Paquet R pour l'estimation d'un mélange de lois de Student multivariées à échelles multiples

Alexis Arnaud, Florence Forbes, Benjamin Lemasson, Emmanuel Barbier

Paquet R pour l'estimation d'un mélange de lois de Student multivariées à échelles multiples.

A. Arnaud[a], **F. Forbes**[a], **B. Lemasson**[b] and **E. Barbier**[b]

[a]Équipe Mistis
Inria Grenoble - Rhône-Alpes
655 avenue de l'Europe, 38330
Montbonnot-Saint-Martin, France
prenom.nom@inria.fr

[b]Équipe 5 - perfusion cérébrale
Grenoble Institut des Neurosciences
31 Chemin Fortuné Ferrini, 38706
La Tronche, France
prenom.nom@ujf-grenoble.fr

Mots clefs : Distribution de Student multivariée à échelles multiples, modèle de mélange, classification non supervisée, algorithme EM.

L'utilisation d'un modèle de mélange de lois est une approche statistique classique en classification non-supervisée. Un mélange fréquemment utilisé pour sa simplicité est le mélange gaussien. Cependant, un tel modèle est sensible aux données atypiques. Pour remédier à cela, nous présentons ici le mélange de lois de Student multivariées à échelles multiples, que nous sommes en train d'incorporer au sein d'un paquet R. Comme nous pouvons le voir sur la figure 1, la classification bivariée en 3 groupes avec les lois (a) de Student à échelles multiples [1] permet de retrouver des classes alongées. En effet, ces lois peuvent gérer des queues de lourdeurs différentes selon les directions alors que les lois gaussiennes(b) et les lois de Student multivariées standards sont contraintes à être symétriques.

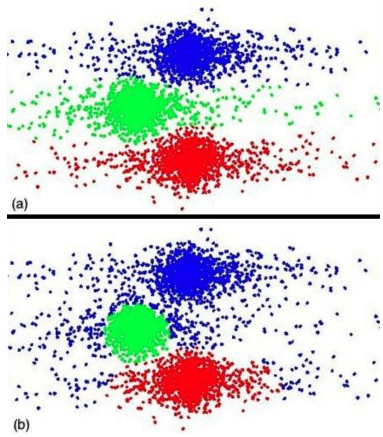


Figure 1

Loi de Student multivariées 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 [1]. 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 $\boldsymbol{Y} \in \mathbb{R}^{\mathrm{M}}$:

$$p_{\mathrm{S}}(\boldsymbol{y}\ ;\ \boldsymbol{\mu},\boldsymbol{T},\nu) = \int_0^\infty \mathcal{N}_{\mathrm{M}}(\boldsymbol{y}\ ;\ \boldsymbol{\mu}, w\,\boldsymbol{T})\,\mathcal{G}\left(w\ ;\ \frac{\nu}{2},\frac{\nu}{2}\right)\mathrm{d}w\ , \tag{1}$$

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 [1] consiste à décomposer la matrice d'échelle 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 :

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

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

Mélange de lois de Student multivariées à échelles multiples

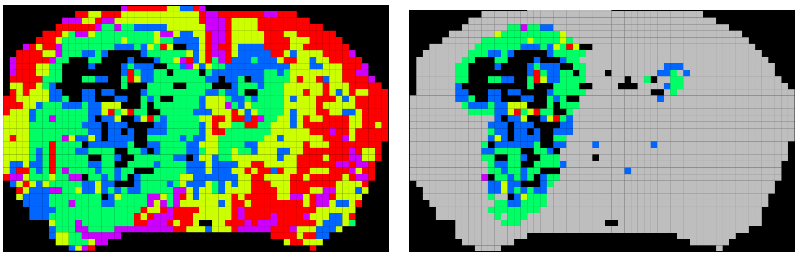
La densité d'un mélange de lois statistiques continues s'explicite simplement comme la somme pondérée des densités de classe. En considérant un mélange à K classes, admettant pour proportions $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_{\text{K}}\}$, nous obtenons la densité suivante pour la variable aléatoire $\boldsymbol{Y} \in \mathbb{R}^{\text{M}}$ issue d'un mélange de lois de Student multivariées à échelles multiples :

$$p_{\text{MSEM}}(\boldsymbol{y} \; ; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{U}, \boldsymbol{D}, \boldsymbol{\nu}) = \sum_{k=1}^{\text{K}} \pi_k \, p_{\text{SEM}}(\boldsymbol{y} \; ; \boldsymbol{\mu}_k, \boldsymbol{U}_k, \boldsymbol{D}_k, \boldsymbol{\nu}_k) \tag{3}$$

L'estimation des paramètres de ce modèle est réalisée via un algorithme d'Expectation-Maximization (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 $k \in \{1, \ldots, \text{K}\}, n = \{1, \ldots, \text{N}\}$. Les détails de l'algorithme EM sont présents dans Forbes et Wraith [1].

Application à la caractérisation de tumeurs cérébrales par IRM multiparamétrique

Nous avons incorporé le modèle de mélange décrit précédement au sein d'un paquet R, non encore disponible sur le CRAN, et nous l'avons testé sur des données d'IRM. Ce jeu de données contient des images IRM multiparamétriques de cerveaux de rats présentant des tumeurs cérébrales (5 paramètres physiologiques sont acquis lors de ces IRM). L'objectif était de délimiter le contour de la tumeur et de la caractériser. Après un ajustement du modèle sur une population de rats sains, l'algorithme a prédit la classe de chaque voxel de nouvelles images IRM présentant des tumeurs. La prédiction de la classe des voxels de ces images est illustrée sur la figure 2a, où nous pouvons voir la tumeur ressortir sur la partie gauche du cerveau, en bleu et vert. Toutefois ces classes ont été ajustées sur une population de référence avec 5 classes et ne permettent pas à elles seules de localiser la tumeur, hors intervention d'un médecin. Ainsi, afin de déterminer les voxels atypiques au sein de cette prédiction, nous utilisons les poids multidimensionnels des lois de Student à échelles multiples utilisées. Nous obtenons ainsi la figure 2b où quasiment seuls les voxels de la tumeur sont sélectionnés (les voxels gris sont non-atypiques par rapport à l'apprentissage). Les résultats ainsi obtenus sont cohérents avec ceux d'une précédente étude [2]. De plus, les centres des classes présentent des valeurs physiologiques moins dégradées par la présence de valeurs extrêmes.



(a) (b)

FIGURE 2

Références

[1] 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*, **24**(6), 971-984.

[2] 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*, **34**(8), 1354-62.