



STATISTIQUE
SCIENCE DES DONNÉES BIOSTATS
UNIVERSITÉ DE MONTPELLIER

Olivier CÔME

Nicolas PRALON

N^o étudiant Olivier CÔME: 22110708

N^o étudiant Nicolas PRALON: 22110681

olivier.come@et u.umontpellier.fr

nicolas.pralon@et u.umontpellier.fr

Master 2

Statistique et Science des Données

Université de Montpellier

Projet

HAX006X : Modèles à variables latentes

Rédigé le 27 mars 2023 en L^AT_EX

Sommaire

1	Introduction	2
2	L'algorithme EM	3
2.1	Implémentation de la fonction simulation	3
2.2	Implémentation de la fonction EM	4
2.3	Étude sur des données simulées	5
2.3.1	Simulation d'un mélange de deux gaussiennes	5
2.3.2	Simulation d'un mélange à quatre gaussiennes	6
2.4	Comparaison de la fonction <i>EM</i> avec <i>mixtools</i> sur des données réelles	7
3	THEME sur des données de Vins	11
4	Conclusion	22
5	Annexes	23
5.1	Script de la fonction <i>simulation</i>	23
5.2	Script de la fonction <i>EM</i>	24
5.3	Script de la fonction <i>plot_distrib</i>	25
6	Bibliographie	26

1 Introduction

Le présent projet s'inscrit dans le cadre de l'unité d'enseignement "HAX006X Modèles à variables latentes". L'objectif est ici d'appliquer sur des données réelles, plusieurs méthodes étudiées durant ce cours. Nous nous intéresserons à deux d'entre elles à savoir la méthode par thèmes et l'algorithme EM (Expectation-Maximization) conçue par Dempster et al., dont l'article est disponible via la source [1].

2 L'algorithme EM

Dans cette section nous allons nous intéresser à l'algorithme EM (Expectation-Maximization) afin d'estimer les paramètres α , μ et σ d'un mélange gaussien. Nous avons implémenté cet algorithme via la fonction *EM* (présente dans notre script *R*) en nous aidant de la source [2]. Afin de tester l'efficacité de notre implémentation, nous allons dans un premier temps l'exécuter sur des données simulées. En effet, dans notre script, nous avons implémenté une autre fonction que nous avons nommée *simulation*. Cette dernière nous permettra de générer de manière aléatoire, un échantillon issu d'un mélange gaussien. Nous décrirons plus en détail cette fonction, en aval. Dans un second temps, nous exécuterons notre algorithme sur de vraies données afin de voir s'il est robuste. Pour cela nous utiliserons le jeu de données galaxies de la librairie MASS et nous le décrirons ultérieurement.

2.1 Implémentation de la fonction simulation

Comme il a été mentionné précédemment nous allons réaliser dans un premier temps une étude sur des données simulées à partir de la fonction *simulation* (le code de son implémentation est disponible en annexes). Décrivons cette dernière, elle prend en argument :

- **dt_param** : Le dataframe contenant les paramètres α , μ et σ
- **n** : La taille de l'échantillon

Elle retourne un vecteur de taille n qui sera l'échantillon du mélange gaussien.

Regardons un plus en détail comment a été conçue cette fonction.

La partie la plus importante et la plus subtile de ce script est celle dans laquelle nous distribuons aléatoirement les $(X_i)_{i \in 1, \dots, n}$ de l'échantillon de sorte à avoir un bon mélange gaussien.

Afin de simplifier les choses, rien de mieux que de prendre un exemple. Dans celui-ci, l'objectif sera de générer un mélange de quatre gaussiennes, ayant pour paramètres respectifs $\theta_1 = (\alpha_1, \mu_1, \sigma_1)$, $\theta_2 = (\alpha_2, \mu_2, \sigma_2)$, $\theta_3 = (\alpha_3, \mu_3, \sigma_3)$ et $\theta_4 = (\alpha_4, \mu_4, \sigma_4)$.

