HAL
open science

Mélanges de lois de Student à Échelles Multiples pour la caractérisation de tumeurs par IRM multiparamétrique

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

To cite this version:

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 SFdS, Jun 2015, Lille, France. hal-01253588

HAL Id: hal-01253588

https://hal.science/hal-01253588v1

Submitted on 11 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mélanges de lois de Student à Échelles Multiples pour la caractérisation de tumeurs par IRM multiparamétrique.

Alexis Arnaud [1] & Florence Forbes [1] & Benjamin Lemasson [2] & Emmanuel Barbier [2] & Nicolas Coquery [3]

[1] *Inria Grenoble - Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot-Saint-Martin, France, prenom.nom@inria.fr*
[2] *GIN, 31 Chemin Fortuné Ferrini, 38706 La Tronche, France, prenom.nom@ujf-grenoble.fr*
[3] *INRA Rennes, Domaine de la Prise, 35590 Saint Gilles, France, prenom.nom@rennes.inra.fr*

Résumé. Dans cette étude nous développons une méthode statistique pour la classification de données IRM multiparamétriques permettant de réaliser un contrôle qualité des données (détection de données atypiques), et de construire un dictionnaire de signatures de tumeurs. Une précédente étude se base sur le modèle de mélange de lois gaussiennes dans lequel on cherche à rassembler des observations en groupes (classes) issus de lois gaussiennes. Ce modèle est connu pour sa sensibilité aux valeurs atypiques qui peuvent sensiblement dégrader la pertinence des groupes obtenus. Nous proposons donc d'utiliser des lois de Student généralisées qui permettent d'attribuer un poids à chaque observation dans chaque dimension, de façon à pondérer l'influence des valeurs atypiques sur la forme des groupes. Il en résulte une plus grande flexibilité dans l'ajustement des classes. Nous utilisons ainsi un algorithme d'Expectation-Maximization (EM), ainsi qu'un critère bayésien de sélection de modèle, que nous appliquons sur un échantillon de 37 rats présentant quatre modèles de tumeur. Nous arrivons à détecter des animaux atypiques avant de construire un dictionnaire de tumeurs discriminant fortement les quatre gliomes considérés.

Mots-clés. Distribution de Student multivariée à échelles multiples, modèle de mélange, classification non supervisée, algorithme EM, IRM multiparamétrique, dictionnaire de tumeurs.

Abstract. In this study we develop a statistical method for the classification of multiparametric MRI, which allows data quality control (atypical observation detection), and provides a tumour signatures dictionary. A previous study has been based on a Gaussian mixture model in which observations are grouped into classes corresponding to Gaussian distributions. This model is known for its sensitivity to outliers which can degrade the relevance of the obtained groups. As an alternative, we propose to use generalized Student distributions which extend the standard multivariate Student distribution by allowing different weights on the different dimensions of each observation.

This allows a finer regulation of the influence of atypical data on the groups shapes. It follows a greater flexibility of the clustering. We use an Expectation-Maximization algorithm (EM), and a Bayesian criterion for the model selection. Our sample consists of 37 rats with 4 tumour models. We manage to detect atypical animals previous to the construction of a tumour dictionary which highly discriminates the 4 tumours.

Keywords. Multivariate multiple scaled Student distribution, mixture model, EM algorithm, clustering, multiparametric MRI, tumor dictionary.

1 Classification de données IRM multiparamétriques

Nous nous intéressons dans cette étude à des données obtenues par IRM multiparamétrique, c'est-à-dire des images IRM où M paramètres physiologiques ont été acquis au cours de la même expérience. Nous disposons donc d'un ensemble de N points en dimension M = 5, chaque point correspond à un voxel dans une image de cerveau de rat. Nous noterons ces observations :

$$\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{\mathrm{N}}\} \quad \text{avec} \quad \boldsymbol{y}_i = (y_{i1}, \ldots, y_{i\mathrm{M}}) \in \mathbb{R}^{\mathrm{M}} \,. \tag{1}$$

En mesurant ces 5 paramètres d'intérêt, nous espérons pouvoir caractériser les différents types de tissus présents dans le cerveau du rat, et à terme le cerveau humain.

