | \boldsymbol{y}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^{2(r-1)})])\ .$$

Then the M-step reduces to:

$$\textbf{M-step:} \quad \Phi^{2(r)} = \arg\max_{\boldsymbol{\phi}^2} E_{q_X^{(r)} q_{\Phi^1}^{(r)}}[\log p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2)]\ .$$

The resulting variational EM algorithm is further specified below in two cases depending on the prior used for the mixing weights.

3.1 No prior on mixing coefficients

This corresponds to a setting adopted by Corduneanu and Bishop [2001] where the mixing coefficients are estimated using type-II maximum likelihood. In this case, the complete likelihood $p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2, \boldsymbol{\Phi}^3)$ writes as

$$p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2, \boldsymbol{\Phi}^3) = p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{A}; \boldsymbol{D}) p(\boldsymbol{W}|\boldsymbol{Z}; \boldsymbol{\alpha}) p(\boldsymbol{Z}; \boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{A}; \boldsymbol{D}, \boldsymbol{m}, \boldsymbol{\Lambda}, \boldsymbol{\lambda}, \boldsymbol{\delta}).$$

Applying the previous general formulas, the E and M steps derive as follows.

E-$\boldsymbol{\Phi}^1$ step. In this setting, the variational posterior has the same structure as the prior distribution (8) . More specifically,
$q^{(r)}_{\boldsymbol{\Phi}^1}(\boldsymbol{\Phi}^1) = \prod_{k=1}^{K} q^{(r)}_{\boldsymbol{\Phi}^1}(\boldsymbol{\Phi}^1_k)$ with $q^{(r)}_{\boldsymbol{\Phi}^1}(\boldsymbol{\Phi}^1_k) = q^{(r)}_{\boldsymbol{\Phi}^1}(\boldsymbol{\mu}_k, \boldsymbol{A}_k) = q^{(r)}_{\mu_k|A_k}(\boldsymbol{\mu}_k \mid \boldsymbol{A}_k)\; q^{(r)}_{A_k}(\boldsymbol{A}_k)$
where

$$q^{(r)}_{\mu_k|A_k}(\boldsymbol{\mu}_k \mid \boldsymbol{A}_k) = \mathcal{N}(\boldsymbol{\mu}_k; \tilde{\boldsymbol{m}}^{(r)}_k, \tilde{\boldsymbol{\Sigma}}^{(r)}_k) \tag{10}$$

$$q^{(r)}_{A_k}(\boldsymbol{A}_k) = \prod_{m=1}^{M} \mathcal{G}(A_{km}; \tilde{\lambda}^{(r)}_{km}, \tilde{\delta}^{(r)}_{km}) \; . \tag{11}$$

Variational posteriors are defined via variational parameters $\tilde{\boldsymbol{m}}^{(r)}_k$, $\tilde{\boldsymbol{\Sigma}}^{(r)}_k$ and $\tilde{\lambda}^{(r)}_{km}, \tilde{\delta}^{(r)}_{km}$. These parameters involve $q^{(r-1)}_X$ via $q^{(r-1)}_{Z_i}(k) = q^{(r-1)}_X(Z_i = k)$ and

$$\tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki} = diag(\tilde{w}^{(r-1)}_{ki1}, \ldots, \tilde{w}^{(r-1)}_{kiM}) \tag{12}$$

where $\tilde{w}^{(r-1)}_{kim} = E_{q^{(r-1)}_X}[W_{im}|Z_i = k] = \tilde{\alpha}^{(r-1)}_{km}/\tilde{\gamma}^{(r-1)}_{kim}$. The specific expressions of $\tilde{\alpha}^{(r-1)}_{km}$ and $\tilde{\gamma}^{(r-1)}_{kim}$ are given in eq. (20) and (21) of the E-$\boldsymbol{X}$ step below but using values from iteration $(r-1)$. The covariance matrix $\tilde{\boldsymbol{\Sigma}}^{(r)}_k$ depends on $\boldsymbol{A}_k$ as in the prior (6) while after simplification $\tilde{\boldsymbol{m}}^{(r)}_k$ does not:

$$\tilde{\boldsymbol{\Sigma}}^{(r)}_k = \boldsymbol{D}^{(r-1)}_k \tilde{\boldsymbol{N}}^{(r)-1} \boldsymbol{A}^{-1}_k \boldsymbol{D}^{(r-1)T}_k \tag{13}$$

$$\begin{aligned}\tilde{\boldsymbol{m}}^{(r)}_k &= \tilde{\boldsymbol{\Sigma}}_k \boldsymbol{D}^{(r-1)}_k \boldsymbol{A}_k \left(\boldsymbol{\Lambda}_k \boldsymbol{D}^{(r-1)T}_k \boldsymbol{m}_k + \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki} \boldsymbol{D}^{(r-1)T}_k \boldsymbol{y}_i \right) \\ &= \boldsymbol{D}^{(r-1)}_k \tilde{\boldsymbol{N}}^{(r)-1} \left(\boldsymbol{\Lambda}_k \boldsymbol{D}^{(r-1)T}_k \boldsymbol{m}_k + \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki} \boldsymbol{D}^{(r-1)T}_k \boldsymbol{y}_i \right) \end{aligned} \tag{14}$$

$$\text{with} \qquad \tilde{\boldsymbol{N}}^{(r)}_k = \boldsymbol{\Lambda}_k + \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki} \; . \tag{15}$$

Then (11) is defined by the parameters:

$$\tilde{\lambda}^{(r)}_{km} = \lambda_{km} + 1/2 \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \tag{16}$$

$$\tilde{\delta}^{(r)}_{km} = \delta_{km} + 1/2 \; [\boldsymbol{M}_k]_{m,m} \tag{17}$$

where $[\boldsymbol{M}_k]_{m,m}$ denotes the m^{th} diagonal element on the following matrix $\boldsymbol{M}_k$:

$$\begin{aligned}\boldsymbol{M}_k = \boldsymbol{D}^{(r-1)T}_k &\left(\sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \boldsymbol{y}_i (\boldsymbol{y}_i - \tilde{\boldsymbol{m}}^{(r)}_k)^T \boldsymbol{D}^{(r-1)}_k \tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki} \right) \\ &+ \boldsymbol{D}^{(r-1)T}_k \boldsymbol{m}_k (\boldsymbol{m}_k - \tilde{\boldsymbol{m}}^{(r)}_k)^T \boldsymbol{D}^{(r-1)}_k \boldsymbol{\Lambda}_k \; . \end{aligned}$$

E-$\boldsymbol{X}$ step. $q^{(r)}_X(\boldsymbol{X}) = \prod_{i=1}^{N} q^{(r)}_{X_i}(\boldsymbol{W}_i, \boldsymbol{Z}_i)$ with

$$q^{(r)}_{W_i|Z_i}(\boldsymbol{W}_i|Z_i = k) = \prod_{m=1}^{M} q^{(r)}_{W_{im}|Z_i}(\boldsymbol{W}_{im}|\boldsymbol{Z}_i = k) = \prod_{m=1}^{M} \mathcal{G}(\boldsymbol{W}_{im}; \tilde{\alpha}^{(r)}_{km}, \tilde{\gamma}^{(r)}_{kim}) \tag{18}$$

$$q_{Z_i}^{(r)}(Z_i = k) = q_{Z_i}^{(r)}(k) \propto \pi_k^{(r-1)} \exp(\tilde{\rho}_k^{(r)}/2) \prod_{m=1}^{M} \frac{\Gamma(\tilde{\alpha}_{km}^{(r)})}{\Gamma(\alpha_{km}^{(r-1)})\ \tilde{\gamma}_{kim}^{(r)\tilde{\alpha}_{km}^{(r)}}} . \quad (19)$$

The right-hand side term above is easy to normalized. The variational parameters $\tilde{\alpha}_{km}^{(r)}, \tilde{\gamma}_{kim}^{(r)}$ are given by:

$$\tilde{\alpha}_{km}^{(r)} = \alpha_{km}^{(r-1)} + \frac{1}{2} \quad (20)$$

$$\tilde{\gamma}_{kim}^{(r)} = 1 + \frac{1}{2}\left(\tilde{A}_{km}^{(r)}[\boldsymbol{D}_k^{(r-1)T}(\boldsymbol{y}_i - \tilde{\boldsymbol{m}}_k^{(r)})]_m^2 + [\tilde{\boldsymbol{N}}_k^{(r)}]_{m,m}^{-1}\right) \quad (21)$$

where $\tilde{\boldsymbol{N}}_k^{(r)}$ is given in (15), $\tilde{\rho}_k^{(r)}$ and $\tilde{\boldsymbol{A}}_{km}^{(r)}$ are easily computed from (11) (Υ is the Digamma function):

$$\tilde{\rho}_k^{(r)} = E_{q_{A_k}^{(r)}}[\log|\boldsymbol{A}_k|] = \sum_{m=1}^{M} \Upsilon(\tilde{\lambda}_{km}^{(r)}) - \log \tilde{\delta}_{km}^{(r)} \quad (22)$$

$$\tilde{\boldsymbol{A}}_k^{(r)} = E_{q_{A_k}^{(r)}}[\boldsymbol{A}_k] \quad \text{that is for } m = 1 \dots M, \quad \tilde{\boldsymbol{A}}_{km}^{(r)} = \tilde{\lambda}_{km}^{(r)}/\tilde{\delta}_{km}^{(r)} . \quad (23)$$

M-step. The M-step divides into 3 sub-steps, where $\boldsymbol{\pi}, \boldsymbol{D}$ and $\boldsymbol{\alpha}$ are updated separately. The sum $\sum_{i=1}^{N} q_{Z_i}^{(r)}(k)$ is denoted by $n_k^{(r)}$.

M-$\boldsymbol{\pi}$-step. This step leads to the standard formula for mixtures. For $k = 1 \dots K$, $\boldsymbol{\pi}$ is updated as:

$$\pi_k^{(r)} = \sum_{i=1}^{N} q_{Z_i}^{(r)}(k)/N = n_k^{(r)}/N .$$

M-$\boldsymbol{\alpha}$- step. This step is less standard but equivalent to the update found in non Bayesian mixture of multiple scale distributions. The details can be found in the Supplementary material of [Forbes and Wraith, 2014]. In practice $\boldsymbol{\alpha}$ is updated as follows. The estimates do not exist in closed form, but are given as a solution of the equations below, for each $k = 1 \dots K$ and $m = 1 \dots M$:

$$\Upsilon(\alpha_{km}) = \Upsilon(\tilde{\alpha}_{km}^{(r)}) - \frac{1}{n_k^{(r)}} \sum_{i=1}^{N} q_{Z_i}^{(r)}(k) \log\left(\tilde{\gamma}_{kim}^{(r)}\right)$$

The resolution of these equations in α_{km} provides the $\alpha_{km}^{(r)}$.

M-$\boldsymbol{D}$-step. Each $\boldsymbol{D}_k$ can be updated separately as follows. Intermediate quantities are introduced to simplify the notation. For $i = 1 \dots N+1$:

$$\begin{aligned} \forall i = 1 \dots N, \quad \boldsymbol{V}_{ki}^{(r)} &= q_{Z_i}^{(r)}(k)(\boldsymbol{y}_i - \tilde{\boldsymbol{m}}_k^{(r)})(\boldsymbol{y}_i - \tilde{\boldsymbol{m}}_k^{(r)})^T \\ \boldsymbol{V}_{k(N+1)}^{(r)} &= (\boldsymbol{m}_k - \tilde{\boldsymbol{m}}_k^{(r)})(\boldsymbol{m}_k - \tilde{\boldsymbol{m}}_k^{(r)})^T \\ \tilde{\boldsymbol{\Delta}}_{k(N+1)}^{(r)} &= \boldsymbol{\Lambda}_k \end{aligned}$$

As already defined in (12) and (15),

$$\forall i = 1 \dots N, \quad \tilde{\boldsymbol{\Delta}}_{ki}^{(r)} = diag(\tilde{w}_{ki1}^{(r)}, \dots, \tilde{w}_{kiM}^{(r)})$$

and $\tilde{\boldsymbol{N}}_k^{(r+1)} = \boldsymbol{\Lambda}_k + \sum_{i=1}^{N} q_{Z_i}^{(r)}(k) \tilde{\boldsymbol{\Delta}}_{ki}^{(r)}$.

$$\begin{aligned}\boldsymbol{D}_k^{(r)} = \arg\min_{\boldsymbol{D}_k \in \mathcal{O}} \sum_{i=1}^{N+1} & trace(\boldsymbol{D}_k \tilde{\boldsymbol{\Delta}}_{ki}^{(r)} \tilde{\boldsymbol{A}}_k{}^{(r)} \boldsymbol{D}_k^T \boldsymbol{V}_{ki}^{(r)}) \\ & + E_{q_A^{(r)}}[trace(\boldsymbol{D}_k \tilde{\boldsymbol{N}}_k^{(r+1)} \boldsymbol{A}_k \boldsymbol{D}_k^T \ \boldsymbol{D}_k^{(r-1)} (\tilde{\boldsymbol{N}}_k^{(r)})^{-1} \boldsymbol{A}_k^{-1} \boldsymbol{D}_k^{(r-1)T})] \ .\end{aligned}$$

The exact computation of the expectation above is feasible but would result in an expression where the elements of $\boldsymbol{D}_k$ would be separated. As an alternative, we consider J *i.i.d* simulations of $\boldsymbol{A}_k$ according to distribution $q_{A_k}^{(r)}$ which is a product of Gamma distributions given in (11). Denoting by $\boldsymbol{A}_{kj}$ for $j = 1 \dots J$ these simulations, $\boldsymbol{D}_k^{(r)}$ can be approximated by,

$$\begin{aligned}\boldsymbol{D}_k^{(r)} \approx \arg\min_{\boldsymbol{D}_k \in \mathcal{O}} \sum_{i=1}^{N+1} & trace(\boldsymbol{D}_k \tilde{\boldsymbol{\Delta}}_{ki}^{(r)} \tilde{\boldsymbol{A}}_k{}^{(r)} \boldsymbol{D}_k^T \boldsymbol{V}_{ki}^{(r)}) \\ & + \tfrac{1}{J} \sum_{j=1}^{J} trace(\boldsymbol{D}_k \tilde{\boldsymbol{N}}_k^{(r+1)} \boldsymbol{A}_{kj} \boldsymbol{D}_k^T \ \boldsymbol{D}_k^{(r-1)} (\tilde{\boldsymbol{N}}_k^{(r)})^{-1} \boldsymbol{A}_{kj}^{-1} \boldsymbol{D}_k^{(r-1)T}) \ .\end{aligned}$$

The Monte-Carlo approximation of the expectation has the advantage to allow for the optimization of $\boldsymbol{D}_k$ on the Stiefel manifold $\mathcal{O}$. In [Celeux and Govaert, 1995, Forbes and Wraith, 2014, Wraith and Forbes, 2015], an algorithm by Flury and Gautschi [1986] was used but we consider here a more recent procedure using an accelerated line search method proposed by Browne and McNicholas [2014].

3.2 Dirichlet prior on mixing coefficients

In this section $\boldsymbol{\pi}$ is now considered as a random variable. More specifically the prior over $\boldsymbol{\Phi}^1$ writes

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{A}; \boldsymbol{\tau}, \boldsymbol{D}) = p(\boldsymbol{\pi}; \boldsymbol{\tau}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k | \boldsymbol{A}_k; \boldsymbol{D}_k) p(\boldsymbol{A}_k) \tag{24}$$

where $p(\boldsymbol{\pi}; \boldsymbol{\tau}) = \mathcal{D}(\boldsymbol{\pi}; \tau_1, \dots, \tau_K) = \dfrac{\Gamma(\sum_{k=1}^K \tau_k)}{\prod_{k=1}^K \Gamma(\tau_k)} \prod_{k=1}^{K} \pi_k^{\tau_k - 1}$ is a Dirichlet distribution.

With this modification, only the $p(\boldsymbol{Z}|\boldsymbol{\pi})\ p(\boldsymbol{\pi}; \boldsymbol{\tau})$ term changes in the complete likelihood which becomes,

$$p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2, \boldsymbol{\Phi}^3) = p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{A}; \boldsymbol{D}) p(\boldsymbol{W}|\boldsymbol{Z}; \boldsymbol{\alpha}) p(\boldsymbol{Z}|\boldsymbol{\pi})\ p(\boldsymbol{\pi}; \boldsymbol{\tau}) p(\boldsymbol{\mu}, \boldsymbol{A}; \boldsymbol{D}, \boldsymbol{m}, \boldsymbol{\Lambda}, \boldsymbol{\lambda}, \boldsymbol{\delta}).$$

Applying the previous general formulas, the E and M steps derive as follows.
E-$\boldsymbol{\Phi}^1$ step. With the same form for the prior and variational posterior, it comes

$$q^{(r)}_{\Phi^1}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{A}) = q^{(r)}(\boldsymbol{\pi}) \prod_{k=1}^{K} q^{(r)}_{\mu_k, \boldsymbol{A}_k}(\boldsymbol{\mu}_k, \boldsymbol{A}_k)$$

where $q^{(r)}_{\mu_k, \boldsymbol{A}_k}(\boldsymbol{\mu}_k, \boldsymbol{A}_k)$ has the same expression as given by (10) and (11). The new term is

$$q^{(r)}_{\boldsymbol{\pi}}(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \tilde{\tau}^{(r)}_1, \ldots, \tilde{\tau}^{(r)}_K)$$

with $\tilde{\tau}^{(r)}_k = \tau_k + \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) = \tau_k + n^{(r-1)}_k$.

E-$\boldsymbol{X}$ step. This step is only partly impacted by the addition of a prior on $\boldsymbol{\pi}$. It comes as in the previous section, $q^{(r)}_X(\boldsymbol{X}) = \prod_{i=1}^{N} q^{(r)}_{X_i}(\boldsymbol{W}_i, \boldsymbol{Z}_i)$ with the term below unchanged and given by (18), (20) and (21),

$$q^{(r)}_{W_i|Z_i}(\boldsymbol{W}_i \mid Z_i = k) = \prod_{m=1}^{M} q^{(r)}_{W_{im}|Z_i}(\boldsymbol{W}_{im} \mid \boldsymbol{Z}_i = k) \;=\; \prod_{m=1}^{M} \mathcal{G}(\boldsymbol{W}_{im}; \tilde{\alpha}^{(r)}_{km}, \tilde{\gamma}^{(r)}_{kim}) \; .$$

In contrast, the posterior on $\boldsymbol{Z}$ is changed into

$$q^{(r)}_{Z_i}(Z_i = k) = q^{(r)}_{Z_i}(k) \propto \tilde{\pi}^{(r)}_k \exp(\tilde{\rho}^{(r)}_k/2) \prod_{m=1}^{M} \frac{\Gamma(\tilde{\alpha}^{(r)}_{km})}{\Gamma(\alpha^{(r-1)}_{km})\, \tilde{\gamma}^{(r)\tilde{\alpha}^{(r)}_{km}}_{kim}} \; . \qquad (25)$$

where the modification reduces to changing $\pi^{(r-1)}_k$ into $\tilde{\pi}^{(r)}_k$ that can be derived from the previous E-$\boldsymbol{\Phi}^1$ step as

$$\log \tilde{\pi}^{(r)}_k = E_{q_{\boldsymbol{\pi}^{(r)}}}[\log \pi_k] = \Upsilon(\tilde{\tau}^{(r)}_k) - \Upsilon(\sum_{k=1}^{K} \tilde{\tau}^{(r)}_l) = \Upsilon(\tau_k + n^{(r)}_k) - \Upsilon(\sum_{k=1}^{K} \tau_l + N) \; .$$

The last term being constant with respect to (r) and k, it follows that $\tilde{\pi}^{(r)}_k$ is proportional to $\exp(\Upsilon(\tau_k + n^{(r)}_k))$ and it is enough to use this later expression in (25).

M-step. The M-step formulation remains the same as before without the M-$\boldsymbol{\pi}$ step.

In what follows, we illustrate the use of the Bayesian formulation and its variational EM implementation on the issue of selecting the number of components in the mixture.

4 Identifying the number of mixture components

When the number of components is unknown, apart from the Reversible Jump Markov Chain Monte Carlo method of Richardson and Green [1997] which allows jumps between different numbers of components, two types of approaches can be distinguished depending on whether the strategy is to increase or decrease the number of components. The first ones can be referred to as greedy algorithms (*e.g.*

Verbeek et al. [2003]) where the mixture is built component-wise, starting with the optimal one-component mixture and increasing the number of components until a stopping criterion is met. In this work, we will rather consider approaches that start from an overfitting model with more components than expected in the data. In this case, as described by Frühwirth-Schnatter [2006], identifiability will be violated in two possible ways. Identifiability issues can arise either because some of the components weights have to be zero (then component-specific parameters cannot be identified) or because some of the components have to be equal (then their weights cannot be identified). In practice, these two possibilities are not equivalent as checking for vanishing components is easier and is likely to lead to more stable behavior than testing for redundant components (see *e.g.* Rousseau and Mengersen [2011]).

Both increasing and decreasing methods can be considered in a Bayesian and maximum likelihood setting. However, in a Bayesian framework, in contrast to maximum likelihood, considering a posterior distribution on the mixture parameters requires integrating out the parameters and this acts as a penalization for more complex models. The posterior is essentially putting mass on the sparsest way to approximate the true density, see *e.g.* Rousseau and Mengersen [2011]. Although the framework of Rousseau and Mengersen [2011] is fully Bayesian with priors on all mixture parameters, it seems that this penalization effect is also effective when only some of the parameters are integrated out. This is observed by Corduneanu and Bishop [2001] who use priors only for the component mean and covariance parameters. This justifies in our setting the investigation of a case with no prior on the mixing weights (Section 3.1).