Les $(\alpha_j)_{j \in \{1, \dots, 4\}}$ étant ici des probabilités, nous avons que :

$$\sum_{j=1}^4 \alpha_j = 1$$

La démarche est la suivante :

- On tire $Z \sim \mathcal{U}(0, 1)$
- Si $Z < \alpha_1$ alors $X \sim \mathcal{N}(\mu_1, \sigma_1)$
- Sinon si $\alpha_1 < Z < \alpha_1 + \alpha_2$ alors $X \sim \mathcal{N}(\mu_2, \sigma_2)$
- Sinon si $\alpha_1 + \alpha_2 < Z < \alpha_1 + \alpha_2 + \alpha_3$ alors $X \sim \mathcal{N}(\mu_3, \sigma_3)$
- Sinon si $\alpha_1 + \alpha_2 + \alpha_3 < Z < \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$ alors $X \sim \mathcal{N}(\mu_4, \sigma_4)$

Notez que notre implémentation marche dans le cas général d'un mélange de J gaussiennes avec $J \in \mathbb{N}^*$.

2.2 Implémentation de la fonction EM

La fonction *EM* est sans aucun doute celle la plus importante de cette section, il est donc primordial de la décrire (le code de son implémentation est disponible en annexes).

Tout d'abord, pour implémenter cette dernière, nous nous sommes fortement aidés du pseudo-code suivant :

Algorithm 1 L'algorithme EM (Dempster et al., 1977).

Entrée(s) : $N \in \mathbb{N}$, $\hat{\theta}_0 \in \Theta$, un jeu de données $x_1 \dots x_n$;

Initialisation ;

1: $k := 1$;

2: **Tant que** $K < N + 1$ **faire**

3: **ETAPE E :** Calculer la fonction $Q(\theta; \hat{\theta}_{k-1}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{\theta}_{k-1}} [\log f(X_i, Z_i, \theta) | X_i = x_i]$;

4: **ETAPE M :** $\hat{\theta}_k = \operatorname{argmax}_{\theta} Q(\theta; \hat{\theta}_{k-1})$;

5: $k \leftarrow k + 1$;

6: **fin du Tant que ;**

7: **retourner** $\hat{\theta}_N$;

La fonction *EM* prend en argument :

- *dt_init*, le dataframe contenant les paramètres initiaux (α_{init} , μ_{init} , σ_{init})
- *X*, les données (réelles ou simulées) issues d'un mélange gaussien
- *K* le nombre d'itérations souhaitées pour l'algorithme

Elle retourne un dataframe contenant les valeurs des paramètres estimées par l'algorithme, à savoir α , μ et σ .

Les formules que nous avons utilisées pour calculer l'étape E et M et qui sont présentées ci dessous sont issues de la source [2].

- Lors de l'étape E nous déterminons la probabilité $\mathbb{P}_{\hat{\theta}}(Z = j | X = X_i)$ via la formule suivante :

$$\mathbb{P}_{\hat{\theta}}(Z = j | X = X_i) = \frac{\alpha_j \times \gamma_{\mu_j, j, v_j}}{\sum_{k=1}^J \alpha_k \times \gamma_{\mu_k, v_k}}$$

- Lors de l'étape M, nous déterminons les estimations des estimateurs du maximum de vraisemblance ($\hat{\alpha}_j, \hat{\mu}_j, \hat{\sigma}_j$) via les formules suivantes :

$$\begin{aligned} \hat{\alpha}_j &= \frac{1}{n} \sum_{i=1}^n (Z = j | X = X_i) \\ \hat{\mu}_j &= \frac{\sum_{i=1}^n X_i (Z = j | X = X_i)}{\sum_{i=1}^n (Z = j | X = X_i)} \\ \hat{v}_j &= \frac{\sum_{i=1}^n (X_i - \hat{\mu}_j)^2 (Z = j | X = X_i)}{\sum_{i=1}^n (Z = j | X = X_i)} \end{aligned}$$

Comme en témoigne la section 3.3 (Preuve de la croissance de la vraisemblance d'une itération à l'autre) de la source [2], il est théoriquement prouvé que l'algorithme permet de faire croître la log-vraisemblance des observations en les paramètres itérativement créés. Cependant, il est important de notifier le fait qu'il n'existe pas de convergence de la suite de paramètres établie par l'algorithme EM. En effet, ces derniers peuvent rester bloqués dans des extremas locaux. On comprend donc qu'il est primordial de choisir de bons paramètres initiaux afin de ne pas être confronté à ce problème. Dans la section consacrée à l'étude sur de vraies données, nous expliciterons la procédure qui a été mise en place pour choisir ces paramètres initiaux.