Pour cela, nous cherchons à regrouper les voxels dont les valeurs physiologiques se ressemblent. Un modèle très utilisé en classification non supervisée est de considérer un modèle de mélange gaussien. Cependant, un tel modèle est sensible aux données atypiques. Pour remédier à cela, nous utiliserons dans ce travail des lois de Student à échelles multiples. Comme nous pouvons le voir sur la figure 1, de telles lois permettent de retrouver des classes alongées car plus tolérantes aux valeurs atypiques.

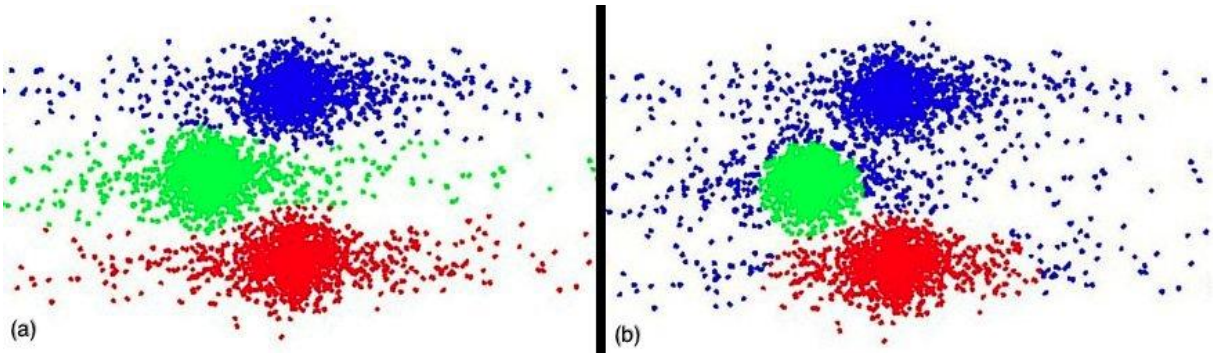

FIGURE 1 – Classification bivariée en 3 groupes avec lois (a) de Student généralisées (Forbes et Wraith 2014) et (b) gaussiennes.

2 Les mélanges de lois de Student à échelles multiples

Une loi de Student à échelles multiples est une généralisation de la loi de Student multivariée standard présentée par Forbes et Wraith 2014. La loi de Student multivariée standard, de paramètre de position $\boldsymbol{\mu} \in \mathbb{R}^{\mathrm{M}}$, d'échelle $\boldsymbol{T} \in \mathcal{M}_{\mathrm{M}\times\mathrm{M}}(\mathbb{R})$ et de degré de liberté $\nu \in \mathbb{R}_*^+$, admet la densité suivante pour la variable aléatoire $\boldsymbol{Y} \in \mathbb{R}^{\mathrm{M}}$:

$$p_{\mathrm{S}}(\boldsymbol{y}\ ;\ \boldsymbol{\mu}, \boldsymbol{T}, \nu) = \int_0^\infty \mathcal{N}_{\mathrm{M}}(\boldsymbol{y}\ ;\ \boldsymbol{\mu}, w\,\boldsymbol{T})\,\mathcal{G}\left(w\ ;\ \frac{\nu}{2}, \frac{\nu}{2}\right) \mathrm{d}w\ , \tag{2}$$

où $\mathcal{N}_{\mathrm{M}}$ est la densité de la loi gaussienne en dimension M de moyenne $\boldsymbol{\mu}$ et de matrice de précision $w\boldsymbol{T}$, et où $\mathcal{G}$ est la densité de la loi gamma de paramètre $\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$. Cette expression de la densité de la loi de Student multivariée peut-être vue comme un modèle de mélange infini de lois gaussiennes, avec comme variable latente W qui est régie par une loi Gamma. Cette dernière est appelée variable de poids dans la suite.