The idea of using overfitting finite mixtures with too many components K has been used in many papers. In a deliberately overfitting mixture model, a sparse prior on the mixture weights will empty superfluous components during estimation [Malsiner-Walli et al., 2016]. To obtain sparse solutions with regard to the number of mixture components, an appropriate prior on the weights $\boldsymbol{\pi}$ has to be selected. Guidelines have been given in previous work when the prior for the weights is a symmetric Dirichlet distribution $\mathcal{D}(\tau_1, \ldots, \tau_K)$ with all τ_k's equal to a value τ_0. To empty superfluous components automatically the value of τ_0 has to be chosen appropriately. In particular, Rousseau and Mengersen [2011] proposed conditions on τ_0 to control the asymptotic behavior of the posterior distribution of an overfitting mixture with respect to the two previously mentioned regimes. One regime in which a high likelihood is set to components with nearly identical parameters and one regime in which some of the mixture weights go to zero. More specifically, if $\tau_0 < d/2$ where d is the dimension of the component specific parameters, when N tends to infinity, the posterior expectation of the weight of superfluous components converges to zero. In practice, N is finite and as observed by Malsiner-Walli et al. [2016], much smaller value of τ_0 are needed (*e.g.* 10^{-5}). It was even observed by Tu [2016] that negative values of τ_0 were useful to induce even more sparsity when the number of observations is too large with respect to the prior impact. Dirichlet priors with negative parameters, although not formally defined, are also mentioned by Figueiredo and Jain [2002]. This latter work does not start from a Bayesian formulation but is based on a Minimum Message Length (MML) principle. Figueiredo and Jain [2002] provide an M-step that performs component annihilation, thus an explicit rule for moving from the current number

of components to a smaller one. A parallel is made with a Dirichlet prior with $\tau_0 = -d/2$ which according to Tu [2016] corresponds to a very strong prior sparsity.

4.1 Single-run number of component selection

In a Bayesian setting with symmetric sparse Dirichlet priors $\mathcal{D}(\tau_0, \ldots, \tau_0)$, the theoretical study of Rousseau and Mengersen [2011] therefore justifies to consider the posterior expectations of the weights $E[\pi_k|\boldsymbol{y}]$ and to prune out the too small ones. In practice this raises at least two additional questions: which expression to use for the estimated posterior means and how to set a threshold under which the estimated means are considered too small. The posterior means estimation is generally guided by the chosen inference scheme. For instance in our variational framework with a Dirichlet prior on the weights, the estimated posterior mean $E[\pi_k|\boldsymbol{y}]$ takes the following form (the (r) notation is removed to signify the convergence of the algorithm),

$$\begin{aligned} E[\pi_k|\boldsymbol{y}] \approx E_{q_{\boldsymbol{\pi}}}[\pi_k] &= \frac{\tilde{\tau}_k}{\sum_{l=1}^{K} \tilde{\tau}_l} \\ &= \frac{\tau_k + n_k}{\sum_{k=1}^{K} \tau_k + N} \end{aligned} \tag{26}$$

where $n_k = \sum_{i=1}^{N} q_{Z_i}(k)$ and $q_{Z_i}(k)$ is given by (25). If we are in the no weight prior case, then the expectation simplifies to

$$\pi_k \approx \frac{n_k}{N} \tag{27}$$

with $q_{Z_i}(k)$ given by (19).

Nevertheless, whatever the inference scheme or prior setting, we are left with the issue of detecting when a component can be set as empty. There is usually a close relationship between the component weight π_k and the number of observations assigned to component k. This later number is itself often replaced by the sum $n_k = \sum_{i=1}^{N} q_{Z_i}(k)$. As an illustration, the choice of a negative τ_0 by Figueiredo and Jain [2002] corresponds to a rule that sets a component weight to zero when $n_k = \sum_{i=1}^{N} q_{Z_i}(k)$ is smaller than $d/2$. This prevents the algorithm from approaching the boundary of the parameter space. When one of the components becomes too weak, meaning that it is not supported by the data, it is simply annihilated. One of the drawbacks of standard EM for mixtures is thus avoided. The rule of Figueiredo and Jain [2002] is stronger than that used by McGrory and Titterington [2007] which annihilates a component when the sum n_k reduces to 1 or the one of Corduneanu and Bishop [2001] which corresponds to the sum n_k lower than a very small fraction of the sample size, *i.e.* $\sum_{i=1}^{N} q_{Z_i}(k)/N < 10^{-5}$ where N varies from 400 to 900 in their experiments. Note that McGrory and Titterington [2007] use a Bayesian framework with variational inference and their rule corresponds to thresholding the variational posterior weights (26) to $1/N$ because they set all τ_k to 0 in their experiments.

In this work, we will also adopt a thresholding approach but note that alternatives have been developed that would worth testing to avoid the issue of setting a threshold for separating large and small weights. In their MCMC sampling, Malsiner-Walli et al. [2016] propose to consider the number of non-empty

components at each iteration and to estimate the number of components as the most frequent number of non-empty components. This is not directly applicable in our variational treatment as it would require to generate hard assignments to components at each iteration instead of dealing with their probabilities. In contrast, we could adopt techniques from the Bayesian non parametrics literature which seek for optimal partitions, such as the criterion of Dahl [2006] using the so-called posterior similarity matrix (Fritsch and Ickstadt [2009]). This matrix could be approximated easily in our case by computing the variational estimate of the probability that two observations are in the same component.

4.2 Tested procedures

We compare three types of single-run methods to estimate the number of components in a mixture of multiple scale distributions.

4.2.1 Fully Bayesian algorithm with sparse Dirichlet prior: "SparseDirichlet"

A first method is directly derived from a fully Bayesian setting with a sparse symmetric Dirichlet prior likely to induce vanishing coefficients as supported by the theoretical results of Rousseau and Mengersen [2011]. This corresponds to the approach adopted in Malsiner-Walli et al. [2016] and McGrory and Titterington [2007]. The difference between the later two being how they check for vanishing coefficients. Our variational inference leads more naturally to the solution of McGrory and Titterington [2007] which is to check the weight posterior means, that is whether at each iteration (r),

$$n_k^{(r)} < (K\tau_0 + N)\rho_t - \tau_0 \tag{28}$$

where ρ_t is the chosen threshold on the posterior means. When ρ_t is set such that (28) leads to $n_k^{(r)} < 1$, this method is referred to, in the next Section, as *SparseDirichlet+πtest*. For comparison, the algorithm run with no intervention is called *SparseDirichlet*.

4.2.2 Type II maximum likelihood on mixing weights: "TypeIIML"

A second method corresponds to the method proposed by Corduneanu and Bishop [2001]: no prior on the weights and a criterion on the estimated weights to detected vanishing coefficients. It corresponds to applying (28) with $\tau_0 = 0$. This method is referred below as *TypeIIML* when the algorithm is run until convergence and *TypeIIML+πtest* when (28) is used at each iteration with $\rho_t = 1/N$.

4.2.3 Free Energy based algorithm: "FEtest"

At last, we consider a criterion based on the free energy (9) to detect components to eliminate. In the no weight prior case, this is to handle potentially redundant components rather than vanishing ones and is also the opportunity to test a greedy potentially accelerated heuristic. This choice is based on the observation that when no prior is used for the weights, we cannot control the hyperparameters ($e.g\ \tau_k$) to

guide the algorithm in the vanishing components regime. Thus the algorithm may as well go to the redundant component regime. The goal is then to test whether this alternative method is likely to handle this behavior. The proposal is to start from a clustering solution with too many components and to try to remove them using a criterion based on the gain in free energy. In this setting, the components that are removed are not necessarily vanishing components but also redundant ones. The free energy expression used is given in Appendix A. With no prior on the weights, the algorithm is referred to as *TypeIIML+FEtest*. The same idea can be applied in the fully Bayesian setting, referred to here as *SparseDirichlet+FEtest*. The heuristic can be described as follows (see the next Section for implementation details).

1. Iteration $r = 0$: Initialization of the $K^{(0)}$ clusters and probabilities using for instance repetitions of k-means or trimmed k-means.
2. Iteration $r \geq 1$:
   (a) E and M steps updating from parameters at iteration $r-1$
   (b) Updating of the resulting Free Energy value
   (c) In parallele, for each cluster $k \in \{1 \dots K^{(r-1)}\}$
      i. Re-normalization of the cluster probabilities when cluster k is removed from current estimates at iteration $r-1$: the sum over the remaining $K^{(r-1)}-1$ clusters must be equal to 1
      ii. Updating of the corresponding E and M steps and computation of the associate Free Energy value
   (d) Selection of the mixture with the highest Free Energy among the $K^{(r-1)}$-component mixture (step (b)) or one of the $(K^{(r-1)}-1)$-component mixtures (step (c)).
   (e) Updating of $K^{(r)}$ accordingly, to $K^{(r-1)}$ or $K^{(r-1)}-1$.
3. When no more cluster deletion occur (*eg.* during 5 steps), we switch to the EM algorithm (*TypeIIML* or *SparseDirichlet*).

5 Experiments

In addition to the 6 methods mentioned above and referred to below as $\mathcal{MP}$ single-run procedures, we consider standard Gaussian mixtures using the Mclust package [Scrucca et al., 2016] including a version with priors on the means and covariance matrices. The Bayesian Information Criterion (BIC) is then used to select the number of components from $K = 1$ to 10. The respective methods are denoted below by *GM+BIC* and *Bayesian GM+BIC*. Regarding mixtures of $\mathcal{MP}$ distributions, we also consider their non Bayesian version, using BIC to select K, denoted below by *MMP+BIC*.

In practice, values need to be chosen for hyperparameters. These include the $\boldsymbol{m}_k$ that are set to 0, the $\boldsymbol{\Lambda}_k$ that are set to $\epsilon \boldsymbol{I}_M$ with ϵ small (set to 10^{-4}) so has to generate a large variance in (6). The δ_{km} are then set to 1 and λ_{km} to values $5 \times 10^{-4} = \lambda_1 < \lambda_2 < \ldots < \lambda_M = 10^{-3}$. When necessary, the τ_k's are set to 10^{-3} to favor sparse mixtures.

Initialization is also an important step in EM algorithms. All single-run methods are initialized with the same initial cluster assignments using $K = 10$ obtained with 10 repetitions of trimmed k-means for 10 iterations each and excluding 10% of outliers. For Gaussian mixtures, the initialization procedure is that embedded

in Mclust. For each run of the procedures, we allow 100 re-starts and choose the best one after 1000 iterations. For $\mathcal{MP}$ models, initial values of the α_{km}'s are set to 1.

Another important point for single-run procedures, is how to finally enumerate remaining components. For simplicity, we report components that are expressed by the maximum a posteriori (MAP) rule, which means components for which there is at least one data point assigned to them with the highest probability.

5.1 Simulated data

We first start with some simulated data from a mixture of $\mathcal{MP}$ distributions in dimension 2 with 3 components respectively centered at $[0, -2]$, $[0, 0]$ and $[0, 2]$ with the same scale matrix $[2, 0; 0, 0.2]$ and parameters $\alpha_{k1} = 1$ and $\alpha_{k2} = 100$ for $k = 1, 2, 3$. This example is a $\mathcal{MP}$ version of a Gaussian mixture used by Corduneanu and Bishop [2001] and McGrory and Titterington [2007] (see Figure 1 (a)). The sample size is $N = 900$ and 10 samples are simulated and used to test the different procedures. Table 1 summarizes the final observed or selected number of components for each procedure. For Gaussian mixtures $K = 6$ and 7 are the most selected values by BIC. The presence of data points in the tails induces the addition of components to capture all points that cannot be well explained by the 3 main visual components (Figures 1 (c) and (d)). In the MMP case, tails are rather well captured but BIC is hesitating between the 4 and 3 component solutions. For the 6 $\mathcal{MP}$ single-run procedures, the final number of components is almost always 3 and the clusterings are all very similar to the one shown in Figure 1 (e) (*TypeIIML+πtest* case). Mean computational times over the 10 repetitions are reported in seconds in Table 1 (last column). The procedure using the $\mathcal{MP}$ mixture and BIC is the longest due to repetitive runs for each value of K between 1 and 10. Computational gain is observed as expected when using one of the 4 procedures with component elimination. In particular, combining a sparse Dirichlet prior and free energy-based elimination seems to provide the largest gain with a running time (525s) divided by 5 compared to the *MMP+BIC* procedure (2767s).

Procedure (10 runs)	Selected number of components										Average time
	1	2	3	4	5	6	7	8	9	10	(in seconds)
GM+BIC	.	.	.	2	.	3	3	.	1	1	604
BayesianGM+BIC	.	.	.	8	1	.	.	.	.	1	252
MMP+BIC	.	.	5	5	.	.	.	.	.	.	2767
TypeIIML	.	.	9	1	.	.	.	.	.	.	1894
TypeIIML+πtest	.	.	10	.	.	.	.	.	.	.	718
TypeIIML+FEtest	.	.	10	.	.	.	.	.	.	.	748
SparseDirichlet	.	.	9	1	.	.	.	.	.	.	1891
SparseDirichlet+πtest	.	.	10	.	.	.	.	.	.	.	678
SparseDirichlet+FEtest	.	.	10	.	.	.	.	.	.	.	**525**

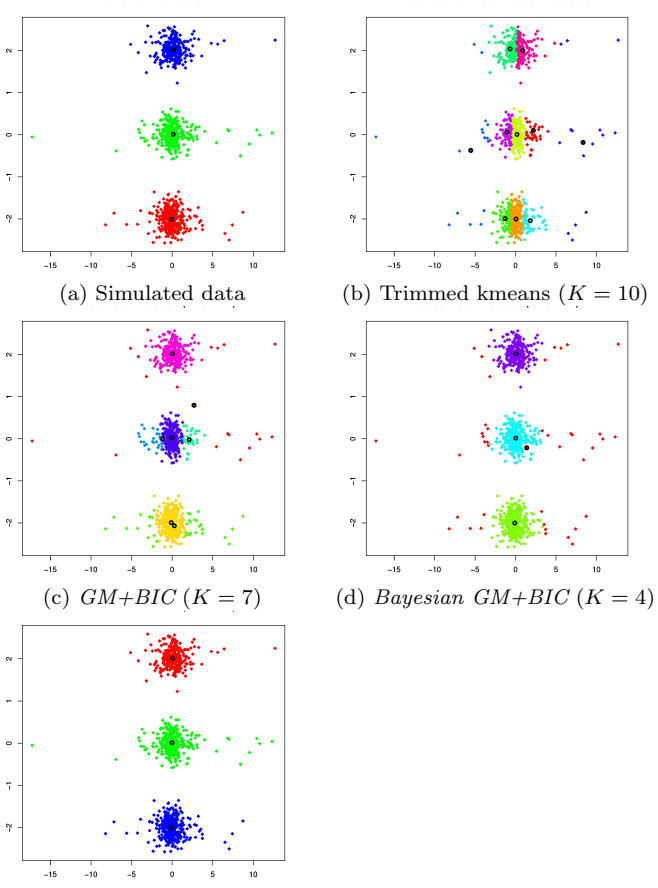
Table 1 Final observed or selected number of components for each procedure on 10 samples, and mean computational times in seconds.

A second example consists of 3 similar $\mathcal{MP}$ distributions but with closer means namely $[0, -1]$, $[0, 0]$ and $[0, 1]$ and $\alpha_{k1} = 2$, $\alpha_{k2} = 100$ for $k = 1, 2, 3$ (Figure 2 (a)). In terms of clustering and computational times, conclusions are similar as illustrated in Figure 2 and Table 2. All $\mathcal{MP}$ methods find $K = 3$ most of the time including the procedure using BIC.

5.2 Standard dataset

5.2.1 Old Faithful Geyser data

This data set contains 272 observations on 2 variables which are the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park. The scatter plot in Figure 3 (a) shows two moderately



(a) Simulated data (b) Trimmed kmeans ($K = 10$)

(c) *GM+BIC* ($K = 7$) (d) *Bayesian GM+BIC* ($K = 4$)

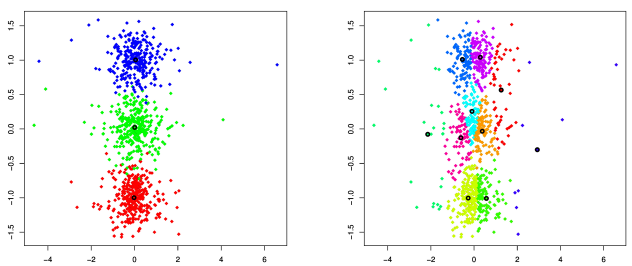
(e) $\mathcal{MP}$ mixture ($K = 3$)

Fig. 1 (a): Mixture of 3 $\mathcal{MP}$ distributions with $N = 900$, (b): 10 component initialization using trimmed k-means, (c): *GM+BIC* clustering ($K = 7$), (d): *Bayesian GM+BIC* clustering ($K = 4$), (e) $\mathcal{MP}$ mixture clustering ($K = 3$).

separated groups. For model selection this example has been studied in particular by Stephens [2000] with Gaussian mixtures. It was found that when more than 2 clusters are fit the extra components are there to model the deviation from normality in the two obvious groups rather than to model interpretable extra clusters. This is consistent with what we observe with Gaussian models (*Mclust+BIC* and *Bayesian Mclust+BIC*), finding 3 components (Figure 3 (c)) while our $\mathcal{MP}$ model with BIC and the free energy based elimination procedure show consistently 2 selected clusters (Figure 3 (d)). All other methods select 3 components but the clustering differs from that of the Gaussian models (Figures 3 (c) and (e)). All procedures were initialized with a 10 component assignment similar to that shown in Figure 3 (b). In terms of complexity, it appears that the fastest procedures are the free energy based ones (*FEtest*) (176s) with a time divided by 5 compared to the *MMP+BIC* one (1012s).

6 Discussion

We investigated, in the context of mixtures of non-Gaussian distributions, different single-run procedures to select automatically the number of components. The most time efficient single run strategies boil down to replace a discrete model choice (*e.g.* selecting K using BIC) by a continuous control parameter, here the threshold ρ_t over the posterior weights or the gain in free energy. The advantage of single run procedures is to avoid time consuming comparison of scores for each model. However, there are different ways to implement this idea: full Bayesian set-



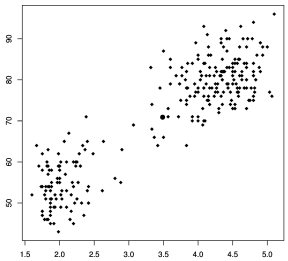
(a) Simulated data (b) Trimmed kmeans ($K = 10$)

(c) *BayesianGM+BIC* ($K = 4$) (d) $\mathcal{MP}$ mixture ($K = 3$)

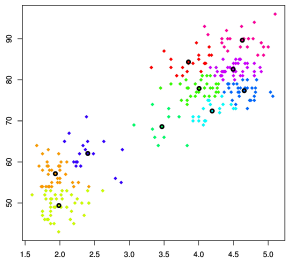
Fig. 2 a): Mixture of 3 closer $\mathcal{MP}$ distributions with $N = 900$, (b): 10 component initialization using trimmed k-means, (c): *Bayesian GM+BIC* clustering ($K = 4$), (d): $\mathcal{MP}$ mixture clustering ($K = 3$).

Procedure (10 runs)	Selected number of components										Average time (in seconds)
	1	2	3	4	5	6	7	8	9	10	
GM+BIC	.	.	.	10	.	.	.	.	.	.	860
BayesianGM+BIC	.	.	.	9	1	.	.	.	.	.	405
MMP+BIC	.	.	10	.	.	.	.	.	.	.	2584
TypeIIML	.	.	10	.	.	.	.	.	.	.	1917
TypeIIML+πtest	.	.	10	.	.	.	.	.	.	.	719
TypeIIML+FEtest	.	.	10	.	.	.	.	.	.	.	771
SparseDirichlet	.	.	10	.	.	.	.	.	.	.	1907
SparseDirichlet+πtest	.	.	10	.	.	.	.	.	.	.	**688**
SparseDirichlet+FEtest	.	.	10	.	.	.	.	.	.	.	751

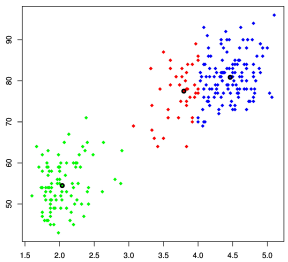
Table 2 Final observed or selected number of components for each procedure on 10 samples, and mean computational times in seconds.



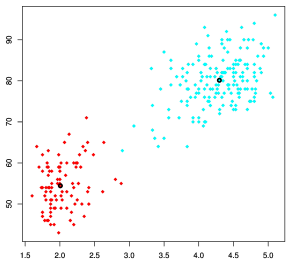
(a) Old Faithful data



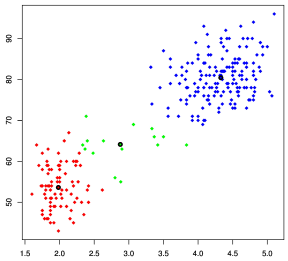
(b) Trimmed kmeans ($K = 10$)



(c) *GM+BIC* ($K = 3$)



(d) *MMP+BIC & FEtest* ($K = 2$)



(e) Other single-run methods ($K = 3$)

Fig. 3 (a): Old Faithful data, (b): 10 component initialization using trimmed k-means, (c): Gaussian mixture clustering as selected by BIC ($K = 3$), (d): $\mathcal{MP}$ clustering obtained with BIC and free energy based elimination ($K = 2$); (e): $\mathcal{MP}$ clustering for the other single-run procedures ($K = 3$).

tings which have the advantage to be supported by some theoretical justification [Rousseau and Mengersen, 2011] and Type II maximum likelihood as proposed by Corduneanu and Bishop [2001]. In this work, we proposed in addition another heuristic based on the gain in free energy. On preliminary experiments, we observed that both in terms of selection and computation time, Type II maximum likelihood on the weights was competitive with the use of a Dirichlet prior. In addition, both approaches were more efficient than BIC comparisons, in particular for $\mathcal{MP}$ mixtures which are more costly to estimate than Gaussian mixtures. Free energy based methods appeared to be slightly faster but not very significantly so compared to posterior weight thresholding methods. To confirm these observations, more tests in particular on larger and real data sets would be required to better compare and understand the various characteristics of each procedures.