2.3 Étude sur des données simulées

Cette sous-section sera consacrée à l'étude menée sur des données simulées à partir de notre fonction *simulation*.

2.3.1 Simulation d'un mélange de deux gaussiennes

Nous avons ici décidé de générer un échantillon de taille 100 issu d'un mélange de deux gaussiennes de lois respectives $\mathcal{N}(\mu_1, \sigma_1) = \mathcal{N}(50, 11)$ et $\mathcal{N}(\mu_2, \sigma_2) = \mathcal{N}(220, 50)$. La densité associée à cet échantillon a été estimée de manière non paramétrique à partir d'une méthode à noyau et elle a été tracée sur la figure 1.

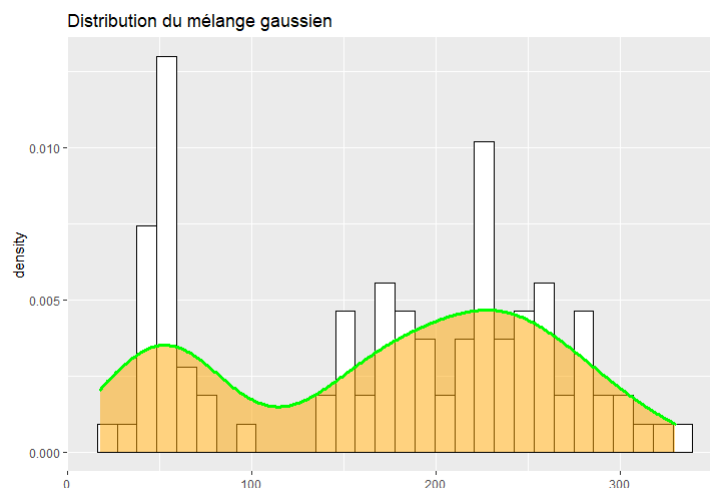


FIGURE 1 – Densité d'un mélange à 2 gaussiennes estimée par une méthode à noyau

Nous avons ensuite appliqué notre algorithme (fonction EM) sur cet échantillon. Le tableau 1 contient les valeurs des vrais paramètres de ce mélange simulé. Comme il a été mentionné précédemment, le choix des paramètres initiaux est crucial si l'on veut que l'algorithme estime correctement les paramètres du mélange. Le tableau 2 contient les valeurs des paramètres initiaux utilisés. Comme vous pouvez le voir en observant ces deux tableaux, nous avons choisi des paramètres initiaux assez proches des vrais paramètres (sauf pour la variance du 2ème mélange) de manière à ne pas être bloqué dans des extremas locaux. Les résultats obtenus par notre algorithme sont affichés sur la capture d'écran de la figure 2.

Comme on peut le voir sur la figure 2, les valeurs estimées par notre implémentation sont proches de celles des vrais paramètres (voir tableau 1). Ces résultats nous montrent donc que notre fonction *EM* fonctionne correctement, ce qui est rassurant.