La généralisation de Forbes et Wraith 2014 consiste à décomposer la matrice de précision en éléments propres $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathrm{t}}$, avec $\boldsymbol{U}$ la matrice orthogonale des vecteurs propres normés, et $\boldsymbol{D}$ la matrice diagonale des valeurs propres. Ceci permet de remplacer la variable latente réelle de poids $W \in \mathbb{R}^+$ par une variable latente M-dimensionnelle $\boldsymbol{W} \in \mathbb{R}_+^{\mathrm{M}}$. La loi de Student à échelles multiples admet ainsi la densité suivante :

$$p_{\mathrm{SEM}}(\boldsymbol{y}\ ;\ \boldsymbol{\mu}, \boldsymbol{U}, \boldsymbol{D}, \boldsymbol{\nu}) = \int_0^\infty \ldots \int_0^\infty \mathcal{N}_{\mathrm{M}}\left(\boldsymbol{y}\ ;\ \boldsymbol{\mu}, \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{D}\boldsymbol{U}^{\mathrm{t}}\right) \prod_{m=1}^{\mathrm{M}} \mathcal{G}\left(w_m\ ;\ \frac{\nu_m}{2}, \frac{\nu_m}{2}\right)\ \mathrm{d}w_1 \ldots \mathrm{d}w_{\mathrm{M}} \tag{3}$$

où $\boldsymbol{\Delta} = \mathrm{diag}(\boldsymbol{W}) = \mathrm{diag}(w_1,\ \ldots,\ w_{\mathrm{M}})$ est la matrice diagonale de poids.

Nous nous intéressons ici à un mélange de lois de Student à échelles multiples, puisque nous souhaitons rassembler les voxels d'une image en classes, chaque classe étant régie par une loi différente. La densité d'un mélange de lois statistiques continues s'explicite simplement comme la somme pondérée des densités de classe. En considérant un mélange à K classes, admettant pour proportion respective $\boldsymbol{\pi} = \{\pi_1,\ \ldots,\ \pi_{\mathrm{K}}\}$, nous obtenons la densité suivante pour la varialbe aléatoire $\boldsymbol{Y} \in \mathbb{R}^{\mathrm{M}}$:

$$p(\boldsymbol{y}\ ;\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{U}, \boldsymbol{D}, \boldsymbol{\nu}) = \sum_{k=1}^{\mathrm{K}} \pi_k\ p_{\mathrm{SEM}}(\boldsymbol{y}\ ;\ \boldsymbol{\mu}_k, \boldsymbol{U}_k, \boldsymbol{D}_k, \boldsymbol{\nu}_k) \tag{4}$$

avec comme paramètres $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1,\ \ldots,\ \boldsymbol{\mu}_{\mathrm{K}}\}$, $\boldsymbol{U} = \{\boldsymbol{U}_1,\ \ldots,\ \boldsymbol{U}_{\mathrm{K}}\}$, $\boldsymbol{D} = \{\boldsymbol{D}_1,\ \ldots,\ \boldsymbol{D}_{\mathrm{K}}\}$, et $\boldsymbol{\nu} = \{\boldsymbol{\nu}_1,\ \ldots,\ \boldsymbol{\nu}_{\mathrm{K}}\}$. Nous notons ainsi $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{U}, \boldsymbol{D}, \boldsymbol{\nu}\}$.

L'estimation des paramètres de ce modèle est réalisée via un algorithme d'Estimation-Maximisation (EM), en introduisant une deuxième variable latent Z, en plus du poids

$\boldsymbol{W}$. Ceci permet d'associer une classe à chaque observation $\boldsymbol{y}_n$ de notre échantillon : $\Pr(z_n = k) = \pi_k$ pour $k \in \{1, \ldots, \mathrm{K}\}, n = \{1, \ldots, \mathrm{N}\}$. Pour chaque observation $\boldsymbol{y}_n$, nous obtenons donc la densité conditionnelle suivante :

$$p(\boldsymbol{y}_n \,|\, Z_n = k \,;\, \boldsymbol{\theta}) = \int_0^\infty \int_0^\infty \mathcal{N}_{\mathrm{M}}\left(\boldsymbol{y}_n \,;\, \boldsymbol{\mu}_k, \boldsymbol{U}_k \boldsymbol{\Delta}_n \boldsymbol{D}_k \boldsymbol{U}_k^{\mathrm{t}}\right) \prod_{m=1}^{\mathrm{M}} \mathcal{G}\left(w_{nm} \,;\, \frac{\nu_{km}}{2}, \frac{\nu_{km}}{2}\right) dw_{n1} \ldots dw_{nM} \,, \tag{5}$$