A Expression of the free energy

The expression of the free energy at each iteration is needed to apply the procedure mentioned in section 4.2. It is given here in the absence of weight prior (Section 3.1). The free energy expression differs only slightly when a Dirichlet prior is added (see Appendix B). The free energy can be decomposed into two terms. At each iteration (r),

$$\mathcal{F}(q^{(r)}, \boldsymbol{\Phi}^{2(r)}) = E_{q^{(r)}}[\log p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^{2(r)})] - E_{q^{(r)}}[\log q^{(r)}(\boldsymbol{X}, \boldsymbol{\Phi}^1)] \,,$$

where the second term is made of entropies and the first term has been already computed in the M-step.

A.1 Entropy terms

In this section, we provide the expression of $-E_{q^{(r)}}[\log q^{(r)}(\boldsymbol{X}, \boldsymbol{\Phi}^1)]$ when it is equal to

$$\begin{aligned} -E_{q^{(r)}}[\log q^{(r)}(\boldsymbol{X}, \boldsymbol{\Phi}^1)] &= H[q^{(r)}_{\boldsymbol{X}}] + H[q^{(r)}_{\boldsymbol{\Phi}^1}] \\ &= \sum_{i=1}^{N} H[q^{(r)}_{X_i}] + \sum_{k=1}^{K} H[q^{(r)}_{\boldsymbol{\Phi}^1_k}] \,. \end{aligned}$$

In the expression above, $H[q^{(r)}_{\boldsymbol{\Phi}^1_k}]$ is the entropy of the Normal-Wishart distribution defined in (10) and (11),

$$\begin{aligned} H[q^{(r)}_{\boldsymbol{\Phi}^1_k}] = \frac{1}{2} \Bigg(& M \log 2\pi e - \log |\tilde{\boldsymbol{N}}^{(r)}_k| - \sum_{m=1}^{M} \Upsilon(\tilde{\lambda}^{(r)}_{km}) - \log \tilde{\delta}^{(r)}_{km} \Bigg) \\ & + \sum_{m=1}^{M} \left(\tilde{\lambda}^{(r)}_{km} - \log \tilde{\delta}^{(r)}_{km} + \log \Gamma(\tilde{\lambda}^{(r)}_{km}) + (1 - \tilde{\lambda}^{(r)}_{km}) \Upsilon(\tilde{\lambda}^{(r)}_{km}) \right) \,, \end{aligned}$$

where $\tilde{\boldsymbol{N}}^{(r)}_k$ is given by equation (15), $\tilde{\lambda}^{(r)}_{km}$ and $\tilde{\delta}^{(r)}_{km}$ by (16) and (17).

Then each term $H[q^{(r)}_{X_i}]$ is the sum of a product-of-Gamma entropy and a multinomial entropy,

$$\begin{aligned} H[q^{(r)}_{X_i}] = & \sum_{k=1}^{K} q^{(r)}_{Z_i}(k) \sum_{m=1}^{M} \left(\tilde{\alpha}^{(r)}_{km} + \log \Gamma(\tilde{\alpha}^{(r)}_{km}) + (1 - \tilde{\alpha}^{(r)}_{km}) \Upsilon(\tilde{\alpha}^{(r)}_{km}) - \log \tilde{\gamma}^{(r)}_{kim} \right) \\ & - \sum_{k=1}^{K} q^{(r)}_{Z_i}(k) \log q^{(r)}_{Z_i}(k) \end{aligned}$$

where $q^{(r)}_{Z_i}(k)$ is given by (19), $\tilde{\alpha}^{(r)}_{km}$ by (20) and $\tilde{\gamma}^{(r)}_{kim}$ by (21).

A.2 M-step terms

The term $E_{q^{(r)}}[\log p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^{2(r)})]$ decomposes into five terms,

$$E_{q^{(r)}}[\log p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^{2(r)})] = E_q[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{D}^{(r)})] + E_{q^{(r)}_{Z,W}}[\log p(\boldsymbol{W}|\boldsymbol{Z}; \boldsymbol{\alpha}^{(r)})] + E_{q^{(r)}_{Z}}[\log p(\boldsymbol{Z}; \boldsymbol{\pi}^{(r)})] + E_{q^{(r)}_{\boldsymbol{\mu},\boldsymbol{A}}}[\log p(\boldsymbol{\mu}|\boldsymbol{A}; \boldsymbol{D}^{(r)})] + E_{q^{(r)}_{\boldsymbol{A}}}[\log p(\boldsymbol{A})] \, .$$

The five terms are detailed in turn below,

$$E_{q^{(r)}}[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{D}^{(r)})] = -1/2 \sum_{i=1}^{N} \sum_{k=1}^{K} q^{(r)}_{Z_i}(k) \left(M \log 2\pi - \tilde{\rho}^{(r)}_k - \sum_{m=1}^{M} (\Upsilon(\tilde{\alpha}^{(r)}_{km}) - \log \tilde{\gamma}^{(r)}_{kim}) + (\tilde{m}^{(r)}_k - \mathbf{y}_i)^T \boldsymbol{D}^{(r)}_k \bar{\boldsymbol{\Delta}}^{(r)}_{ki} \tilde{\boldsymbol{A}}^{(r)}_k \boldsymbol{D}^{(r)T}_k (\tilde{m}^{(r)}_k - \mathbf{y}_i) + trace(\bar{\boldsymbol{\Delta}}^{(r)}_{ki} (\tilde{\boldsymbol{N}}^{(r)}_k)^{-1}) \right)$$

with $\bar{\boldsymbol{\Delta}}^{(r)}_{ki}$ given in (12), $\tilde{m}^{(r)}_k$ in (14), $\tilde{\boldsymbol{N}}^{(r)}_k$ in (15), $\tilde{\rho}^{(r)}_k$ and $\tilde{\boldsymbol{A}}^{(r)}_k$ in (22) and (23), $\tilde{\alpha}^{(r)}_{km}$ and $\tilde{\gamma}^{(r)}_{kim}$ in (20) and (21), $\boldsymbol{D}^{(r)}$ is the solution of the M-$\boldsymbol{D}$-step. .

$$E_{q^{(r)}_{Z,W}}[\log p(\boldsymbol{W}|\boldsymbol{Z}; \boldsymbol{\alpha}^{(r)})] = \sum_{i=1}^{N} \sum_{k=1}^{K} q^{(r)}_{Z_i}(k) \sum_{m=1}^{M} \left(-\log \Gamma(\alpha^{(r)}_{km}) + (\alpha^{(r)}_{km} - 1)(\Upsilon(\tilde{\alpha}^{(r)}_{km}) - \log \tilde{\gamma}^{(r)}_{kim}) - \frac{\tilde{\alpha}^{(r)}_{km}}{\tilde{\gamma}^{(r)}_{kim}} \right)$$

where $q^{(r)}_{Z_i}(k)$ is given in (19), $\alpha^{(r)}_{km}$ are the solutions of the M-$\boldsymbol{\alpha}$ step, $\tilde{\alpha}^{(r)}_{km}$ and $\tilde{\gamma}^{(r)}_{kim}$ are given in (20) and (21).

$$E_{q^{(r)}_{Z}}[\log p(\boldsymbol{Z}; \boldsymbol{\pi}^{(r)})] = (\sum_{k=1}^{K} n^{(r)}_k \log n^{(r)}_k) - N \log N$$

with $n^{(r)}_k = \sum_{i=1}^{N} q^{(r)}_{Z_i}(k)$.

$$E_{q^{(r)}_{\boldsymbol{\mu},\boldsymbol{A}}}[\log p(\boldsymbol{\mu}|\boldsymbol{A}; \boldsymbol{D}^{(r)})] = -1/2 \sum_{k=1}^{K} M \log 2\pi - \log |\boldsymbol{\Lambda}_k| - \tilde{\rho}^{(r)}_k + (\tilde{m}^{(r)}_k - m_k)^T \boldsymbol{D}^{(r)}_k \boldsymbol{\Lambda}_k \tilde{\boldsymbol{A}}^{(r)}_k \boldsymbol{D}^{(r)T}_k (\tilde{m}^{(r)}_k - m_k) + trace(\boldsymbol{\Lambda}_k (\tilde{\boldsymbol{N}}^{(r)}_k)^{-1})$$

where $\tilde{\boldsymbol{N}}^{(r)}_k$, $\tilde{m}^{(r)}_k$ are given in (15) and (14), $\tilde{\rho}^{(r)}_k$ is given in (22) and $\tilde{\boldsymbol{A}}^{(r)}_k$ in (23).

$$E_{q^{(r)}_{\boldsymbol{A}}}[\log p(\boldsymbol{A})] = \sum_{k=1}^{K} \sum_{m=1}^{M} \left(\lambda_{km} \log \delta_{km} - \log \Gamma(\lambda_{km}) + (\lambda_{km} - 1)(\Upsilon(\tilde{\lambda}^{(r)}_{km}) - \log \tilde{\delta}^{(r)}_{km}) - \delta_{km} \tilde{A}^{(r)}_{km} \right)$$

with $\tilde{A}^{(r)}_{km}$ given in (23), $\tilde{\lambda}^{(r)}_{km}$ and $\tilde{\delta}^{(r)}_{km}$ in (16) and (17).

B Free energy expression with a Dirichlet prior on the mixture weights

B.1 Entropy terms

The entropy terms are the same as in the previous Section with an additional term that corresponds to the entropy of $q_\pi^{(r)}$:

$$H[q_\pi^{(r)}] = \log B(\tilde{\boldsymbol{\tau}}^{(r)}) - (K - \tilde{\tau}_0^{(r)})\Upsilon(\tilde{\tau}_0^{(r)}) - \sum_{k=1}^{K}(\tilde{\tau}_k^{(r)} - 1)\Upsilon(\tilde{\tau}_k^{(r)})$$

where $B(\tilde{\boldsymbol{\tau}}^{(r)}) = \dfrac{\prod_{k=1}^{K}\Gamma(\tilde{\tau}_k^{(r)})}{\Gamma(\tilde{\tau}_0^{(r)})}$ and $\tilde{\tau}_0^{(r)} = \sum_{k=1}^{K}\tilde{\tau}_k^{(r)}$.

B.2 M-step terms

Similarly to the previous Section A, the term $E_{q^{(r)}}[\log p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^{2(r)})]$ decomposes now into six terms,

$$\begin{aligned} E_{q^{(r)}}[\log p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^{2(r)})] = {} & E_q[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{D}^{(r)})] + E_{q_{Z,W}^{(r)}}[\log p(\boldsymbol{W}|\boldsymbol{Z}; \boldsymbol{\alpha}^{(r)})] + E_{q_Z^{(r)} q_\pi^{(r)}}[\log p(\boldsymbol{Z}; \boldsymbol{\pi})] \\ & + E_{q_{\boldsymbol{\mu},\boldsymbol{A}}^{(r)}}[\log p(\boldsymbol{\mu}|\boldsymbol{A}; \boldsymbol{D}^{(r)})] + E_{q_{\boldsymbol{A}}^{(r)}}[\log p(\boldsymbol{A})] + E_{q_\pi^{(r)}}[\log p(\boldsymbol{\pi}; \boldsymbol{\tau})] \ . \end{aligned}$$

where the last term is an additional term not present in Section A and the third term has changed and is now

$$E_{q_Z^{(r)} q_\pi^{(r)}}[\log p(\boldsymbol{Z}; \boldsymbol{\pi})] = \left(\sum_{k=1}^{K} n_k^{(r)} \Upsilon(\tilde{\tau}_k^{(r)})\right) - N\Upsilon(\tilde{\tau}_0^{(r)})$$

The new term is,

$$E_{q_\pi^{(r)}}[\log p(\boldsymbol{\pi}; \boldsymbol{\tau})] = -\log B(\boldsymbol{\tau}) + (K - \tau_0)\Upsilon(\tilde{\tau}_0^{(r)}) + \sum_{k=1}^{K}(\tau_k - 1)\Upsilon(\tilde{\tau}_k^{(r)})$$

where $\tau_0 = \sum_{k=1}^{K}\tau_k$. All other terms have already been computed in Section A.

References

C. Archambeau and M. Verleysen. Robust Bayesian clustering. *Neural Networks*, 20(1):129–138, 2007.

H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pages 21–30, 1999.

J. Banfield and A.E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49(3):803–821, 1993.

R. Browne and P. McNicholas. Orthogonal stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*, 24, 03 2014.

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28 (5):781–793, 1995.

A. Corduneanu and C. Bishop. Variational Bayesian Model Selection for Mixture Distributions. In *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, page 2734. Morgan Kaufmann, January 2001.

D. B Dahl. Model-based clustering for expression data via a Dirichlet process mixture model, in Bayesian Inference for Gene Expression and Proteomics. 2006.

T. Eltoft, T. Kim, and T-W. Lee. Multivariate Scale Mixture of Gaussians Modeling. In Justinian Rosca, Deniz Erdogmus, Jose Principe, and Simon Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 799–806. Springer Berlin / Heidelberg, 2006.

M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

B. N. Flury and W. Gautschi. An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly Diagonal Form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184, 1986.

F. Forbes and D. Wraith. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering. *Statistics and Computing*, 24(6):971–984, 2014.

B. C. Franczak, C. Tortora, R. P. Browne, and P. D. McNicholas. Unsupervised learning via mixtures of skewed distributions with hypercube contours. *Pattern Recognition Letters*, 58: 69–76, 2015.

A. Fritsch and K. Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–391, 06 2009.

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Verlag, 2006.

P. D. Hoff. A Hierarchical Eigenmodel for Pooled Covariance Estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(5):971–992, 2009.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, vol.2, 2nd edition*. John Wiley & Sons, New York, 1994.

G. Malsiner-Walli, , S. Frühwirth-Schnatter, and B. Grün. Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26(1):303–324, Jan 2016.

C. A. McGrory and D. M. Titterington. Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions. *Comput. Stat. Data Anal.*, 51(11):5352–5367, July 2007.

S. Richardson and P. J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

L. Scrucca, M. Fop, T. B. Murphy, and A.E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016.

M. Stephens. Bayesian analysis of mixture models with an unknown number of componentsan alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74, 02 2000.

K. Tu. Modified Dirichlet Distribution: Allowing Negative Parameters to Induce Stronger Sparsity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1986–1991, 2016.

J. Verbeek, N. Vlassis, and B. Kröse. Efficient Greedy Learning of Gaussian Mixture Models. *Neural Computation*, 15(2):469–485, 2003.

D. Wraith and F. Forbes. Location and scale mixtures of Gaussians with flexible tail behaviour: Properties, inference and application to multivariate clustering. *Computational Statistics & Data Analysis*, 90:61–73, 2015.

CHAPITRE 5

COUPLAGE DU MODÈLE DE MÉLANGE DE LOIS DE PEARSON TYPE VII À MÉLANGE D'ÉCHELLES MULTIPLES AVEC UN CHAMP DE MARKOV LATENT POUR LA PRISE EN COMPTE DE DÉPENDANCES SPATIALES

Dans ce chapitre, nous présentons nos travaux sur le couplage d'un modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples (modèle MMSP) avec un champ de Markov latent, ceci afin de prendre la compte la dépendance spatiale de données IRM. La section 5.1 présente la structure spatiale de telles données, ainsi que l'ajout d'un champ de Markov au modèle MMSP. L'estimation de ce nouveau modèle est détaillée section 5.2, et les résultats sur données simulées et données IRM sont présentés section 5.3.

5.1 Modèle MMSP avec champ de Markov latent pour des données spatialement structurées

Modélisation de la structure spatiale sous-jacente des données IRM.
La segmentation automatique des images IRM est une étape importante et délicate, comme nous avons pu le voir dans le chapitre 3. Une des difficultés réside dans le fait que l'acquisition des données IRM en un voxel est impactée par la position spatiale de ce dernier. En effet, des inhomogénéités du champ magnétique sont possibles et modifient ou dégradent les mesures. En de tels voxels, considérer les voxels proches est un moyen de contrebalancer en partie ces artefacts. Les champs

de Markov sont une modélisation statistique possible pour la prise en compte des informations spatiales, comme dans le cas de l'estimation des modèles de mélange pour la segmentation d'image ([28]).

Dans cette partie, nous allons ainsi étendre le modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP) précédemment utilisé au cas de données spatialement dépendantes. Dans la suite, nous considèrerons des données issues d'acquisition IRM afin d'illustrer nos propos, puisque cela correspond à l'application finale visée.

Bien que l'acquisition IRM soit régulière et basée sur une grille, les régions d'intérêt que nous souhaitons segmenter ne présentent pas nécessairement une structure de grille. C'est par exemple le cas pour les données d'IRM cérébrales où seuls les voxels formant le cerveau sont gardés (chapitre 3, Figure 3). Nous considérons donc une représentation plus générale de structure : les voxels forment les sommets d'un graphe $\mathcal{G}$, et sont reliés entre eux par des arrêtes s'ils sont dits voisins. Le graphe $\mathcal{G}$ défini ainsi un système de voisinage : seuls les voisins d'un voxel vont influer sur sa classification. Un voisinage simple est de marquer deux voxels comme voisins s'ils sont adjacents lors de l'acquisition IRM. Ce voisinage correspond alors à un système des 8-plus proches voisins pour une acquisition 2D en considérant comme adjacents des voxels côte à côte sur les axes orthogonaux et les diagonales. Pour plus de détails sur les champs de Markov, le lecteur peut consulter [29]. Dans la suite, nous allons étendre le modèle de mélange précédemment utilisé en imposant un champ de Markov sur les observations au travers d'une distribution de Gibbs sur les variables latentes de classe, section 5.1. En effet, tout champ de Markov est équivalent à une distribution de Gibbs par le théorème de Hammersley-Clifford, or celle-ci s'exprimant sous forme de distribution est plus facile à manipuler et à intégrer au sein d'un modèle de mélange. De plus, en imposant une contrainte spatiale uniquement sur les variables latentes de classe, nous allons pouvoir réaliser une estimation du modèle, section 5.2, similaire au cas sans dépendance spatiale. Enfin, nous présentons section 5.3 des résultats sur données simulées pour mesurer l'impact du champ de Markov dans l'estimation du modèle de mélange proposé ; un premier résultat sur des données réelles d'IRM est également présenté.

Modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples avec champ de Markov latent

Nous reprenons ici le modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples, décrit dans la section 2 de l'article TMI au chapitre 3, pour y incorporer une distribution de Gibbs. Pour cela, considérons un échantillon $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, issu d'un mélange à K composantes, associé aux variables latentes de classe $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N\}$, reliant chaque observation à sa composante au sein

du mélange, et aux variables latentes de poids $\boldsymbol{W} = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{\mathrm{N}}\}$, explicitant les lois à mélange d'échelles multiples comme un mélange infini. De plus, dans le cadre du mélange, les observations $\boldsymbol{y}$ sont conditionnellement indépendantes par rapport aux variables de classe $\boldsymbol{Z}$. La loi jointe de ce mélange s'écrit :

$$p(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z} \,;\, \boldsymbol{\phi}) \;=\; p\left(\boldsymbol{y} \,|\, \boldsymbol{W} = \boldsymbol{w}, \boldsymbol{Z} = \boldsymbol{z} \,;\, \boldsymbol{\psi}^1\right) p\left(\boldsymbol{w} \,|\, \boldsymbol{Z} = \boldsymbol{z} \,;\, \boldsymbol{\psi}^1\right) p\left(\boldsymbol{z} \,;\, \boldsymbol{\psi}^2\right) \tag{5.1}$$

avec $\boldsymbol{\phi} = \left\{\boldsymbol{\psi}^1, \boldsymbol{\psi}^2\right\}$ contenant les paramètres du mélange ($\boldsymbol{\psi}^1$) et des classes ($\boldsymbol{\psi}^2$).

Dans le chapitre 3, la loi des variables de classe $p\left(\boldsymbol{z} \,;\, \boldsymbol{\psi}^2\right)$ se factorisait sous la forme d'un produit de lois multinomiales partageant les même proportions de classes $\boldsymbol{\psi}^2 = (\pi_1, \ldots, \pi_{\mathrm{K}})$:

$$p\left(\boldsymbol{z} \,;\, \boldsymbol{\psi}^2\right) = \prod_{i=1}^{\mathrm{N}} \mathcal{M}\left(z_i \,;\, 1, \pi_1, \ldots, \pi_{\mathrm{K}}\right)$$

Cette factorisation entraîne l'indépendance des variables $\boldsymbol{z}$. Afin d'inclure une dépendance spatiale entre les observations $\boldsymbol{y}$ au travers des variables de classe $\boldsymbol{z}$, nous remplaçons ces lois multinomiales par une distribution de Gibbs. La distribution des variables de classes $p\left(\boldsymbol{z} \,;\, \boldsymbol{\psi}^2\right)$ devient ainsi :

$$p\left(\boldsymbol{z} \,;\, \boldsymbol{\psi}^2\right) = p_G\left(\boldsymbol{z} \,;\, \boldsymbol{\tau}, \beta\right) = \mathcal{K}\left(\boldsymbol{\tau}, \beta\right)^{-1} \exp\left[H\left(\boldsymbol{z} \,;\, \boldsymbol{\tau}, \beta\right)\right] \tag{5.2}$$

en notant maintenant les paramètres $\boldsymbol{\psi}^2 = (\boldsymbol{\tau}, \beta)$, avec $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{\mathrm{K}}\}$ les paramètres du champ externe, et β le paramètre d'interaction locale.

Nous choisissons comme fonction énergie :

$$H\left(\boldsymbol{z} \,;\, \boldsymbol{\tau}, \beta\right) = \sum_{i=1}^{\mathrm{N}} \left[\tau_{z_i} + \frac{\beta}{2} \sum_{l \in \Omega(i)} \mathbb{I}_{z_i}\left(z_l\right)\right] \tag{5.3}$$

avec $\Omega(i)$ le voisinage de l'observation i [1], et $\mathcal{K}$ la constante de normalisation définie par : $\mathcal{K}\left(\boldsymbol{\tau}, \beta\right) = \sum_{\boldsymbol{z}} \exp\left[H\left(\boldsymbol{z} \,;\, \boldsymbol{\tau}, \beta\right)\right]$.

Il est nécessaire de poser une contrainte sur le paramètre $\boldsymbol{\tau}$ pour des raisons d'identifiabilité des classes. Nous pouvons supposer par exemple que $\tau_1 = 0$, ou encore que $\sum_{k=1}^{\mathrm{K}} e^{\tau_k} = 1$ pour retrouver la contrainte sur les proportions de classes dans le cas non markovien (ce qui revient alors à choisir $\beta = 0$). C'est cette dernière contrainte qui sera retenue par la suite pour les tests numériques. À noter que la fonction d'énergie choisie ne dépend pas des coordonnées spatiales des observations,

1. Dans le cas des données IRM, $\Omega(i)$ est l'ensemble des voxels qui sont voisins du voxel i.

la classe z_i de l'observation i ne dépend que de la classe de ses voisins $\Omega(i)$, en plus des paramètres $(\boldsymbol{\tau}, \beta)$ de la distribution de Gibbs.

Nous avons ainsi la représentation hiérarchique suivante :

$$\begin{aligned} \boldsymbol{Z} &\sim p_G(\boldsymbol{\tau}, \beta) \\ \text{et } \forall i \in \{1, \ldots, \mathrm{N}\} &: \\ \boldsymbol{W}_i \,|\, Z_i = k &\sim \mathcal{G}(\alpha_{k,1}, 1) \otimes \ldots \otimes \mathcal{G}(\alpha_{k,\mathrm{M}}, 1) \\ \boldsymbol{Y}_i \,|\, \boldsymbol{W}_i = \boldsymbol{w}_i, Z_i = k &\sim \mathcal{N}_{\mathrm{M}}\left(\boldsymbol{\mu}_k, \boldsymbol{D}_k \Delta_{\boldsymbol{w}_i}^{-1} \boldsymbol{A}_k^{-1} \boldsymbol{D}_k^{\mathrm{t}}\right) \end{aligned}$$

avec les paramètres $\boldsymbol{\phi} = \left\{\boldsymbol{\psi}^1, \boldsymbol{\psi}^2\right\}$ où $\boldsymbol{\psi}^1 = \{\psi_1^1, \ldots, \psi_{\mathrm{K}}^1\}$ les paramètres du mélange qui ne changent pas par rapport à la version sans dépendance spatiale : $\psi_k^1 = \{\boldsymbol{\mu}_k, \boldsymbol{T}_k, \boldsymbol{\alpha}_k\}$ le vecteur de position $\boldsymbol{\mu}_k$, la matrice d'échelle se décomposant en valeurs singulières $\boldsymbol{T}_k = \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^{\mathrm{t}}$, et le vecteur des degrés de liberté $\boldsymbol{\alpha}_k$. À noter un changement de notation par soucis de simplicité : nous noterons maintenant $\Delta_{\boldsymbol{w}_i} = \mathrm{diag}(\boldsymbol{w}_i)$ la matrice diagonale des poids intervenant dans la loi de Pearson type VII à mélange d'échelles multiples.

5.2 Inférence du modèle par un algorithme d'Espérance-Maximisation variationnel

L'estimation d'un mélange de lois de Pearson type VII à mélange d'échelles multiples peut se faire par maximum de vraisemblance en utilisant l'algorithme d'Espérance-Maximisation (EM), tel que décrit par [7]. Cependant, cette approche n'est pas directement applicable ici car la constante de normalisation $\mathcal{K}(\boldsymbol{\tau}, \beta)$ de la distribution de Gibbs empêche un calcul explicite de la distribution a posteriori des variables de classe $\boldsymbol{Z}$ conditionnellement aux observations $\boldsymbol{y}$. Il est possible de contourner cette difficulté en approchant la distribution a posteriori $p(\boldsymbol{w}, \boldsymbol{z} \,|\, \boldsymbol{y} \,;\, \boldsymbol{\phi})$ par une approximation variationnelle $q(\boldsymbol{w}, \boldsymbol{z}) = \prod_{i=1}^{\mathrm{N}} q_{\boldsymbol{W}_i, Z_i}(\boldsymbol{w}_i, z_i)$.

Cette factorisation permet de réduire la complexité des interactions entre variables latentes au sein de $p(\boldsymbol{w}, \boldsymbol{z} \,|\, \boldsymbol{y} \,;\, \boldsymbol{\phi})$, et ainsi de permettre une estimation explicite par maximum a posteriori (MAP) de l'approximation $q(\boldsymbol{w}, \boldsymbol{z})$. Nous pouvons alors dériver un algorithme EM pour l'estimation de $q(\boldsymbol{w}, \boldsymbol{z})$. À l'itération (r), la valeur courante des paramètres du modèle $\boldsymbol{\phi}$ est fixée à $\boldsymbol{\phi}^{(r-1)}$, et l'algorithme EM se décompose en deux étapes classiques, en notant $\mathcal{D}_q$ l'ensemble des distributions ayant la factorisation voulue $q(\boldsymbol{w}, \boldsymbol{z}) = \prod_{i=1}^{\mathrm{N}} q_{\boldsymbol{W}_i, Z_i}(\boldsymbol{w}_i, z_i)$, et $\mathcal{D}_{\boldsymbol{\phi}}$ l'espace des

paramètres du modèle :

$$\textbf{étape E :} \qquad q^{(r)} = \underset{q \in \mathcal{D}_q}{\arg\max}\, \mathcal{F}\left(q, \boldsymbol{\phi}^{(r-1)}\right) \tag{5.4}$$

$$\textbf{étape M :} \qquad \boldsymbol{\phi}^{(r)} = \underset{\phi \in \mathcal{D}_{\boldsymbol{\phi}}}{\arg\max}\, \mathcal{F}\left(q^{(r)}, \boldsymbol{\phi}\right) \tag{5.5}$$

où $\mathcal{F}$ correspond à l'énergie libre du modèle :

$$\mathcal{F}(q, \boldsymbol{\phi}) = \mathbb{E}_q\left[\log p(\boldsymbol{y}, \boldsymbol{W}, \boldsymbol{Z}\ ;\ \boldsymbol{\phi}) - \log q(\boldsymbol{W}, \boldsymbol{Z})\right]\ . \tag{5.6}$$

L'expression de l'énergie libre du modèle est donnée en annexe B.4 afin de contrôler l'exécution de l'implémentation numérique du modèle. En effet, l'énergie libre doit augmenter à chaque itération de l'algorithme EM, en ce sens cela permet de s'assurer de la bonne exécution de l'algorithme. De plus, puisqu'il s'agit de maximiser l'énergie libre, celle-ci peut servir de critère d'arrêt à l'algorithme EM.

Du fait de la factorisation de l'approximation variationnelle, l'étape E se décompose en N étapes, une par observation. Ainsi, lors de la mise à jour de la distribution $q_{\boldsymbol{W}_i, Z_i}$ à l'itération (r), les variables latentes autres que $\boldsymbol{W}_i$ et Z_i, c'est-à-dire $\boldsymbol{W}_{\setminus i} = \{\boldsymbol{W}_l\ ;\ l = 1, \ldots, \mathrm{N}, l \neq i\}$ et $\boldsymbol{Z}_{\setminus i} = \{Z_l\ ;\ l = 1, \ldots, \mathrm{N}, l \neq i\}$ sont intégrées grâce à leur distribution associée $q^{(r-1)}_{\boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i}} = \prod_{l=1\,, l \neq i}^{\mathrm{N}} q^{(r-1)}_{\boldsymbol{W}_l, Z_l}$. L'étape E se décompose ainsi comme suit :

étape E- $(\boldsymbol{W}_i, Z_i)$

$$q^{(r)}_{\boldsymbol{W}_i, Z_i}(\boldsymbol{w}_i, z_i) \quad \propto \quad \exp\left[\mathrm{E}_{q^{(r-1)}_{\boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i}}} \ln p\left(\boldsymbol{w}_i, z_i \,|\, \boldsymbol{y}, \boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i}\ ;\ \boldsymbol{\phi}^{(r-1)}\right)\right] \tag{5.7}$$

De manière similaire, de part la décomposition de la distribution jointe :

$$\begin{aligned} \ln p(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z}\ ;\ \boldsymbol{\phi}) \quad = \quad & \ln p(\boldsymbol{y} \,|\, \boldsymbol{W} = \boldsymbol{w}, \boldsymbol{Z} = \boldsymbol{z}\ ;\ \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}) \\ & + \ln p(\boldsymbol{w} \,|\, \boldsymbol{Z} = \boldsymbol{z}\ ;\ \boldsymbol{\alpha}) \\ & + \ln p(\boldsymbol{z}\ ;\ \boldsymbol{\tau}, \beta) \end{aligned} \tag{5.8}$$

l'étape M se décompose en sous-étapes :

$$\begin{aligned}
&\textbf{étape M-}\ (\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}) \\
\left(\boldsymbol{\mu}^{(r)}, \boldsymbol{D}^{(r)}, \boldsymbol{A}^{(r)}\right) &= \underset{\boldsymbol{\mu},\boldsymbol{D},\boldsymbol{A}}{\arg\max}\, \mathrm{E}_{q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}} \left[\ln p\left(\boldsymbol{y} \,|\, \boldsymbol{W}, \boldsymbol{Z}\ ;\ \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}\right)\right] \qquad (5.9) \\
&\textbf{étape M-}\ (\boldsymbol{\alpha}) \\
\boldsymbol{\alpha}^{(r)} &= \underset{\boldsymbol{\alpha}}{\arg\max}\, \mathrm{E}_{q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}} \left[\ln p\left(\boldsymbol{W} \,|\, \boldsymbol{Z}\ ;\ \boldsymbol{\alpha}\right)\right] \qquad (5.10) \\
&\textbf{étape M-}\ (\boldsymbol{\tau}, \beta) \\
\left(\boldsymbol{\tau}^{(r)}, \beta^{(r)}\right) &= \underset{(\boldsymbol{\tau},\beta)}{\arg\max}\, \mathrm{E}_{q^{(r)}_{\boldsymbol{Z}}} \left[\ln p\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)\right] \qquad (5.11)
\end{aligned}$$

Étape variationnelle d'espérance - Étape E-$(\boldsymbol{W}_i, Z_i)$

La factorisation utilisée pour l'approximation variationnelle entraîne, pour la distribution a posteriori des variables latentes, la même structure conjuguée que dans le modèle sans dépendance spatiale de [7]. Le détail des calculs est donné en annexe B.1 et permet d'obtenir les expressions suivantes :

$$\begin{aligned}
q^{(r)}_{\boldsymbol{W}_i, Z_i}\left(\boldsymbol{w}_i, z_i\right) &= q^{(r)}_{\boldsymbol{W}_i \,|\, Z_i = z_i}\left(\boldsymbol{w}_i\right) \cdot q^{(r)}_{Z_i}\left(z_i\right) \\
\text{avec } q^{(r)}_{Z_i}\left(k\right) &= \mathcal{M}\left(k\ ;\ 1, \widetilde{\pi}^{(r)}_{i,1}, \ldots, \widetilde{\pi}^{(r)}_{i,\mathrm{K}}\right),\ k = 1, \ldots, \mathrm{K} \qquad (5.12) \\
\text{et } q^{(r)}_{\boldsymbol{W}_i \,|\, Z_i = k}\left(\boldsymbol{w}_i\right) &= \prod_{m=1}^{\mathrm{M}} \mathcal{G}\left(w_{i,m}\ ;\ \widetilde{\gamma}^{(r)}_{k,m}, \widetilde{\delta}^{(r)}_{i,k,m}\right) \qquad (5.13)
\end{aligned}$$

La différence étant dans l'expression des paramètres des lois multinomiales relatives aux variables de classe. Celles-ci contiennent la dépendance spatiale entre observations, ce qui correspond à l'approximation par champ moyen :

$$\widetilde{\pi}^{(r)}_{i,k} = \frac{\exp\left[\widetilde{\boldsymbol{\tau}}^{(r)}_{i,k} + \frac{\beta^{(r-1)}}{2} \sum_{l \in \Omega(i)} q^{(r-1)}_{Z_l}\left(k\right)\right]}{\sum_{j=1}^{\mathrm{K}} \exp\left[\widetilde{\boldsymbol{\tau}}^{(r)}_{i,j} + \frac{\beta^{(r-1)}}{2} \sum_{l \in \Omega(i)} q^{(r-1)}_{Z_l}\left(j\right)\right]} \qquad (5.14)$$

avec

$$\widetilde{\tau}^{(r)}_{i,k} = \tau^{(r-1)}_{k} + \sum_{m=1}^{\mathrm{M}} \left\{ \frac{1}{2} \ln\left(\left[A^{(r-1)}_{k}\right]_{m,m}\right) - \widetilde{\gamma}^{(r)}_{k,m} \ln\left(\widetilde{\delta}^{(r)}_{i,k,m}\right) + \ln\left(\frac{\Gamma\left(\widetilde{\gamma}^{(r)}_{k,m}\right)}{\Gamma\left(\alpha^{(r-1)}_{k,m}\right)}\right) \right\}$$

Les paramètres des lois Gamma étant donnés quant à eux par :

$$\widetilde{\gamma}_{k,m}^{(r)} = \alpha_{k,m}^{(r-1)} + \frac{1}{2} \tag{5.15}$$

$$\widetilde{\delta}_{i,k,m}^{(r)} = 1 + \frac{1}{2}\left[\boldsymbol{A}_k^{(r-1)}\boldsymbol{D}_k^{(r-1)^{\mathrm{t}}}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r-1)}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r-1)}\right)^{\mathrm{t}}\boldsymbol{D}_k^{(r-1)}\right]_{m,m} \tag{5.16}$$

Étape variationnelle de maximisation

Le modèle avec dépendance spatiale diffère du modèle sans dépendance spatiale uniquement au niveau de la distribution des variables de classes $\boldsymbol{Z}$. Ainsi, les distributions des poids $\boldsymbol{W}$ et des observations $\boldsymbol{y}$ étant exprimées conditionnellement à $\boldsymbol{Z}$, l'inférence des paramètres associés, intervenant lors de l'étape de maximisation, est identique entre les modèles avec et sans dépendance spatiale. Seuls les paramètres liés à $\boldsymbol{Z}$ sont inférés différemment du fait que nous ayons remplacé une distribution multinomiale commune à toutes les classes par une distribution de Gibbs.

Étape variationnelle M-$(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})$

En posant :

$$\widetilde{\Delta}_{i,k}^{(r)} = \mathrm{E}_{q_{\boldsymbol{W}_i \mid Z_i=k}^{(r)}}(\Delta_{\boldsymbol{W}_i}) = \mathrm{diag}\left(\frac{\widetilde{\gamma}_{k,1}^{(r)}}{\widetilde{\delta}_{i,k,1}^{(r)}}, \ldots, \frac{\widetilde{\gamma}_{k,\mathrm{M}}^{(r)}}{\widetilde{\delta}_{i,k,\mathrm{M}}^{(r)}}\right) \tag{5.17}$$

nous retrouvons, suite aux calculs en annexe B.2, les estimations par maximum de vraisemblance présentés dans [7] :

$$\mu_{k,m}^{(r)} = \frac{\sum_{i=1}^{\mathrm{N}} \widetilde{\pi}_{i,k}^{(r)}\left[\boldsymbol{D}_k^{(r-1)}\widetilde{\Delta}_{\boldsymbol{W}_i,k}^{(r)}\boldsymbol{D}_k^{(r-1)^{\mathrm{t}}}\boldsymbol{y}_i\right]_m}{\sum_{j=1}^{\mathrm{N}} \widetilde{\pi}_{j,k}^{(r)}\left[\widetilde{\Delta}_{j,k}^{(r)}\right]_{m,m}} \text{ , pour } m = 1, \ldots, \mathrm{M}$$

$$\boldsymbol{D}_k^{(r)} = \arg\min_{\boldsymbol{D}_k} \sum_{i=1}^{\mathrm{N}} \widetilde{\pi}_{i,k}^{(r)} \ \mathrm{tr}\left[\boldsymbol{D}_k\widetilde{\boldsymbol{\Delta}}_{i,k}^{(r)}\boldsymbol{A}_k^{(r-1)}\boldsymbol{D}_k^{\mathrm{t}}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r)}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r)}\right)^{\mathrm{t}}\right]$$

la forme ainsi obtenue peut être résolue par l'algorithme ALS ([30]).

$$A_{k,m}^{(r)} = \frac{\sum_{j=1}^{\mathrm{N}} \widetilde{\pi}_{j,k}^{(r)}}{\sum_{i=1}^{\mathrm{N}} \widetilde{\pi}_{i,k}^{(r)}\left[\widetilde{\Delta}_{i,k}^{(r)}\right]_{m,m}\left[\boldsymbol{D}_k^{(r)^{\mathrm{t}}}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r)}\right)\right]_m^2} \text{ , pour } m = 1, \ldots, \mathrm{M}$$

Étape variationnelle M-$\boldsymbol{\alpha}$

En réécrivant l'équation (5.10) pour faire apparaître la maximisation des degrés de liberté dimension par dimension, annexe B.2, nous obtenons la même formulation que [7], ce qui permet d'écrire l'expression de $\alpha_{k,m}^{(r)}$ comme solution de l'équation suivante :

$$\Upsilon\left(\alpha_{k,m}\right) = \Upsilon\left(\widetilde{\gamma}_{k,m}^{(r)}\right) - \frac{\sum_{i=1}^{\mathrm{N}} \widetilde{\pi}_{i,k}^{(r)} \ln\left(\widetilde{\delta}_{i,k,m}^{(r)}\right)}{\sum_{j=1}^{\mathrm{N}} \widetilde{\pi}_{j,k}^{(r)}} \tag{5.18}$$

en notant Υ la fonction digamma (dérivée du logarithme de la fonction gamma). Celle-ci n'ayant pas d'inverse sous forme explicite, nous résolvons cette équation par un optimisation numérique de type méthode de Newton ou de Halley, en utilisant respectivement la dérivée première et seconde de la fonction digamma.

Étape variationnelle M-$(\boldsymbol{\tau}, \beta)$

L'estimation des paramètres de la distribution de Gibbs n'admettant pas d'expression explicite, nous réalisons cette estimation à nouveau par résolution numérique via un méthode de descente de gradient. Notons $\nabla_{(\boldsymbol{\tau},\beta)}$ le gradient de la densité de Gibbs (équation 5.2) par rapport aux paramètres $\boldsymbol{\tau}$ et β, ainsi que $\nabla^2_{(\boldsymbol{\tau},\beta)}$ la matrice hessienne. Nous avons les relations suivantes :

$$\begin{aligned}
\nabla_{(\boldsymbol{\tau},\beta)} \mathrm{E}_{q_{\boldsymbol{Z}}^{(r)}}\left[\ln p\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)\right] &= \mathrm{E}_{q_{\boldsymbol{Z}}^{(r)}}\left[\nabla_{(\boldsymbol{\tau},\beta)} H\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)\right] \\
&- \mathrm{E}_{p(\boldsymbol{z}\ ;\ \boldsymbol{\tau},\beta)}\left[\nabla_{(\boldsymbol{\tau},\beta)} H\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)\right] \qquad (5.19) \\
\nabla^2_{(\boldsymbol{\tau},\beta)} \mathrm{E}_{q_{\boldsymbol{Z}}^{(r)}}\left[\ln p\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)\right] &= \mathrm{E}_{q_{\boldsymbol{Z}}^{(r)}}\left[\nabla^2_{(\boldsymbol{\tau},\beta)} H\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)\right] \\
&- \mathrm{E}_{p(\boldsymbol{z}\ ;\ \boldsymbol{\tau},\beta)}\left[\nabla^2_{(\boldsymbol{\tau},\beta)} H\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)\right] \\
&- \mathrm{var}_{p(\boldsymbol{z}\ ;\ \boldsymbol{\tau},\beta)}\left[\nabla_{(\boldsymbol{\tau},\beta)} H\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)\right] . \qquad (5.20)
\end{aligned}$$

À noter que la dernière espérance dans la formule (5.19) ainsi que la variance formule (5.20) sont calculées par rapport à la distribution de Gibbs telle que défini dans la formule (5.2).