avec $\boldsymbol{\Delta}_n = \mathrm{diag}\,(w_{n1}, \ldots, w_{n\mathrm{M}})$. La différence par rapport au cas du mélange de lois gaussiennes réside dans la présence de poids multidimensionnels. L'étape E de calcul des densités conditionnelles passe par le calcul des quantités suivantes lors de l'itération r :

$$\boldsymbol{w}_{nk}^{(r)} = \left(w_{nk1}^{(r)}, \ldots, w_{nk\mathrm{M}}^{(r)}\right) \tag{6}$$

$$\text{avec } w_{nkm}^{(r)} = \mathbb{E}\left[W_{nm} \,|\, Z_n = k, \boldsymbol{y}_n \,;\, \boldsymbol{\theta}^{(r)}\right] = \frac{\nu_{km}^{(r)} + 1}{\nu_{km}^{(r)} + \left[\boldsymbol{D}_k^{(r)}\right]_m \left[\boldsymbol{U}_k^{(r)^{\mathrm{t}}} \left(\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(r)}\right)\right]_m^2} \tag{7}$$

Le terme $\left[\boldsymbol{D}_k^{(r)}\right]_m \left[\boldsymbol{U}_k^{(r)^{\mathrm{t}}} \left(\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(r)}\right)\right]_m^2$ dans l'équation (6) peut s'interpréter comme le carré de la distance de Mahalanobis entre $\left[\boldsymbol{U}_k^{(r)^{\mathrm{t}}} \boldsymbol{y}_n\right]_m$ et $\left[\boldsymbol{U}_k^{(r)^{\mathrm{t}}} \boldsymbol{\mu}_k^{(r)}\right]_m$ pour la variance $\left[\boldsymbol{D}_k^{(r)}\right]_m^{-1}$. Ce terme est donc grand pour les valeurs atypiques, celles-ci s'écartant du centre de la classe. Ainsi plus une observation est loin du centre d'une classe, et moins elle contribue à l'ajustement de la classe. Pour plus de détails sur le reste de l'algorithme EM voir Forbes et Wraith 2014.

3 Application à la reconnaissance de tumeurs

Nous appliquons ce modèle à des IRM de cerveaux de rats présentant quatre modèles de gliomes : 9L ($n = 5$), C6 de deux laboratoires (C6a, $n = 6$; C6b, $n = 13$), et F98 ($n = 13$). Les images IRM ont été acquises suivant un protocole similaire à Coquery et al. 2014. Les paramètres d'intérêt pour chaque rat sont : le coefficient de diffusion apparent (ADC), le volume sanguin cérébral (CBV), le débit sanguin cérébral (CBF), la perméabilité des vaisseaux (AUC), et le taux d'oxygénation des tissus (StO2). Des régions d'intérêt (ROI : 1 tumeur à gauche, 2 saines à droite) ont été manuellement délimitées sur les images pondérée T2, puis reportées sur les cartes physiologiques. Les voxels de toutes les ROI sont rassemblés et classifiés ensemble de façon à regrouper les voxels présentant des caractéristiques micro-vasculaires similaires.

La classification des voxels nécessite de fixer le nombre K de classes à estimer. Pour cela nous minimisons le critère d'information bayésien (BIC), ce dernier présentant un

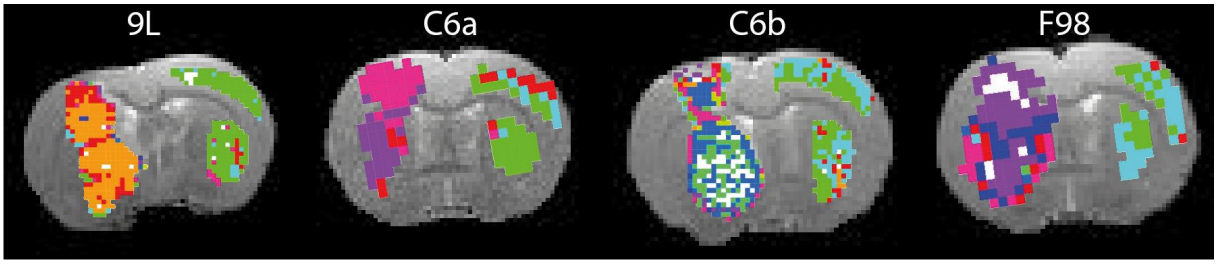

FIGURE 2 – Classification de cerveaux de rats (10 classes) pour 4 modèles de gliome (9L, C6a, C6b, F98), illustrée sur une tranche d'IRM.