La fonction énergie $H\left(\boldsymbol{z}\ ;\ (\boldsymbol{\tau}, \beta)\right)$ est linéaire en les paramètres $(\boldsymbol{\tau}, \beta)$, ce qui entraîne que les termes en $\nabla^2_{(\boldsymbol{\tau},\beta)} H\left(\boldsymbol{Z}\ ;\ \boldsymbol{\tau}, \beta\right)$ sont nuls et donc que la matrice hessienne est semi-négative. Nous sommes donc ramenés à chercher le maximum d'une fonction concave.

Cependant, la constante de normalisation $\mathcal{K}$ rend les formules (5.19) et (5.20) non explicites. Pour contourner cette difficulté, nous remplaçons la distribution de Gibbs $p_G\left(\boldsymbol{z}\ ;\ \boldsymbol{\tau}, \beta\right)$ utilisée dans le modèle par une approximation dont la forme est induite par l'approximation variationnelle de la distribution a posteriori des

variables de classes, formule (5.12). Nous obtenons alors l'approximation suivante :

$$q_{\boldsymbol{Z}}^{prior}\left(\boldsymbol{z}\ ;\ \boldsymbol{\tau},\beta\right)=\prod_{i=1}^{\mathrm{N}}q_{Z_i}^{prior}\left(z_i\ ;\ \boldsymbol{\tau},\beta\right) \tag{5.21}$$

avec $q_{Z_i}^{prior}\left(z_i\ ;\ \boldsymbol{\tau},\beta\right)$ défini par :

$$\begin{aligned}
&q_{Z_i}^{prior}\left(z_i\ ;\ \boldsymbol{\tau},\beta\right)\\
&\quad\propto\ \mathrm{E}_{\prod_{j\neq i}q_{Z_j}^{(r)}}\left[H\left(Z_1,\ \ldots,\ Z_{i-1},z_i,Z_{i+1},\ \ldots,\ Z_{\mathrm{N}}\right)\right]\\
&\quad\propto\ \mathrm{E}_{\prod_{j\neq i}q_{Z_j}^{(r)}}\left[\tau_{z_i}+\frac{\beta}{2}\sum_{l\in\Omega(i)}\mathbb{I}_{z_i}\left(Z_l\right)+\sum_{j\neq i}\tau_{z_j}+\frac{\beta}{2}\left(\sum_{j\neq i,j\in\Omega(i)}\sum_{l\in\Omega(j)}\mathbb{I}_{Z_j}\left(Z_l\right)+\underbrace{\sum_{j\neq i,j\notin\Omega(i)}\sum_{l\in\Omega(j)}\mathbb{I}_{Z_j}\left(Z_l\right)}_{\text{indépendant de }z_i}\right)\right]\\
&\quad\propto\ \tau_{z_i}+\frac{\beta}{2}\sum_{l\in\Omega(i)}q_{Z_l}^{(r)}\left(z_i\right)+\frac{\beta}{2}\mathrm{E}_{\prod_{j\neq i}q_{Z_j}^{(r)}}\left[\sum_{j\neq i,j\in\Omega(i)}\left(\mathbb{I}_{Z_j}\left(z_i\right)+\underbrace{\sum_{l\in\Omega(j),l\neq i}\mathbb{I}_{Z_j}\left(Z_l\right)}_{\text{indépendant de }z_i}\right)\right]\\
&\quad\propto\ \tau_{z_i}+\frac{\beta}{2}\sum_{l\in\Omega(i)}q_{Z_l}^{(r)}\left(z_i\right)+\frac{\beta}{2}\sum_{j\in\Omega(i)}\mathrm{E}_{q_{Z_j}^{(r)}}\left[\mathbb{I}_{Z_j}\left(z_i\right)\right]\\
&\quad\propto\ \tau_{z_i}+\frac{\beta}{2}\sum_{l\in\Omega(i)}q_{Z_l}^{(r)}\left(z_i\right)+\frac{\beta}{2}\sum_{j\in\Omega(i)}q_{Z_j}^{(r)}\left(z_i\right)=\tau_{z_i}+\beta\sum_{l\in\Omega(i)}q_{Z_l}^{(r)}\left(z_i\right)
\end{aligned}$$

Ce qui aboutit à l'expression suivante :

$$q_{Z_i}^{prior}\left(k\ ;\ \boldsymbol{\tau},\beta\right)=\frac{\exp\left[\tau_k+\beta\sum_{l\in\Omega(i)}q_{Z_l}^{(r)}\left(k\right)\right]}{\sum_{j=1}^{\mathrm{K}}\ \exp\left[\tau_j+\beta\sum_{l\in\Omega(i)}q_{Z_l}^{(r)}\left(j\right)\right]}=\widetilde{\pi}_{i,k}^{prior} \tag{5.22}$$

Cette approximation de la distribution de Gibbs basée sur l'approximation de la loi a posteriori a été proposée par [28] et utilisée également par [31].

Finalement, nous approchons le gradient en (5.19) par résolution numérique du

système à $K+1$ équations suivant :

$$\begin{cases} \sum_{i=1}^{\mathrm{N}} \sum_{k=1}^{\mathrm{K}} \left[\widetilde{\pi}_{i,k}^{(r)} \sum_{l \in \Omega(i)} \widetilde{\pi}_{l,k}^{(r)} - \widetilde{\pi}_{i,k}^{prior} \sum_{l \in \Omega(i)} \widetilde{\pi}_{l,k}^{prior} \right] \\ \sum_{i=1}^{\mathrm{N}} \left(\widetilde{\pi}_{1,k}^{(r)} - \widetilde{\pi}_{i,1}^{prior} \right) \\ \vdots \\ \sum_{i=1}^{\mathrm{N}} \left(\widetilde{\pi}_{i,\mathrm{K}}^{(r)} - \widetilde{\pi}_{i,\mathrm{K}}^{prior} \right) \end{cases}$$

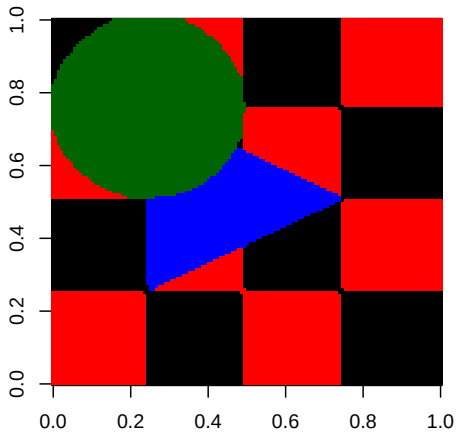
5.3 Résultats numériques

Application sur données simulées

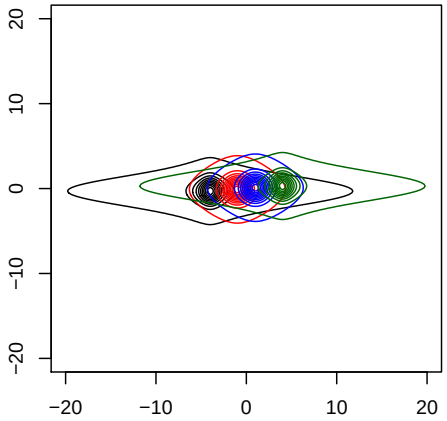
Nous avons d'abord testé l'impact de l'ajout du champ de Markov sur des données simulées issues d'un mélange de 4 lois de Pearson type VII à mélange d'échelles multiples en dimension 2. Pour cela, nous avons utilisé une image carrée de 128 pixels de côté, Figure 5.1-A, chaque pixel étant étiqueté de 1 à 4 pour représenter son appartenance à une classe. Nous avons ainsi les classes 1 (en noir) et 2 (en rouge) qui forment un damier, la classe 3 (en bleu) qui représente un morceau de triangle, et la classe 4 (en vert) qui représente un disque. L'image a ensuite été bruitée en simulant en chaque voxel une observation issue du modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples (MMSP) en fonction de l'étiquette du voxel : les classes sont décalées de centres $\boldsymbol{\mu}_k = [-4, -0.3]$, $[-1, -0.1]$, $[1, 0.1]$ et $[4, 0.3]$, la matrice d'échelle est commune fixée à l'identité $\boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T = [1, 0; 0, 1]$, et les degrés de liberté sont choisis pour obtenir une large superposition des classes sur la première dimension $\boldsymbol{\alpha}_k = [2, 100]$, $[10, 100]$, $[10, 100]$ et $[2, 100]$. La Figure 5.1-B montre les niveaux de densité de chacune des 4 lois de Pearson type VII à mélange d'échelles multiples et en particulier les différentes dispersions des distributions en fonction des degrés de liberté choisis. Nous avons ensuite estimé un modèle de mélange à 4 lois de Pearson type VII à mélange d'échelles multiples couplé à un champ de Markov latent à partir des images bruitées obtenues, Figure 5.1-C et Figure 5.1-D. Le champ de Markov est défini sur un voisinage des 8 plus proches voisins en 2 dimensions.

Pour des raisons de temps de calcul, le paramètre β de la distribution de Gibbs n'est pas estimé dans l'algorithme implémenté, et est considéré actuellement comme un paramètre à régler par l'utilisateur. Après une segmentation initiale obtenue à l'aide d'un mélange gaussien à 4 composantes [2], Figure 5.2-A, nous avons fait varier le paramètre β de 0 à 10 avec un pas de 1. Sans interaction locale ($\beta = 0$), Figure 5.2-B, l'image obtenue reste bruitée. En augmentant le paramètre β à 1, Figure 5.2-C, les classes deviennent spatialement plus homogènes. Il néces-

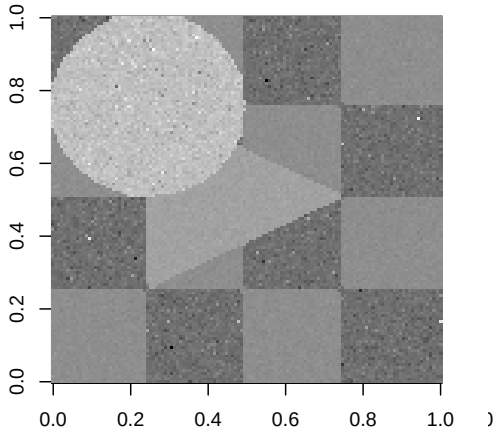
2. Le mélange gaussien est estimé en R grâce au paquet mclust [22].



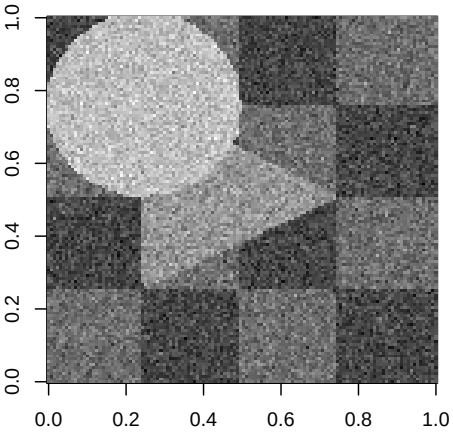
(A) Image utilisée pour la répartition spatiale des observations.



(B) Lignes de niveau des 4 lois de Pearson type VII à mélange d'échelles multiples constituant le mélange utilisé.



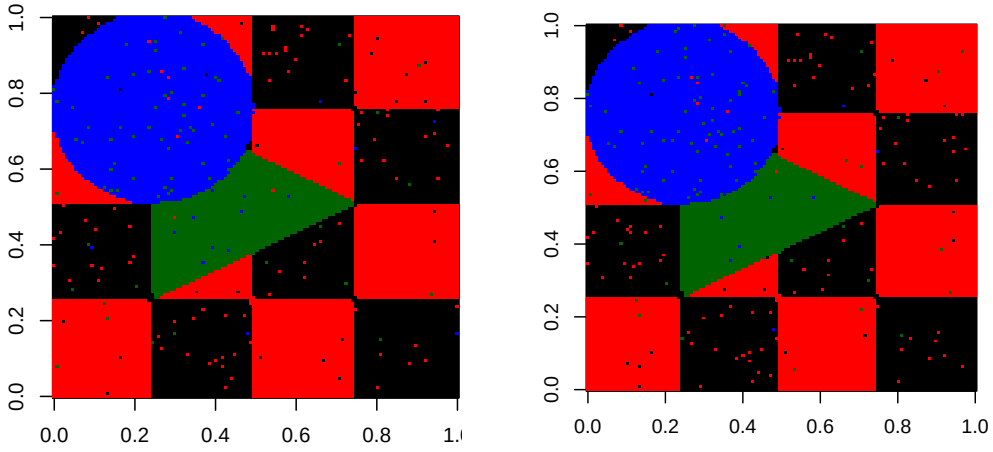
(C) Image bruitée pour la première dimension.



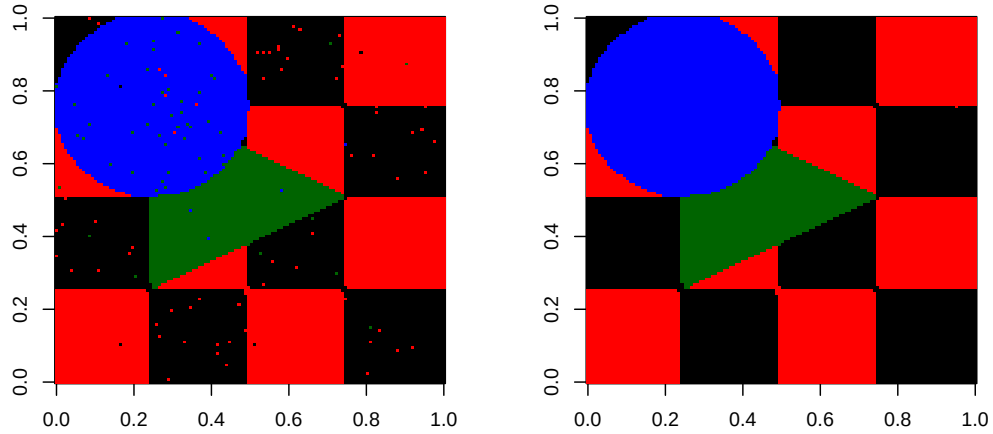
(D) Image bruitée pour la seconde dimension.

FIGURE 5.1 – Simulation d'un mélange de 4 lois de Pearson type VII à mélange d'échelles multiples en dimension 2 sur une image de taille 128 par 128.

saire d'augmenter β jusqu'à 8, Figure 5.2-D, pour retrouver les classes entièrement homogènes de la simulation. Nous retrouvons ainsi le rôle classique du paramètre d'interaction local β qui régule l'appartenance d'une observation à une classe en faisant un compromis entre la proximité de l'observation $\boldsymbol{y}$ au centre de la classe, dans l'espace des variables, et la concordance de la classe z associée à l'observation $\boldsymbol{y}$ avec celle des voisins de $\boldsymbol{y}$.



(A) Initialisation. (B) $\beta = 0$



(C) $\beta = 1$ (D) $\beta = 8$

FIGURE 5.2 – (A) Initialisation du mélange avec un modèle de mélange gaussien à 4 composantes, (B)-(D) Estimation du modèle MMSP avec champ de Markov avec un paramètre d'interaction local $\beta = 0$, 1 et 8.

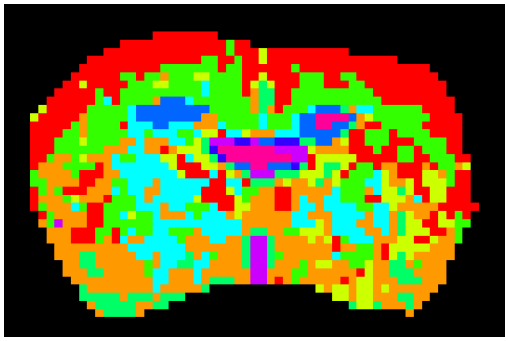
Application sur données réelles

Nous travaillons actuellement à l'application du modèle MMSP avec champ de Markov aux données d'IRM cérébrale de rats avec ou sans tumeurs, tel que présenté dans le chapitre 3. Une des difficultés majeures rencontrées réside dans le temps de calcul particulièrement long pour l'estimation du mélange avec dépendance spatiale. De ce fait, nous ne présentons ici qu'un résultat préliminaire de comparaison des modèles avec et sans champ de Markov sur données réelles. À noter également que la dépendance spatiale est appliquée coupe par coupe car l'acquisition IRM n'a été réalisée que sur 5 coupes.

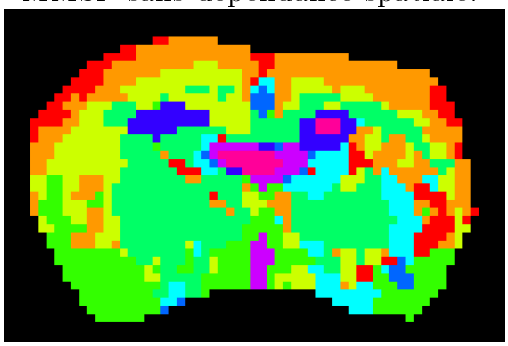
Sans dépendance spatiale, Figure 5.3-A, la classification obtenue grâce à un mélange de 10 lois de Pearson type VII à mélange d'échelles multiples présente une certaine cohérence spatiale avec par exemple le cortex en rouge, les ventricules et vaisseaux sanguins en bleu et mauve. Toutefois, la classification reste pixélisée avec des voxels isolés (leur classe est différente de celle de leurs voisins), ainsi que des frontières entre classes très morcelées. En ajoutant une interaction locale avec un $\beta = 1$, Figure 5.3-B, la cohérence spatiale est plus facilement identifiable, les groupes de voxels sont plus homogènes avec moins de voxels isolés et des frontières entre classes plus lisses. Lorsque l'interaction locale devient trop forte, par exemple $\beta = 10$ Figure 5.3-C, la cohérence spatiale devient très discutable, et nous retrouvons que la classification a été guidée presque uniquement par la concordance locale des classes, ce qui fait disparaître un grand nombre de détails présents dans l'image, tels que les vaisseaux sanguins au bas de la coupe (classe mauve dans les deux cas précédents).

5.4 Discussion et conclusion

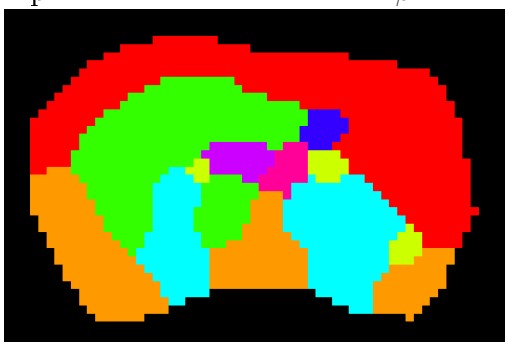
Nous avons étendu dans ce chapitre le modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples au cas de données présentant de la dépendance spatiale au travers de l'ajout d'un champ de Markov. Ce dernier se traduisant par une distribution de Gibbs sur les variables latentes de classe, une estimation simple du modèle a été proposée en dérivant un algorithme EM à partir de celui développé par [7]. L'algorithme EM proposé repose sur une approximation variationnelle de la distribution a posteriori des variables latentes, du fait de l'utilisation de la distribution de Gibbs dans le modèle. Nous avons pu vérifier sur des données simulées que l'interaction locale présente au sein de la distribution de Gibbs permet de prendre en compte la répartition spatiale des données, et peut contrebalancer la distribution des données dans l'espace des variables observées ; comme c'est le cas pour une interaction locale très élevée qui rend prépondérante la position spatiale des données. Concernant les données d'IRM cérébrale de rats utilisées dans le chapitre 3, l'ajout du champ de Markov sur un voisinage basé sur les 8 plus proches voisins montre un net lissage des classes qui augmente la lisibilité et la cohérence spatiale de la classification. Toutefois le choix du paramètre d'interaction locale reste primordial dans la classification obtenue, et son estimation automatique est la prochaine étape dans nos applications. Il sera également intéressant de tester l'influence du voisinage sur les classifications obtenues, en considérant par exemple les 4 plus proches voisins ou encore un voisinage 3D ou de dimensions supérieures. L'optimisation de l'implémentation numérique est aussi un point majeur de la suite de nos travaux, en particulier afin de pouvoir évaluer en temps raisonnable notre modèle sur des données IRM issues de patients



(A) Classification avec le modèle MMSP sans dépendance spatiale.



(B) Classification avec dépendance spatiale : interaction locale $\beta = 1$.



(C) Classification avec une forte dépendance spatiale : interaction locale $\beta = 10$.

FIGURE 5.3 – Classification des données d'IRM cérébrales de rats issue d'un mélange, avec ou sans champ de Markov, de 10 lois de Pearson type VII à mélange d'échelles multiples. La coupe présentée correspond à la coupe centrale d'un rat sain.

humains, pour lesquels le volume de données est d'un ordre de grandeur supérieur

à celui des données d'IRM de rats.

CHAPITRE 6

DÉVELOPPEMENTS INFORMATIQUES

Dans ce chapitre nous présentons les outils numériques mis au point afin de réaliser l'estimation des différents modèles statistiques élaborés, et leurs applications aux données d'IRM cérébrales fournies par le GIN.

Paquet R pour l'estimation du modèle de mélange de lois de Student à mélange d'échelles multiples

En collaboration avec Stéphane Despreaux [1], nous avons développé un paquet R pour réaliser l'estimation du modèle de mélange de lois de Student à mélange d'échelles multiples au cours des deux premières années de thèse. Nommé Mixture of Multiple Scaled Distributions (MMSD), ce paquet est disponible à l'adresse :

`http://www-ljk.imag.fr/membres/Stephane.Despreaux/MMSD/Download/0.7/`

Le cœur du code a été réalisé en C++ avec une interface avec R grâce au paquet Rcpp [32]. Ce paquet reprend le code R préalablement développé par [7], en y apportant des optimisations fines en terme de complexité algorithmique et de gestion de la mémoire. Le paquet possède également une version parallèle via l'utilisation d'OpenMP [2] en parallélisant le code sur le nombre de classes.