minimum marqué contrairement au cas du mélange gaussien de Coquery et al. 2014. Nous obtenons K = 10, ce qui donne une classification (figure 2) cohérente avec celle de Coquery et al. 2014, mais avec des groupes dont les centres sont moins déviés par les valeurs extrêmes, et dont les variances sont réduites. L'estimation des poids (eq.(6)) de chaque observation permet d'identifier des classes où toutes les observations sont proches du centre, ainsi que des classes avec un nuage d'observations plus étiré. La classification ainsi réalisée permet également de construire une signature pour chaque ROI en considérant la proportion de chaque classe au sein d'une ROI (figure 3). Les ROI saines présentent une signature similaire au sein de chaque type de rats (« Striat », figure 3). Cette similarité des ROI saines liée au type permet de détecter des rats atypiques, ce qui est le cas pour l'un des rats (« Outlier striat », figure 3). Au contraire, les ROI de tumeurs présentent une grande différence entre elles, y compris entre les deux souches C6 (« C6a » et « C6b », figure 3). Le pouvoir discriminant des signatures établies est évalué par une procédure leave-one-out : la reconnaissance est quasi parfaite ($> 97\%$) avec une seule tumeur F98 mal-classée en tumeur « C6b ».

4 Conclusion

L'utilisation d'un mélange de lois de Student à échelles multiples permet de prendre en compte des informations concernant l'éloignement d'une observation au centre d'une classe et ceci pour chaque paramètre. Nous avons utilisé la classification ainsi obtenue afin de construire des signatures pour chaque ROI, ce qui a notamment permis de détecter et d'écarter un rat atypique. De plus, ces signatures ont fortement discriminé les modèles de tumeurs, y compris les deux gliomes C6 utilisés, ce qui nous laisse penser que ce modèle statistique pourrait détecter des dérives de lignées cellulaires entre laboratoires.

Un travail parallèle est en cours pour l'ajout d'une composante spatiale au modèle, via un champ de Markov caché ; ceci serait utilisé pour faire de la segmentation de tumeurs ou du contrôle qualité sur la délimitation manuelle des ROI. De même nous cherchons à intégrer au modèle le choix du nombre de classes à considérer dans le mélange, de façon à

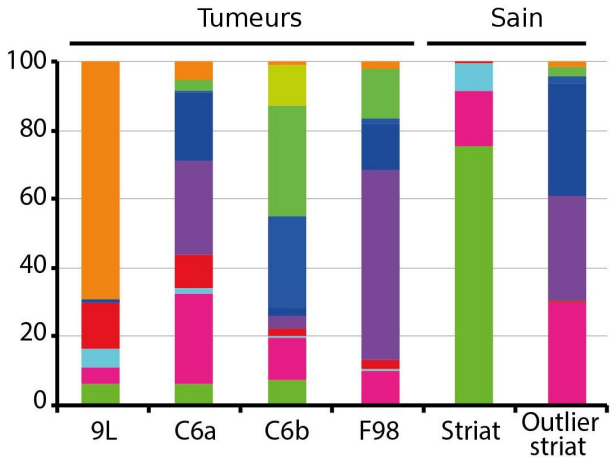

FIGURE 3 – Signature des ROI représentée comme la moyenne des proportions respectives de chaque classe dans chaque ROI. La dernière colonne montre la signature de la partie saine d'un rat atypique.

remplacer le critère du BIC. Pour cela, nous travaillons sur une approche bayésienne avec une approximation variationnelle de la vraisemblance.

Bibliographie

[1] Coquery N, Francois O, Lemasson B, Debacker C, Farion R, Rémy C, et Barbier E (2014), *Microvascular MRI and unsupervised clustering yields histology-resembling images in two rat models of glioma*, Journal of Cerebral Blood Flow & Metabolism, volume 34, numéro 8, 1354–62.
[2] Forbes F, et Wraith D (2014), *A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights : Application to robust clustering*, Statistics and Computing, volume 24, numéro 6, 971–984.