Scripts R pour le protocole d'analyse d'IRM

Le protocole d'analyse d'IRM présenté dans le chapitre 3 a été élaboré tout au long de la thèse sous forme de scripts R de façon à gérer entre autre l'importation et le pré-traitement des données, la détection d'anomalies, l'apprentissage du modèle de signatures, ou encore le post-traitement spatial. Le paquet MMSD est un

1. Ingénieur de Recherche CNRS au Laboratoire Jean Kuntzmann
2. https ://www.openmp.org/

élément clé de ce protocole puisqu'il permet l'estimation des différents modèles de mélange présents au sein du protocole. Tout au long du protocole, des rapports sont générés automatiquement via le paquet R Knitr [33] de façon à contrôler l'ajustement des modèles statistiques et à sauvegarder les résultats produits.

Une conversion des données de segmentation au format NIfTI [3] a été intégrée au protocole grâce à Véronica Munoz Ramirez [4] afin de pouvoir utiliser des outils classiques de visualisation en IRM tels que itk-SNAP [34]. De plus, l'ensemble du protocole peut être déployé sur une grille de calculs utilisant OAR [5] comme gestionnaire de ressources et de tâches.

Code C++ pour l'estimation des extensions bayésienne et markovienne du modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples

Les extensions bayésienne (chapitre 4) et markovienne (chapitre 5) du modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples ont été implémentées directement en C++ avec également une interface avec R via les paquets Rcpp [32] et RcppArmadillo [35], ce dernier permettant de faciliter l'écriture des calculs d'algèbre linéaire. Le développement du code bayésien a débuté lors du projet de recherche à l'Université McGill avec le professeur Russell Steele, et s'est poursuivit ensuite en collaboration avec Steven Quinito Masnada [6] au cours de la quatrième année de thèse, avec qui nous avons également développé le code markovien. Ces deux codes étant proches, un des objectifs à moyen terme est de les fusionner et de les incorporer dans le protocole d'analyse développé. Pour le moment, seul le code markovien est en cours d'intégration avec le protocole.

3. https ://nifti.nimh.nih.gov/
4. Doctorante de l'équipe Neuroimagerie Fonctionnelle et Perfusion Cérébrale au GIN.
5. http ://oar.imag.fr/
6. Ingénieur au sein de l'équipe MISTIS à INRIA.

CHAPITRE 7

CONCLUSIONS

7.1 Apports de la thèse

Tout au long de la thèse, nous avons eu en fil conducteur l'objectif de développer un nouvel outil automatique d'aide au diagnostic en exploitant les modèles statistiques de mélange appliqués à l'imagerie médicale.

Nous avons en premier lieu développé un protocole entièrement automatique et générique permettant la localisation et la caractérisation conjointe de tissus non physiologiques. À partir de données de référence, le protocole réalise l'apprentissage d'un premier modèle de mélange représentatif des données de référence (sujets sains), modèle qui est ensuite utilisé afin de détecter les anomalies présentes dans un second jeu de données (sujets atteints de pathologies). Une fois les anomalies localisées, le protocole utilise un deuxième modèle de mélange afin d'ajuster un modèle représentatif des types tissulaires constituant les anomalies. À partir des proportions de ce mélange, le protocole construit un modèle de signatures d'anomalies. L'ajustement des différents modèles de mélange se fait en utilisant l'heuristique de pente de [20] afin de déterminer automatiquement le nombre de composantes des mélanges, nombre de composantes généralement fixé par l'utilisateur, comme dans [18], [13], [19], ou basé sur la maximisation du critère BIC, comme [16], critère délicat à exploiter lorsque le maximum se situe sur un plateau. De même, l'heuristique de pente utilisée permet de déterminer automatiquement les différents niveaux d'anormalité présents dans les deux jeux de données, avant de déterminer la meilleure séparation de ces données en deux sous-populations : anormale contre non-anormale. Nous avons validé le concept de ce protocole sur des données d'IRM de cerveaux de rats sains et porteurs de différentes tumeurs au

travers d'un article publié dans le journal IEEE Transactions on Medical Imaging. L'utilisation de mélanges de lois de Student à échelles multiples, dont l'une des particularités est de tolérer des niveaux de valeurs extrêmes différents par dimension, a permis d'améliorer les localisations (AIR plus élevé de 6.4%) et les taux de bonne caractérisation des anomalies (taux plus élevé de 5.6%) par rapport à l'utilisation plus classique des lois Normales.

D'un point de vue biologique, nous avons constaté une variabilité importante au sein des modèles de tumeurs, comme par exemple la tumeur C6 dont la signature d'anomalie fluctue fortement d'un individu à l'autre (cf. Figure (8), de l'article TMI au chapitre 3). La question du contrôle qualité des données utilisées en entrée est ainsi remise en avant. La variabilité observée peut être dû à la simple variabilité biologique des tumeurs C6 qui, étant instables, présentent une forte hétérogénéité. Ou bien, il s'agit d'une variabilité imputable à l'acquisition IRM, et il est alors nécessaire de ne conserver que des rats présentant des développements homogènes de la tumeur C6. Dans les deux cas, il faut augmenter la taille de l'échantillon des rats porteurs de tumeurs C6 afin de s'assurer soit de l'homogénéité des groupes, soit d'une couverture suffisante de la variabilité biologique constatée. Ce qui est valable à cette échelle de validation sur petit animal l'est d'autant plus lors de la constitution de groupes de patients lors d'essais pharmacologiques préclinique.

Bien que la validation de notre protocole ait été réalisée grâce à un échantillon de 32 rats pour la partie apprentissage, et 21 rats pour la partie d'évaluation, soit un volume de donnés plus important que dans les études précédentes telles que [17], [13], [16], il s'agit d'un petit échantillon d'un point de vue statistique. Le passage à des données d'IRM chez l'homme sera un moyen de valider définitivement les concepts du protocole, mais cela pose également le problème du temps de calcul pour de très larges quantités de données. Une réponse possible est proposée dans les perspectives méthodologiques de la partie suivante.

Nous avons également cherché à améliorer la brique statistique au cœur du protocole développé qui est l'estimation des modèles de mélange, et plus particulièrement le choix du nombre de composantes à considérer. En se basant sur les résultats théoriques de [24], nous avons transcrit le modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples dans un cadre bayésien au travers de deux extensions : la première sans mettre de loi a priori sur les proportions du mélange, la seconde avec un a priori de Dirichlet. L'idée commune de ces modèles est d'initialiser le modèle de mélange avec plus de composantes que nécessaire, afin d'exploiter une caractéristique asymptotique de l'algorithme EM. Ce dernier, sous certaines conditions de régularité, vide à la convergence les classes surnuméraires, ce qui permet d'obtenir un algorithme d'estimation de mélange avec une sélection conjointe du nombre de composantes. Pour les deux extensions proposées,

nous avons présenté deux stratégies d'accélération de l'algorithme EM. Les tests effectués sur données simulées et données réelles montrent une meilleure efficacité (résultats plus robustes) et un temps de calcul raccourci (facteur 3 à 5 possible) de notre stratégie d'accélération par rapport à l'approche plus classique de sélection par maximisation du critère BIC.

Il sera nécessaire de valider les différentes stratégies de sélection de modèles sur une plus large variété de données simulées, entre autre pour vérifier les résultats obtenus lorsque la dimension des données augmente, ce qui revient à considérer plus de cartes IRM dans nos applications. Une application aux données d'IRM de cerveaux de rats et d'hommes est également prévue, avec notamment la comparaison par rapport aux critères de choix de modèles actuellement utilisés, tels que le critère BIC et l'heuristique de pente.

Enfin, nous avons étendu le modèle de mélange de lois de Pearson type VII à mélange d'échelles multiples pour la prise en compte de la dépendance spatiale. Pour cela, nous avons couplé le modèle de mélange avec un champ de Markov latent de façon à ce que deux observations spatialement proches influent l'une sur l'autre quand à leur classification. Cette influence réciproque est définie en utilisant une distribution de Gibbs sur les variables latentes de classe, via l'utilisation d'un paramètre de champ externe et d'un paramètre d'interaction locale. Ce dernier paramètre permet de pondérer la part d'information issue de la structure spatiale et celle issue des variables mesurées. Les premiers tests sur données simulées montrent le comportement habituel des champs de Markov : l'ajout du paramètre d'interaction locale permet de lisser la classification obtenue.

Une validation de l'impact de cet ajout dans le protocole développé est maintenant nécessaire, notamment afin de déterminer si cela améliore seulement la localisation des anomalies ou également la caractérisation de ces dernières. Il reste aussi à automatiser l'estimation du paramètre d'interaction locale, car ce dernier est pour le moment fixé par l'utilisateur, sachant que cette automatisation a un coût non négligeable en terme de temps de calcul.

7.2 Perspectives

De nombreuses perspectives possibles ont été relevées au cours de la thèse. Nous détaillerons en premier celles d'ordre méthodologique, avant d'indiquer quelques perspectives quant à d'autres applications médicales envisagées.

Aspects méthodologiques.

Le protocole développé ayant mis en avant une homogénéité inégale entre les groupes de rats utilisés pour les tests, il serait pertinent d'ajouter un contrôle

qualité des données avant analyse. Cela permettrait d'exclure en amont des patients atypiques ou au contraire cela justifierait d'augmenter la taille des groupes, et cela afin de couvrir plus efficacement la variabilité biologique de la pathologie considérée. De même, il serrait intéressant d'ajouter une estimation de la certitude du diagnostic posé, c'est-à-dire de la localisation de la zone anormale et du type d'anomalie prédite, cela permettrait en particulier de détecter des patients présentant une anormalité non incluse dans les échantillons d'apprentissage.

La validation du protocole proposé peut être raffinée en exploitant des données histologiques, seules données pouvant servir de vérité terrain, même si elles présentent des difficultés d'exploitation. Leur première utilisation serait pour vérifier la localisation automatique des anomalies : actuellement cette vérification est faite par comparaison avec une délinéation manuelle sur l'IRM anatomique, or celle-ci varie d'un expérimentateur à l'autre de même qu'au cours du temps pour un même expérimentateur. Une seconde utilisation serait pour confirmer la concordance du modèle de signatures d'anomalies avec les types tissulaires, en particulier : peut on associer une classe avec un type tissulaire ? Les signatures basées sur les classes prendraient alors un sens biologique plus fort.

Les méthodes de type "deep learning" ont énormément évolué ces dernières années, au point de dominer par exemple le challenge BraTs où il est question de délimiter 4 structures au sein de tumeurs à l'aide des données IRM mises à disposition. Un point actuellement essentiel de ces techniques est l'étiquetage de toutes les données d'apprentissage, de façon à pouvoir en extraire des résumés discriminants. Or, notre protocole se base sur le fait qu'au contraire l'information à connaître a priori est minimale : uniquement le type de chaque patient, et non le type de chaque voxel. Un couplage des deux approches pourrait être envisagé : le protocole développé faisant l'étiquetage automatique tandis qu'une méthode de type "deep learning" réalise la caractérisation des anomalies. Cependant, on risque alors de perdre l'interprétabilité des signatures actuellement fournies par notre protocole ; dans ce cas l'intérêt des méthodes de type "deep learning", lorsqu'elles sont plus performantes, serait de mettre en évidence la marge de progression encore possible des modèles statistiques considérés.

Dans cette thèse, nous avons cherché à obtenir systématiquement des modèles plus efficaces que ceux déjà testés, par exemple en utilisant le modèle de mélange de lois de Student à mélange d'échelles multiples au lieu du mélange gaussien. Une vision alternative est non pas de considérer uniquement le meilleur modèle individuel, mais au contraire de considérer le résultat de tous les modèles développés et de prendre une décision basée sur ces résultats, comme par exemple

avec le vote majoritaire. Cela permet ainsi de combiner les forces des différents modèles. Sur le challenge BraTs, ces méthodes agrégatives surpassent systématiquement les méthodes individuelles. Il se pose alors la question du temps de calcul lorsque l'on souhaite utiliser conjointement plusieurs algorithmes itératifs, l'algorithme développé pour l'estimation d'un mélange de lois de Pearson type VII à mélange d'échelles multiples couplé avec un champ de Markov est à lui seul particulièrement lent.

Concernant les temps de calcul, tous les tests réalisés au cours de la thèse ont été fait sur de petits échantillons au vue des volumes collectés quotidiennement sur une plate-forme IRM. Il est envisageable de vouloir évaluer le protocole développé sur une cohorte de 2000 patients, par exemple afin de proposer de nouveaux biomarqueurs de tumeurs ou de nouvelles cartes IRM. Cependant, le passage à l'humain est toujours en cours de calibration est pose déjà des problèmes de gestion de la mémoire des ordinateurs pour une dizaine de patients. Augmenter de 2 ordres de grandeur le nombre de patients nécessiterait de revoir d'un, l'implémentation du protocole, notamment par une optimisation plus fine, et de deux, la structure même des algorithmes EM utilisés afin d'utiliser des variantes convergeant plus rapidement.

Chacune des perspectives précédemment évoquées va comporter son lot de tests, sur données simulées et données réelles, afin d'évaluer les performances respectives et déterminer s'il y a bien une amélioration. Il serait bien d'établir un protocole de test qui serait commun à tous ces développements de façon à pouvoir positionner facilement ces méthodes. À l'instar du challenge BraTs, il est nécessaire de former une base de données publiques respectant un formalisme donné de façon à permettre l'ajout de nouvelles données au cours du temps. Il faut également définir des métriques, telle que le Dice ou l'ARI, pour comparer la qualité et la concordance des localisations automatiques, si possible via l'utilisation de données histologiques. Pour la caractérisation automatique des signatures, le taux de bonnes prédictions du type d'anomalie est évidemment un critère de première importance, mais la concordance avec le type tissulaire gagnerait aussi à être incorporée par exemple.

Enfin, toute une réflexion, en partenariat avec des radiologues, sur la conception d'une interface radiologue/algorithme est à développer, afin d'intégrer les outils statistiques mis au point dans le quotidien du radiologue. Le résultat seul est-il suffisant s'il est accompagné d'intervalle de confiance ? Quels résultats intermédiaires sont à rendre accessibles ? Par exemple, la carte d'anomalie du chapitre 3 permet de mettre en évidence l'ampleur des altérations dues à une tumeur.

Applications Médicales

Deux applications du protocole développé sont actuellement en cours d'évaluation sur des données IRM chez l'homme : une portant sur des tumeurs cérébrales, l'autre sur la maladie de Parkinson. Le premier objectif de ces essais est d'évaluer l'apport des différentes cartes IRM dans la localisation et la prédiction des pathologies, notamment pour optimiser les séquences d'acquisition IRM en fournissant des recommandations quant aux paramètres IRM à considérer. À plus long terme, un second objectif est d'apporter une aide dans la prise en charge des traitements, en fournissant un suivi au cours du temps de l'évolution des pathologies, ainsi que sur de possibles développements de celles-ci.

Pour finir, deux autres applications sont envisagées sur l'accident vasculaire et le traumatisme crânien, pour lesquels nous avons à la fois des données sur le petit animal et chez l'homme.

BIBLIOGRAPHIE

[1] L. M. DeAngelis, "Brain Tumors", *New England Journal of Medicine*, t. 344, n° 2, p. 114-123, janvier 2001.

[2] A. Drevelegas et N. Papanikolaou, "Imaging of Brain Tumors with Histological Correlations", in, A. Drevelegas, éd. Springer Berlin Heidelberg, 2011, chap. Imaging Modalities in Brain Tumors, p. 13-33.

[3] P. Y. Wen, D. R. Macdonald, D. A. Reardon, T. F. Cloughesy, A. G. Sorensen, E. Galanis, J. DeGroot, W. Wick, M. R. Gilbert, A. B. Lassman, C. Tsien, T. Mikkelsen, E. T. Wong, M. C. Chamberlain, R. Stupp, K. R. Lamborn, M. A. Vogelbaum, M. J. van den Bent et S. M. Chang, "Updated Response Assessment Criteria for High-Grade Gliomas : Response Assessment in Neuro-Oncology Working Group", *Journal of Clinical Oncology*, t. 28, n° 11, p. 1963-1972, avril 2010.

[4] J. Simon, D. Li, A. Traboulsee, P. Coyle, D. Arnold, F. Barkhof, J. Frank, R. Grossman, D. Paty, E. Radue et J. Wolinsky, "Standardized MR Imaging Protocol for Multiple Sclerosis : Consortium of MS Centers Consensus Guidelines", *American Journal of Neuroradiology*, t. 27, n° 2, p. 455-461, 2006.

[5] A. Sottoriva, I. Spiteri, S. G. M. Piccirillo, A. Touloumis, V. P. Collins, J. C. Marioni, C. Curtis, C. Watts et S. Tavaré, "Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics", *Proceedings of the National Academy of Sciences of the United States of America*, t. 110, n° 10, p. 4009-4014, février 2013.

[6] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe et J. Verweij, "New response evaluation criteria in solid tumours : Revised RECIST gui-

deline (version 1.1)", *European Journal of Cancer*, t. 45, n° 2, p. 228-247, 2009.

[7] F. Forbes et D. Wraith, "A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights : Application to robust clustering", *Statistics and Computing*, t. 24, n° 6, p. 971-984, 2014.

[8] C. Maugis et B. Michel, "Slope heuristics for variable selection and clustering via Gaussian mixtures", Inria, Research, juin 2008.

[9] C. Pierpaoli, "Quantitative brain MRI", *Topics in magnetic resonance imaging*, t. 21, p. 63, 2010.

[10] O. Wu, R. M. Dijkhuizen et A. G. Sorensen, "Multiparametric MR Imaging of Brain Disorders", *Topics in Magnetic Resonance Imaging*, t. 21, n° 2, p. 129-138, avril 2010.

[11] A. Waldman, A. Jackson, S. Price, C. Clark, T. Booth, D. Auer, P. Tofts, D. Collins, M. O Leach et J. Rees, "Quantitative imaging biomarkers in neuro-oncology", *Nature Reviews Clinical oncology*, t. 6, p. 445-454, 2009.

[12] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes et K. Van Leemput, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", *IEEE Transactions on Medical Imaging*, t. 34, n° 10, p. 1993-2024, octobre 2015.

[13] J. Juan-Albarracín, E. Fuster-Garcia, J. V. Manjón, M. Robles, F. Aparici, L. Martí-Bonmatí et J. M. García-Gómez, "Automated Glioblastoma Segmentation Based on a Multiparametric Structured Unsupervised Classification", *PLoS ONE*, t. 10, n° 5, p. 1-20, mai 2015.

[14] L. Macyszyn, H. Akbari, J. M. Pisapia, X. Da, M. Attiah, V. Pigrish, Y. Bi, S. Pal, R. V. Davuluri, L. Roccograndi, N. Dahmane, M. Martinez-Lage, G. Biros, R. L. Wolf, M. Bilello, D. M. O'Rourke et C. Davatzikos, "Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques", *Neuro-Oncology*, t. 18, n° 3, p. 1-9, juillet 2015.

[15] E. I. Zacharaki, S. Wang, S. Chawla, D. S. Yoo, R. Wolf, E. R. Melhem et C. Davatzikos, "Classification of Brain Tumor Type and Grade Using MRI Texture and Shape in a Machine Learning Scheme", *Magnetic Resonance in Medicine*, t. 62, n° 6, p. 1609-1618, 2009.

[16] N. Coquery, O. François, B. Lemasson, C. Debacker, R. Farion, C. Rémy et E. L. Barbier, "Microvascular MRI and unsupervised clustering yields histology-resembling images in two rat models of glioma", *Journal of Cerebral Blood Flow & Metabolism*, t. 34, n° 8, p. 1354-1362, mai 2014.

[17] J. K. Boult, M. Borri, A. Jury, S. Popov, G. Box, L. Perryman, S. A. Eccles, C. Jones et S. P. Robinson, "Investigating intracranial tumour growth patterns with multiparametric MRI incorporating Gd-DTPA and USPIO-enhanced imaging", *NMR in Biomedicine*, t. 29, n° 11, p. 1608-1617, 2016.

[18] P. Katiyar, M. R. Divine, U. Kohlhofer, L. Quintanilla-Martinez, B. Schölkopf, B. J. Pichler et J. A. Disselhorst, "A Novel Unsupervised Segmentation Approach Quantifies Tumor Tissue Populations Using Multiparametric MRI : First Results with Histological Validation", *Molecular Imaging and Biology*, t. 19, n° 3, p. 391-397, 2017.

[19] J. H. Rasmussen, M. Nørgaard, A. E. Hansen, I. R. Vogelius, M. C. Aznar, H. H. Johannesen, J. Costa, A. M. E. Engberg, A. Kjær, L. Specht et B. M. Fischer, "Feasibility of Multiparametric Imaging with PET/MR in Head and Neck Squamous Cell Carcinoma", *Journal of Nuclear Medicine*, t. 58, n° 1, p. 69-74, 2017.

[20] J.-P. Baudry, C. Maugis et B. Michel, "Slope heuristics : overview and implementation", *Statistics and Computing*, t. 22, n° 2, p. 455-470, 2012.

[21] W. N. Venables et B. D. Ripley, *Modern Applied Statistics with S*, 4e éd. Springer-Verlag New York, 2002.

[22] L. Scrucca, M. Fop, T. B. Murphy et A. E. Raftery, "mclust 5 : clustering, classification and density estimation using Gaussian finite mixture models", *The R Journal*, t. 8, n° 1, p. 205-233, 2017.

[23] L. Bergé, C. Bouveyron et S. Girard, "HDclassif : An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data", *Journal of Statistical Software*, t. 46, n° 1, p. 1-29, 2012.

[24] J. Rousseau et K. Mengersen, "Asymptotic behaviour of the posterior distribution in overfitted mixture models", *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, t. 73, n° 5, p. 689-710, novembre 2011.

[25] A. Corduneanu et C. M. Bishop, "Variational Bayesian Model Selection for Mixture Distributions", in *Artificial Intelligence and Statistics 2001 : Proceedings of the Eighth International Workshop*, Morgan Kaufmann, 2001, p. 27-34.

[26] M. H. C. Law, M. A. T. Figueiredo et A. K. Jain, "Simultaneous feature selection and clustering using mixture models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, t. 26, n° 9, p. 1154-1166, 2004.

[27] G. Malsiner-Walli, S. Frühwirth-Schnatter et B. Grün, "Model-based clustering based on sparse finite Gaussian mixtures", *Statistics and Computing*, t. 26, n° 1, p. 303-324, janvier 2016.

[28] G. Celeux, F. Forbes et N. Peyrard, "EM procedures using mean field-like approximations for Markov model-based image segmentation", *Pattern recognition*, t. 36, n° 1, p. 131-144, 2003.

[29] J. Blanchet, "Modèles markoviens et extensions pour la classification de données complexes", Thèse, Université Joseph-Fourier - Grenoble I, 2007.

[30] R. Browne et P. McNicholas, "Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models", *Statistics and Computing*, t. Published online, doi :10.1007/s11222-012-9364-2, p. 1-8, 2012.

[31] L. Chaari, T. Vincent, F. Forbes, M. Dojat et P. Ciuciu, "Fast Joint Detection-Estimation of Evoked Brain Activity in Event-Related fMRI Using a Variational Approach", *IEEE Transactions on Medical Imaging*, t. 32, n° 5, p. 821-837, mai 2013.

[32] D. Eddelbuettel et R. François, "Rcpp : Seamless R and C++ Integration", *Journal of Statistical Software*, t. 40, n° 8, p. 1-18, 2011.

[33] Y. Xie, *Dynamic Documents with R and knitr*, 2nd. Boca Raton, Florida : Chapman et Hall/CRC, 2015, ISBN 978-1498716963.

[34] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee et G. Gerig, "User-Guided 3D Active Contour Segmentation of Anatomical Structures : Significantly Improved Efficiency and Reliability", *Neuroimage*, t. 31, n° 3, p. 1116-1128, 2006.

[35] D. Eddelbuettel et C. Sanderson, "RcppArmadillo : Accelerating R with high-performance C++ linear algebra", *Computational Statistics and Data Analysis*, t. 71, p. 1054-1063, mars 2014.

[36] A. A. Taha et A. Hanbury, "Metrics for evaluating 3D medical image segmentation : analysis, selection, and tool", *BMC Medical Imaging*, t. 15, n° 1, p. 29, 2015.

[37] L. R. Dice, "Measures of the amount of ecologic association between species", *Ecology*, t. 26, n° 3, p. 297-302, 1945.

[38] L. Hubert et P. Arabie, "Comparing partitions", *Journal of Classification*, t. 2, n° 1, p. 193-218, 1985.

[39] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods", *Journal of the American Statistical Association*, t. 66, n° 336, p. 846-850, décembre 1971.

[40] F. Forbes et N. Peyrard, "Hidden Markov Random Field Model Selection Criteria Based on Mean Field-Like Approximations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, t. 25, n° 9, p. 1089-1101, 2003.

ANNEXE A

MÉTRIQUES POUR LA COMPARAISON DE SEGMENTATIONS

L'étude de nouvelles cartes IRM ou de nouvelles techniques de classification pour la segmentation s'accompagne d'un besoin de comparer les segmentations obtenues entre elles, de façon à déterminer les cas où une approche est meilleure qu'une autre. Pour cela, une segmentation de référence, souvent obtenue manuellement ou par analyse histologique, est utilisée pour quantifier la concordance d'une autre segmentation avec celle-ci. Plusieurs mesures de concordance sont possibles (Taha and Hanbury [36]), quelques unes sont décrites ici car étant utilisées dans ce manuscrit.

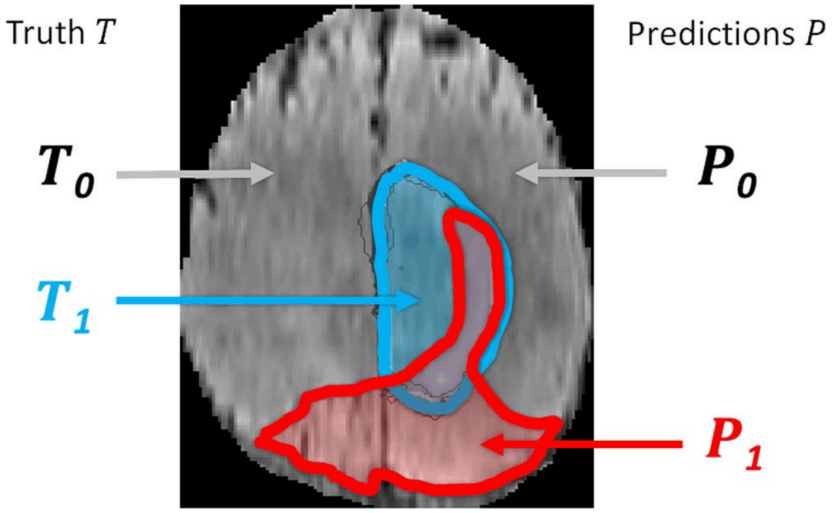


FIGURE A.1 – Figure extraite de Menze et al 2015.

Tout d'abord, il est possible de définir la concordance de deux segmentations en quantifiant le recouvrement des deux partitions obtenues. Pour cela, on utilise un tableau de contingence entre la segmentation manuelle (en ligne, considéré comme la référence) et la segmentation automatique (en colonne, considéré comme la prédiction à évaluer). En considérant que l'objectif est de délimiter une zone d'intérêt (éventuellement non connexe), chaque segmentation admet deux modalités de voxels : ceux non sélectionnés et dits négatifs, et ceux sélectionnés dits positifs. Cela induit quatre groupes de voxels dont on reporte la taille comme suit :

VN : le nombre de voxels négatifs déclarés effectivement comme négatifs
FN : le nombre de voxels positifs déclarés à tort négatifs
FP : le nombre de voxels négatifs déclarés à tort positifs
VP : le nombre de voxels positifs déclarés effectivement comme positifs
n_N : le nombre de voxels négatifs
n_P : le nombre de voxels positifs
n_{pN} : le nombre de voxels prédits comme négatifs
n_{pP} : le nombre de voxels prédits comme positifs

et qui fournit la table suivante :

Tableau A.1 –

		Segmentation automatique		
		Négatif	Positif	*Total*
Segmentation manuelle	Négatif	VN	FP	n_N
	Positif	FN	VP	n_P
	Total	n_{pN}	n_{pP}	n

Il est alors possible de définir des mesures de recouvrement en se basant sur la table de contingence, comme le coefficient de DICE ou l'indice de Rand ajusté.

Le coefficient de DICE (ou score F1 - Dice [37]) mesure la similarité de deux ensembles, ici la similarité entre la segmentation manuelle (vérité de l'expert) et la segmentation automatique produite (prédiction du type des voxels). Il s'agit d'une moyenne pondérée de la précision et la sensibilité de la localisation de la région d'intérêt. Le coefficient de DICE est une valeur entre 0 et 1, plus le coefficient est grand et meilleure est la concordance. Une valeur de 0 signifie qu'aucun voxel positif n'a été détecté ; une valeur de 1 que seuls les voxels effectivement positifs ont été détectés (concordance parfaite).

$$\text{DICE} = \frac{2VP}{2VP + FP + FN} = \frac{2VP}{n_P + n_{pP}} \tag{A.1}$$

L'Indice de Rand Ajusté (Adjusted Rand Index, ARI - Hubert et Arabie [38])

mesure la concordance entre deux partitions (ici les segmentations manuelles et automatiques). Il s'agit d'une mesure dérivée de l'indice de Rand (Rand [39]) qui calcule la concordance comme étant le nombre de paires de voxels co-groupés dans les deux segmentations (c'est-à-dire dont les éléments sont dans une même classe dans chacune des segmentations), plus le nombre de paires de voxels non co-groupés dans les deux segmentations. L'indice de Rand prend ses valeurs entre 0 (concordance sur aucune paire de voxels) et 1 (concordance sur toutes les paires de voxels), les valeurs le plus proche de 1 étant les meilleures. La version ajustée de cet indice est une renormalisation pour avoir des valeurs entre -1 et 1, où la valeur nulle correspond à la valeur attendue dans le cas de segmentations aléatoires issues de distributions hyper-géométriques.

$$\mathrm{ARI} = \frac{\binom{\mathrm{VN}}{2} + \binom{\mathrm{FN}}{2} + \binom{\mathrm{FP}}{2} + \binom{\mathrm{VP}}{2} - \left[\binom{n_N}{2} + \binom{n_P}{2}\right]\left[\binom{n_{pN}}{2} + \binom{n_{pP}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\binom{n_N}{2} + \binom{n_P}{2} + \binom{n_{pN}}{2} + \binom{n_{pP}}{2}\right] - \left[\binom{n_N}{2} + \binom{n_P}{2}\right]\left[\binom{n_{pN}}{2} + \binom{n_{pP}}{2}\right]/\binom{n}{2}} \quad \text{(A.2)}$$

Outre les indices de recouvrement, il est possible de définir la concordance de deux segmentations via l'étude de l'écart entre leurs formes. Par exemple, la distance de Hausdorff est la plus grande distance minimale séparant deux ensembles bornés (ici les segmentations manuelles et automatiques). Cette distance permet d'évaluer le plus grand écart entre les bordures des ensembles considérés. Pour deux ensembles bornés X et Y, la distance de Hausdorff s'exprime via :

$$\begin{aligned} \mathrm{Hausdorff}(X,Y) &= \max\left\{\sup_{y\in Y} d(X,y), \sup_{x\in X} d(x,Y)\right\} && \text{(A.3)} \\ &= \max\left\{\sup_{y\in Y}\inf_{x\in X} d(x,y), \sup_{x\in X}\inf_{y\in Y} d(x,y)\right\} && \text{(A.4)} \end{aligned}$$

avec d la distance entre deux points (par exemple la distance euclidienne).

ANNEXE B

DÉTAILS DE L'INFÉRENCE DU MODÈLE DE MÉLANGE DE LOIS DE PEARSON TYPE VII À MÉLANGE D'ÉCHELLES MULTIPLES COUPLÉ À UN CHAMP DE MARKOV LATENT

B.1 Étape variationnelle E-$(\boldsymbol{W}_i, Z_i)$

Dans la suite, le symbole $\propto$ indique que les équations sont vraies à une constante additive près, constante que nous omettons pour plus de clarté :

$$
\begin{aligned}
& \ln p\left(\boldsymbol{w}_i, z_i \,|\, \boldsymbol{y}_i, \boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) \\
\propto\ & \ln p\left(\boldsymbol{y}_i \,|\, \boldsymbol{W}_i = \boldsymbol{w}_i, Z_i = z_i, \boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) + \ln p\left(\boldsymbol{w}_i \,|\, Z_i = z_i, \boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) \\
& + \ln p\left(z_i \,|\, \boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) \\
\propto\ & \ln p\left(\boldsymbol{y}_i \,|\, \boldsymbol{w}_i, z_i \,;\, \boldsymbol{\phi}^{(r-1)}\right) + \ln p\left(\boldsymbol{w}_i \,|\, z_i \,;\, \boldsymbol{\phi}^{(r-1)}\right) + \ln p\left(z_i \,|\, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) \\
& \text{par indépendance conditionnelle} \\
\propto\ & \sum_{k=1}^{\mathrm{K}} \mathbb{I}_k\left(z_i\right) \ln \mathcal{N}_{\mathrm{M}}\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k^{(r-1)}, \boldsymbol{D}_k^{(r-1)} \left(\Delta_{\boldsymbol{w}_i} \boldsymbol{A}_k^{(r-1)}\right)^{-1} \boldsymbol{D}_k^{(r-1)^{\mathrm{t}}}\right) \\
& + \sum_{k=1}^{\mathrm{K}} \mathbb{I}_k\left(z_i\right) \sum_{m=1}^{\mathrm{M}} \ln \mathcal{G}\left(w_{i,k,m} \,;\, \alpha_{k,m}^{(r-1)}\right) \\
& + \ln H\left(z_i\,, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\tau}^{(r-1)}, \beta^{(r-1)}\right) \\
& \text{en notant : } \ln H\left(z_i\,, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\tau}, \beta\right) = \ln H\left(Z_1, \,\ldots,\, Z_{i-1}, z_i, Z_{i+1}, \,,\, \ldots,\, Z_{\mathrm{N}} \,;\, \boldsymbol{\tau}, \beta\right)
\end{aligned}
$$

$$
\begin{aligned}
& \ln p\left(\boldsymbol{w}_i, z_i \,|\, \boldsymbol{y}_i, \boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) \\
\propto\ & \frac{1}{2}\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\left[-\mathrm{M}\ln(2\pi) + \det\left(\Delta_{\boldsymbol{w}_i}\boldsymbol{A}_k^{(r-1)}\right)\right] \\
& -\frac{1}{2}\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\left[\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r-1)}\right)^{\mathrm{t}} \boldsymbol{D}_k^{(r-1)} \Delta_{\boldsymbol{w}_i} \boldsymbol{A}_k^{(r-1)} \boldsymbol{D}_k^{(r-1)^{\mathrm{t}}} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r-1)}\right)\right] \\
& +\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\left\{\sum_{m=1}^{\mathrm{M}}\left[-\ln\Gamma\left(\alpha_{k,m}^{(r-1)}\right) + \left(\alpha_{k,m}^{(r-1)} - 1\right)\ln(w_{i,m}) - w_{i,m}\right]\right\} \\
& +\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\left\{\tau_k^{(r-1)} + \frac{\beta^{(r-1)}}{2}\sum_{l\in\Omega(i)} \mathbb{I}_k(\boldsymbol{z}_l)\right\} \\
\propto\ & -\frac{1}{2}\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\sum_{m=1}^{\mathrm{M}} w_{i,m}\left[\boldsymbol{A}_k^{(r-1)}\boldsymbol{D}_k^{(r-1)^{\mathrm{t}}}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r-1)}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r-1)}\right)^{\mathrm{t}}\boldsymbol{D}_k^{(r-1)}\right]_{m,m} \\
& +\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\sum_{m=1}^{\mathrm{M}}\left[\frac{1}{2}\ln(w_{i,m}) + \left(\alpha_{k,m}^{(r-1)} - 1\right)\ln(w_{i,m}) - w_{i,m}\right] \\
& +\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\left\{\sum_{m=1}^{\mathrm{M}}\left[\frac{1}{2}\ln\left(\left[A_k^{(r-1)}\right]_{m,m}\right) - \ln\Gamma\left(\alpha_{k,m}^{(r-1)}\right)\right] + \tau_k^{(r-1)} + \frac{\beta^{(r-1)}}{2}\sum_{l\in\Omega(i)} \mathbb{I}_k(\boldsymbol{z}_l)\right\} \\
\propto\ & \sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\sum_{m=1}^{\mathrm{M}}\left\{\widetilde{\gamma}_{k,m}^{(r)}\ln\left(\widetilde{\delta}_{i,k,m}^{(r)}\right) - \ln\Gamma\left(\widetilde{\gamma}_{k,m}^{(r)}\right) + \left(\widetilde{\gamma}_{k,m}^{(r)} - 1\right)\ln(w_{i,m}) - \widetilde{\delta}_{i,k,m}^{(r)} w_{i,m}\right\} \\
& +\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\left\{\sum_{m=1}^{\mathrm{M}}\left[\frac{1}{2}\ln\left(\left[A_k^{(r-1)}\right]_{m,m}\right) - \ln\Gamma\left(\alpha_{k,m}^{(r-1)}\right)\right] + \tau_k^{(r-1)} + \frac{\beta^{(r-1)}}{2}\sum_{l\in\Omega(i)} \mathbb{I}_k(\boldsymbol{z}_l)\right\} \\
& -\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i)\sum_{m=1}^{\mathrm{M}}\left\{\widetilde{\gamma}_{k,m}^{(r)}\ln\left(\widetilde{\delta}_{i,k,m}^{(r)}\right) - \ln\Gamma\left(\widetilde{\gamma}_{k,m}^{(r)}\right)\right\} \\
& \text{avec } \widetilde{\gamma}_{k,m}^{(r)} = \alpha_{k,m}^{(r-1)} + \frac{1}{2} \\
& \text{et} \quad \widetilde{\delta}_{i,k,m}^{(r)} = 1 + \frac{1}{2}\left[\boldsymbol{A}_k^{(r-1)}\boldsymbol{D}_k^{(r-1)^{\mathrm{t}}}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r-1)}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r-1)}\right)^{\mathrm{t}}\boldsymbol{D}_k^{(r-1)}\right]_{m,m}
\end{aligned}
$$

d'où :

$$
\begin{aligned}
& \ln p\left(\boldsymbol{w}_i, z_i \,|\, \boldsymbol{y}_i, \boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) \\
\propto\ & \sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i) \sum_{m=1}^{\mathrm{M}} \left\{ \widetilde{\gamma}_{k,m}^{(r)} \ln\left(\widetilde{\delta}_{i,k,m}^{(r)}\right) - \ln\Gamma\left(\widetilde{\gamma}_{k,m}^{(r)}\right) + \left(\widetilde{\gamma}_{k,m}^{(r)} - 1\right) \ln(w_{i,m}) - \widetilde{\delta}_{i,k,m}^{(r)} w_{i,m} \right\} \\
& + \sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i) \left\{ \widetilde{\tau}_{i,k}^{(r)} + \frac{\beta^{(r-1)}}{2} \sum_{l \in \Omega(i)} \mathbb{I}_k(\boldsymbol{z}_l) \right\} \\
& \text{avec } \widetilde{\tau}_{i,k}^{(r)} = \tau_k^{(r-1)} + \sum_{m=1}^{\mathrm{M}} \left\{ \frac{1}{2} \ln\left(\left[A_k^{(r-1)} \right]_{m,m} \right) - \widetilde{\gamma}_{k,m}^{(r)} \ln\left(\widetilde{\delta}_{i,k,m}^{(r)}\right) + \ln\left(\frac{\Gamma\left(\widetilde{\gamma}_{k,m}^{(r)}\right)}{\Gamma\left(\alpha_{k,m}^{(r-1)}\right)} \right) \right\}
\end{aligned}
$$

En prenant l'espérance suivant $q^{(r-1)}_{\boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i}}$, nous obtenons :

$$
\begin{aligned}
& \mathrm{E}_{q^{(r-1)}_{\boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i}}} \left[\ln p\left(\boldsymbol{w}_i, z_i \,|\, \boldsymbol{y}_i, \boldsymbol{W}_{\setminus i}, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) \right] \\
\propto\ & \mathrm{E}_{q^{(r-1)}_{\boldsymbol{Z}_{\setminus i}}} \left[\mathrm{E}_{q^{(r-1)}_{\boldsymbol{W}_{\setminus i} \,|\, \boldsymbol{Z}_{\setminus i}}} \left[\ln p\left(\boldsymbol{y}_i \,|\, \boldsymbol{W}_i = \boldsymbol{w}_i, Z_i = z_i \,;\, \boldsymbol{\phi}^{(r-1)}\right) + \ln p\left(\boldsymbol{w}_i \,|\, Z_i = z_i \,;\, \boldsymbol{\phi}^{(r-1)}\right) \right] \right] \\
& + \mathrm{E}_{q^{(r-1)}_{\boldsymbol{Z}_{\setminus i}}} \left[\mathrm{E}_{q^{(r-1)}_{\boldsymbol{W}_{\setminus i} \,|\, \boldsymbol{Z}_{\setminus i}}} \left[\ln p\left(z_i \,|\, \boldsymbol{Z}_{\setminus i} \,;\, \boldsymbol{\phi}^{(r-1)}\right) \right] \right] \\
\propto\ & \sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i) \sum_{m=1}^{\mathrm{M}} \left\{ \widetilde{\gamma}_{k,m}^{(r)} \ln\left(\widetilde{\delta}_{i,k,m}^{(r)}\right) - \ln\Gamma\left(\widetilde{\gamma}_{k,m}^{(r)}\right) + \left(\widetilde{\gamma}_{k,m}^{(r)} - 1\right) \ln(w_{i,m}) - \widetilde{\delta}_{i,k,m}^{(r)} w_{i,m} \right\} \\
& + \mathrm{E}_{q^{(r-1)}_{\boldsymbol{Z}_{\setminus i}}} \left[\sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i) \left\{ \widetilde{\tau}_{i,k}^{(r)} + \frac{\beta^{(r-1)}}{2} \sum_{l \in \Omega(i)} \mathbb{I}_k(\boldsymbol{z}_l) \right\} \right] \\
\propto\ & \sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i) \sum_{m=1}^{\mathrm{M}} \left\{ \widetilde{\gamma}_{k,m}^{(r)} \ln\left(\widetilde{\delta}_{i,k,m}^{(r)}\right) - \ln\Gamma\left(\widetilde{\gamma}_{k,m}^{(r)}\right) + \left(\widetilde{\gamma}_{k,m}^{(r)} - 1\right) \ln(w_{i,m}) - \widetilde{\delta}_{i,k,m}^{(r)} w_{i,m} \right\} \\
& + \sum_{k=1}^{\mathrm{K}} \mathbb{I}_k(z_i) \left\{ \widetilde{\tau}_{i,k}^{(r)} + \frac{\beta^{(r-1)}}{2} \sum_{l \in \Omega(i)} q_{Z_l}^{(r-1)}(k) \right\}
\end{aligned}
$$

Ce qui nous permet de reconnaître la factorisation suivante :

$$
\begin{aligned}
q^{(r)}_{\boldsymbol{W}_i, Z_i}(\boldsymbol{w}_i, z_i) &= q^{(r)}_{\boldsymbol{W}_i \mid Z_i = z_i}(\boldsymbol{w}_i) \cdot q^{(r)}_{Z_i}(z_i) \\
&\text{avec } q^{(r)}_{Z_i}(k) = \mathcal{M}\left(1 \,;\, \widetilde{\pi}^{(r)}_{i,1}, \ldots, \widetilde{\pi}^{(r)}_{i,\mathrm{K}}\right),\ k = 1, \ldots, \mathrm{K} \\
&\text{et } q^{(r)}_{\boldsymbol{W}_i \mid Z_i = k}(\boldsymbol{w}_i) = \prod_{m=1}^{\mathrm{M}} \mathcal{G}\left(w_{i,m} \,;\, \widetilde{\gamma}^{(r)}_{k,m}, \widetilde{\delta}^{(r)}_{i,k,m}\right) \\
&\text{où } \widetilde{\pi}^{(r)}_{i,k} = \frac{\exp\left[\widetilde{\boldsymbol{\tau}}^{(r)}_{i,k} + \frac{\beta^{(r-1)}}{2} \sum_{l \in \Omega(i)} q^{(r-1)}_{Z_l}(k)\right]}{\sum_{j=1}^{\mathrm{K}} \exp\left[\widetilde{\boldsymbol{\tau}}^{(r)}_{i,j} + \frac{\beta^{(r-1)}}{2} \sum_{l \in \Omega(i)} q^{(r-1)}_{Z_l}(j)\right]}
\end{aligned}
$$

B.2 Étape variationnelle M-$(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})$

$$
\begin{aligned}
\left(\boldsymbol{\mu}^{(r)}, \boldsymbol{D}^{(r)}, \boldsymbol{A}^{(r)}\right) &= \underset{\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}}{\arg\max}\, \mathrm{E}_{q^{(r)}_{\boldsymbol{W}, \boldsymbol{Z}}}\left[\ln p\left(\boldsymbol{y} \mid \boldsymbol{W}, \boldsymbol{Z} \,;\, \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}\right)\right] \\
&= \underset{\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}}{\arg\max} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{\mathrm{N}} q^{(r)}_{Z_i}(k)\, \mathrm{E}_{q^{(r)}_{\boldsymbol{W}_i \mid Z_i = k}}\left[\ln p\left(\boldsymbol{y}_i \mid \boldsymbol{W}_i, Z_i = k \,;\, \boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k\right)\right]
\end{aligned}
$$

d'où :

$$
\left(\boldsymbol{\mu}^{(r)}_k, \boldsymbol{D}^{(r)}_k, \boldsymbol{A}^{(r)}_k\right) = \underset{\boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k}{\arg\max} \underbrace{\sum_{i=1}^{\mathrm{N}} q^{(r)}_{Z_i}(k)\, \mathrm{E}_{q^{(r)}_{\boldsymbol{W}_i \mid Z_i = k}}\left[\ln p\left(\boldsymbol{y}_i \mid \boldsymbol{W}_i, Z_i = k \,;\, \boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k\right)\right]}_{f_1(\boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k)}
$$

$$
\begin{aligned}
f_1\left(\boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k\right) &= \frac{1}{2} \sum_{i=1}^{\mathrm{N}} q^{(r)}_{Z_i}(k)\, \mathrm{E}_{q^{(r)}_{\boldsymbol{W}_i \mid Z = k}}\left[-\mathrm{M} \ln(2\pi) + \ln\det\left(\Delta_{\boldsymbol{W}_i}\right) + \ln\det\left(\boldsymbol{A}_k\right)\right] \\
&\quad - \frac{1}{2} \sum_{i=1}^{\mathrm{N}} q^{(r)}_{Z_i}(k)\, \mathrm{E}_{q^{(r)}_{\boldsymbol{W}_i \mid Z = k}}\left[\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k\right)^{\mathrm{t}} \boldsymbol{D}_k \Delta_{\boldsymbol{W}_i} \boldsymbol{A}_k \boldsymbol{D}_k^{\mathrm{t}} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k\right)\right] \\
&= \frac{1}{2} \sum_{i=1}^{\mathrm{N}} q^{(r)}_{Z_i}(k) \left[\ln\det\left(\boldsymbol{A}_k\right) - \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k\right)^{\mathrm{t}} \boldsymbol{D}_k \widetilde{\Delta}_{i,k} \boldsymbol{A}_k \boldsymbol{D}_k^{\mathrm{t}} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k\right)\right]
\end{aligned}
$$

à une constante additive près, indépendante de $(\boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k)$ et en posant :

$$\widetilde{\Delta}_{i,k}^{(r)} = \mathrm{E}_{q_{\boldsymbol{W}_i \mid Z_i=k}^{(r)}} \left(\Delta_{\boldsymbol{W}_i}\right) = \mathrm{diag}\left(\frac{\widetilde{\gamma}_{k,1}^{(r)}}{\widetilde{\delta}_{i,k,1}^{(r)}}, \ldots, \frac{\widetilde{\gamma}_{k,\mathrm{M}}^{(r)}}{\widetilde{\delta}_{i,k,\mathrm{M}}^{(r)}}\right)$$

Nous pouvons alors utiliser les estimations par maximum de vraisemblance présentés dans [7], ce qui donne les expressions suivantes pour chaque dimension $m \in \mathrm{M}$:

$$\mu_{k,m}^{(r)} = \frac{\sum_{i=1}^{\mathrm{N}} \widetilde{\pi}_{i,k}^{(r)} \left[\boldsymbol{D}_k^{(r-1)} \widetilde{\Delta}_{i,k}^{(r)} \boldsymbol{D}_k^{(r-1)^{\mathrm{t}}} \boldsymbol{y}_i\right]_m}{\sum_{j=1}^{\mathrm{N}} \widetilde{\pi}_{j,k}^{(r)} \left[\widetilde{\Delta}_{j,k}^{(r)}\right]_{m,m}}$$

$$\boldsymbol{D}_k^{(r)} = \arg\min_{\boldsymbol{D}_k} \sum_{i=1}^{\mathrm{N}} \widetilde{\pi}_{i,k}^{(r)} \mathrm{tr}\left[\boldsymbol{D}_k \widetilde{\boldsymbol{\Delta}}_{i,k}^{(r)} \boldsymbol{A}_k^{(r-1)} \boldsymbol{D}_k^{\mathrm{t}} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r)}\right) \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r)}\right)^{\mathrm{t}}\right]$$

la forme ainsi obtenue peut être résolue par l'algorithme ALS ([30]).

$$A_{k,m}^{(r)} = \frac{\sum_{j=1}^{\mathrm{N}} \widetilde{\pi}_{j,k}^{(r)}}{\sum_{i=1}^{\mathrm{N}} \widetilde{\pi}_{i,k}^{(r)} \left[\widetilde{\Delta}_{i,k}^{(r)}\right]_{m,m} \left[\boldsymbol{D}_k^{(r)^{\mathrm{t}}} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(r)}\right)\right]_m^2}$$

B.3 Étape variationnelle M-$\boldsymbol{\alpha}$

$$\begin{aligned}
\boldsymbol{\alpha}^{(r)} &= \arg\max_{\boldsymbol{\alpha}} \mathrm{E}_{q_{\boldsymbol{W},\boldsymbol{Z}}^{(r)}} \left[\ln p\left(\boldsymbol{W} \mid \boldsymbol{Z} \,;\, \boldsymbol{\alpha}\right)\right] \\
&= \arg\max_{\boldsymbol{\alpha}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{\mathrm{N}} q_{Z_i}(k) \, \mathrm{E}_{q_{\boldsymbol{W}_i \mid Z_i=k}^{(r)}} \left[\ln p\left(\boldsymbol{W}_i \mid Z_i = k \,;\, \boldsymbol{\alpha}_k\right)\right] \\
&= \arg\max_{\boldsymbol{\alpha}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{\mathrm{N}} \sum_{m=1}^{\mathrm{M}} q_{Z_i}(k) \, \mathrm{E}_{q_{\boldsymbol{W}_i \mid Z_i=k}^{(r)}} \left[\ln \mathcal{G}\left(W_{i,m} \,;\, \alpha_{k,m}\right)\right]
\end{aligned}$$

d'où :

$$\alpha_{k,m}^{(r)} = \arg\max_{\alpha_{k,m}} \sum_{i=1}^{\mathrm{N}} q_{Z_i}(k) \, \mathrm{E}_{q_{\boldsymbol{W}_i \mid Z_i=k}^{(r)}} \left[\ln \mathcal{G}\left(W_{i,m} \,;\, \alpha_{k,m}\right)\right]$$

Cette maximisation revient à déterminer $\alpha_{k,m}^{(r)}$ comme solution de l'équation suivante ([7]) :

$$\Upsilon\left(\alpha_{k,m}\right) = \Upsilon\left(\widetilde{\gamma}_{k,m}^{(r)}\right) - \frac{\sum_{i=1}^{\mathrm{N}} \widetilde{\pi}_{i,k}^{(r)} \ln\left(\widetilde{\delta}_{i,k,m}^{(r)}\right)}{\sum_{j=1}^{\mathrm{N}} \widetilde{\pi}_{j,k}^{(r)}} \tag{B.1}$$

B.4 Calcul de l'énergie libre $\mathcal{L}\left(Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}\,;\,\boldsymbol{y},\boldsymbol{\phi}^{(r)}\right)$

$$
\begin{aligned}
&\mathcal{L}\left(Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}\,;\,\boldsymbol{y},\boldsymbol{\phi}^{(r)}\right)\\
&\quad=\mathrm{E}_{Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}}\left\{\log p\left(\boldsymbol{y},\boldsymbol{W},\boldsymbol{Z}\,;\,\phi^{(r)}\right)-\log Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}\left(\boldsymbol{W},\boldsymbol{Z}\right)\right\}\\
&\quad=\mathrm{E}_{Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}}\left\{\ln p\left(\boldsymbol{y}\,|\,\boldsymbol{W},\boldsymbol{Z}\,;\,\boldsymbol{\mu}^{(r)},\boldsymbol{D}^{(r)},\boldsymbol{A}^{(r)}\right)+\ln p\left(\boldsymbol{W}\,|\,\boldsymbol{Z}\,;\,\boldsymbol{\alpha}^{(r)}\right)+\ln p\left(\boldsymbol{Z}\,|\,\boldsymbol{\tau}^{(r)},\beta^{(r)}\right)\right\}\\
&\qquad-\mathrm{E}_{Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}}\left\{\log Q^{(r)}_{\boldsymbol{W}\,|\,\boldsymbol{Z}}\left(\boldsymbol{W}\right)+\log Q^{(r)}_{\boldsymbol{Z}}\left(\boldsymbol{Z}\right)\right\}\\
&\quad=\mathrm{E}_{Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}}\left[\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{I}_k\left(Z_i\right)\left\{\frac{-\mathrm{M}}{2}\log\left(2\pi\right)+\frac{1}{2}\log\left|\Delta_{\boldsymbol{W}_i}\boldsymbol{A}^{(r)}_k\right|\right\}\right]\\
&\qquad+\mathrm{E}_{Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}}\left[\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{I}_k\left(Z_i\right)\frac{-1}{2}\left(\boldsymbol{y}_i-\boldsymbol{\mu}^{(r)}_k\right)^{\mathrm{t}}\boldsymbol{D}^{(r)}_k\Delta_{\boldsymbol{W}_i}\boldsymbol{A}^{(r)}_k\left(\boldsymbol{D}^{(r)}_k\right)^{\mathrm{t}}\left(\boldsymbol{y}_i-\boldsymbol{\mu}^{(r)}_k\right)\right]\\
&\qquad+\mathrm{E}_{Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}}\left[\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{I}_k\left(Z_i\right)\sum_{m=1}^{\mathrm{M}}\left\{-\ln\Gamma\left(\alpha^{(r)}_{k,m}\right)+\left(\alpha^{(r)}_{k,m}-1\right)\ln\left(W_{i,m}\right)-W_{i,m}\right\}\right]\\
&\qquad+\mathrm{E}_{Q^{(r)}_{\boldsymbol{Z}}}\left[-\ln\left(\mathcal{K}\left(\boldsymbol{\tau}^{(r)},\beta^{(r)}\right)\right)+\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{I}_k\left(Z_i\right)\left\{\tau^{(r)}_k+\frac{\beta^{(r)}}{2}\sum_{l\in\Omega(i)}\mathbb{I}_k\left(Z_m\right)\right\}\right]\\
&\qquad-\mathrm{E}_{Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}}\left[\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{I}_k\left(Z_i\right)\sum_{m=1}^{\mathrm{M}}\left\{\widetilde{\gamma}^{(r)}_{k,m}\ln\left(\widetilde{\delta}^{(r)}_{i,k,m}\right)-\ln\Gamma\left(\widetilde{\gamma}^{(r)}_{k,m}\right)\right\}\right]\\
&\qquad-\mathrm{E}_{Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}}\left[\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{I}_k\left(Z_i\right)\sum_{m=1}^{\mathrm{M}}\left\{\left(\widetilde{\gamma}^{(r)}_{k,m}-1\right)\ln\left(W_{i,m}\right)-\widetilde{\delta}^{(r)}_{k,m}W_{i,m}\right\}\right]\\
&\qquad-\mathrm{E}_{Q^{(r)}_{\boldsymbol{Z}}}\left[\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{I}_k\left(Z_i\right)\ln\left(\widetilde{\pi}^{(r)}_{i,k}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{L}\left(Q^{(r)}_{\boldsymbol{W},\boldsymbol{Z}}\,;\,\boldsymbol{y},\boldsymbol{\phi}^{(r)}\right)\\
&=\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{(r)}_{i,k}\left\{\frac{-\mathrm{M}}{2}\log\left(2\pi\right)+\frac{1}{2}\sum_{m=1}^{\mathrm{M}}\left\{\Upsilon\left(\widetilde{\gamma}^{(r)}_{k,m}\right)-\ln\left(\widetilde{\delta}^{(r)}_{i,k,m}\right)\right\}+\frac{1}{2}\sum_{m=1}^{\mathrm{M}}\ln\left(\left[A^{(r)}_{k}\right]_{l,m}\right)\right\}\\
&\quad+\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{(r)}_{i,k}\frac{-1}{2}\left(\boldsymbol{y}_i-\boldsymbol{\mu}^{(r)}_{k}\right)^{\mathrm{t}}\boldsymbol{D}^{(r)}_{k}\mathrm{diag}\left(\frac{\widetilde{\gamma}^{(r)}_{k1}}{\widetilde{\delta}^{(r)}_{i,k1}},\,\ldots,\,\frac{\widetilde{\gamma}^{(r)}_{k\mathrm{M}}}{\widetilde{\delta}^{(r)}_{i,k\mathrm{M}}}\right)\boldsymbol{A}^{(r)}_{k}\left(\boldsymbol{D}^{(r)}_{k}\right)^{\mathrm{t}}\left(\boldsymbol{y}_i-\boldsymbol{\mu}^{(r)}_{k}\right)\\
&\quad+\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{(r)}_{i,k}\sum_{m=1}^{\mathrm{M}}\left\{-\ln\Gamma\left(\alpha^{(r)}_{k,m}\right)+\left(\alpha^{(r)}_{k,m}-1\right)\left[\Upsilon\left(\widetilde{\gamma}^{(r)}_{k,m}\right)-\ln\left(\widetilde{\delta}^{(r)}_{i,k,m}\right)\right]-\frac{\widetilde{\gamma}^{(r)}_{k,m}}{\widetilde{\delta}^{(r)}_{i,k,m}}\right\}\\
&\quad-\ln\left(\mathcal{K}\left(\boldsymbol{\tau}^{(r)},\beta^{(r)}\right)\right)+\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{(r)}_{i,k}\left\{\tau^{(r)}_{k}+\frac{\beta^{(r)}}{2}\sum_{l\in\Omega(i)}\widetilde{\pi}^{(r)}_{l,k}\right\}\\
&\quad-\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{(r)}_{i,k}\sum_{m=1}^{\mathrm{M}}\left\{\widetilde{\gamma}^{(r)}_{k,m}\ln\left(\widetilde{\delta}^{(r)}_{i,k,m}\right)-\ln\Gamma\left(\widetilde{\gamma}^{(r)}_{k,m}\right)+\left(\widetilde{\gamma}^{(r)}_{k,m}-1\right)\left[\Upsilon\left(\widetilde{\gamma}^{(r)}_{k,m}\right)-\ln\left(\widetilde{\delta}^{(r)}_{i,k,m}\right)\right]-\widetilde{\gamma}^{(r)}_{k,m}\right.\\
&\quad-\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{(r)}_{i,k}\ln\left(\widetilde{\pi}^{(r)}_{i,k}\right)
\end{aligned}
$$

L'expression précédente n'est pas calculable directement du fait de la constante de normalisation $\mathcal{K}$ de la distribution de Gibbs. Il est donc nécessaire de l'approcher par $\widetilde{\mathcal{K}}$, par exemple via la borne de Gibbs-Bogoliubov-Feynman tel qu'indiqué dans [40].

En réécrivant l'expression de $q^{prior}_{\boldsymbol{Z}}$, formule (5.21), sous la même forme qu'une distribution de Gibbs :

$$
\begin{aligned}
q^{prior}_{\boldsymbol{Z}}\left(\boldsymbol{z}\,;\,\boldsymbol{\tau},\beta\right) &= \left(\mathcal{K}^{prior}\left(\boldsymbol{\tau},\beta\right)\right)^{-1}\exp\left(H^{prior}\left(\boldsymbol{z}\,;\,\boldsymbol{\tau},\beta\right)\right) \qquad \text{(B.2)}\\
\text{avec}\qquad \mathcal{K}^{prior}\left(\boldsymbol{\tau},\beta\right) &= \prod_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\exp\left(\tau_k+\frac{\beta}{2}\sum_{l\in\Omega(i)}\widetilde{\pi}_{l,k}\right)\\
\text{et}\qquad H^{prior}\left(\boldsymbol{z}\,;\,\boldsymbol{\tau},\beta\right) &= \sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{I}_k\left(z_i\right)\left(\tau_k+\frac{\beta}{2}\sum_{l\in\Omega(i)}\widetilde{\pi}_{l,k}\right)
\end{aligned}
$$

l'approximation $\widetilde{\mathcal{K}}$ de $\mathcal{K}$ s'exprime comme suit :

$$
\widetilde{\mathcal{K}}\left(\boldsymbol{\tau},\beta\right) = \mathcal{K}^{prior}\left(\boldsymbol{\tau},\beta\right)\exp\left(\mathbb{E}_{q^{prior}_{\boldsymbol{Z}}}\left[H\left(\boldsymbol{Z}\,;\,\boldsymbol{\tau},\beta\right)+H^{prior}\left(\boldsymbol{Z}\,;\,\boldsymbol{\tau},\beta\right)\right]\right) \qquad \text{(B.3)}
$$

or

$$\begin{aligned}\mathbb{E}_{q^{prior}_{\boldsymbol{Z}}}\left[H\left(\boldsymbol{Z}\,;\,\boldsymbol{\tau},\beta\right)\right] &= \sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{E}_{q^{prior}_{\boldsymbol{Z}}}\left[\mathbb{I}_k\left(z_i\right)\left(\tau_k+\frac{\beta}{2}\sum_{l\in\Omega(i)}\mathbb{I}_k\left(z_l\right)\right)\right]\\ &= \sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{prior}_{i,k}\left(\tau_k+\frac{\beta}{2}\sum_{l\in\Omega(i)}\widetilde{\pi}^{prior}_{l,k}\right)\end{aligned}$$

et

$$\begin{aligned}\mathbb{E}_{q^{prior}_{\boldsymbol{Z}}}\left[H^{prior}\left(\boldsymbol{Z}\,;\,\boldsymbol{\tau},\beta\right)\right] &= \sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\mathbb{E}_{q^{prior}_{\boldsymbol{Z}}}\left[\mathbb{I}_k\left(z_i\right)\left(\tau_k+\frac{\beta}{2}\sum_{l\in\Omega(i)}\widetilde{\pi}_{l,k}\right)\right]\\ &= \sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{prior}_{i,k}\left(\tau_k+\frac{\beta}{2}\sum_{l\in\Omega(i)}\widetilde{\pi}_{l,k}\right)\end{aligned}$$

d'où l'expression finale :

$$\begin{aligned}\widetilde{\mathcal{K}}\left(\boldsymbol{\tau},\beta\right) &= \prod_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\exp\left(\tau_k+\frac{\beta}{2}\sum_{l\in\Omega(i)}\widetilde{\pi}_{l,k}\right)\\ &\quad\times\exp\left[\sum_{i=1}^{\mathrm{N}}\sum_{k=1}^{\mathrm{K}}\widetilde{\pi}^{prior}_{i,k}\left(2\tau_k+\frac{\beta}{2}\sum_{l\in\Omega(i)}\left(\widetilde{\pi}^{prior}_{l,k}+\widetilde{\pi}_{l,k}\right)\right)\right] \quad \text{(B.4)}\end{aligned}$$