	α	μ	σ
Paramètres du 1er mélange	0.4	50	11
Paramètres du 2ème mélange	0.6	220	50

TABLE 1 – Vrais paramètres du mélange

	α_{init}	μ_{init}	σ_{init}
Paramètres du 1er mélange	0.2	30	21
Paramètres du 2ème mélange	0.8	280	160

TABLE 2 – Paramètres initiaux

	mixtureParameters	alpha	mu	sigma
1	parameters of Mixture1	0.4584696	49.08584	9.983414
2	parameters of Mixture2	0.5415304	227.40612	55.429273

FIGURE 2 – Paramètres du mélange gaussien estimés par notre fonction *EM*

2.3.2 Simulation d'un mélange à quatre gaussiennes

Nous avons ici décidé de générer un échantillon de taille 1000 issu d'un mélange de quatre gaussiennes de lois respectives :

- $\mathcal{N}(\mu_1, \sigma_1) = \mathcal{N}(35, 11)$
- $\mathcal{N}(\mu_2, \sigma_2) = \mathcal{N}(350, 22)$
- $\mathcal{N}(\mu_3, \sigma_3) = \mathcal{N}(720, 32)$
- $\mathcal{N}(\mu_4, \sigma_4) = \mathcal{N}(1198, 55)$

La densité associée à cet échantillon a été estimée de manière non paramétrique à partir d'une méthode à noyau et elle a été tracée sur la figure 3.

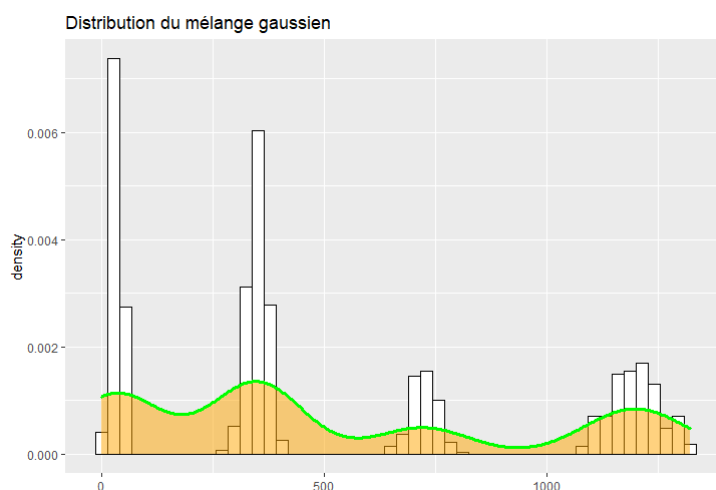


FIGURE 3 – Densité d'un mélange à 4 gaussiennes estimée par une méthode à noyau

Nous avons ensuite appliqué notre algorithme (fonction *EM*) sur cet échantillon. Le tableau 3 contient les valeurs des vrais paramètres de ce mélange simulé. Comme il a été mentionné précédemment, le choix des paramètres initiaux est crucial si l'on veut que l'algorithme estime correctement les paramètres du mélange. Le tableau 4 contient les valeurs des paramètres initiaux utilisés. Comme vous pouvez le voir en observant ces deux tableaux, nous avons choisi des paramètres initiaux assez proches des vrais de manière à ne pas être bloqué dans des extremas locaux. Les résultats obtenus par notre algorithme sont affichés sur la capture d'écran de la figure 4.

Comme on peut le voir sur la figure 4, les valeurs estimées par notre implémentation sont proches de celles des vrais paramètres (voir tableau 3). Ces résultats nous confortent une fois de plus dans l'idée que notre fonction *EM* fonctionne correctement.

	α	μ	σ
Paramètres du 1er mélange	0.30	35	11
Paramètres du 2ème mélange	0.33	350	22
Paramètres du 3ème mélange	0.15	720	32
Paramètres du 4ème mélange	0.22	1198	55

TABLE 3 – Vrais paramètres du mélange

	α	μ	σ
Paramètres du 1er mélange	0.33	30	10
Paramètres du 2ème mélange	0.30	370	25
Paramètres du 3ème mélange	0.17	717	30
Paramètres du 4ème mélange	0.20	1238	57

TABLE 4 – Paramètres initiaux choisis

	mixtureParameters	alpha	mu	sigma
1	parameters of Mixture1	0.296	35.40873	11.22125
2	parameters of Mixture2	0.318	350.93412	21.71632
3	parameters of Mixture3	0.141	722.17459	32.45836
4	parameters of Mixture4	0.245	1199.41438	49.72699

FIGURE 4 – Paramètres du mélange gaussien estimés par notre fonction *EM*

2.4 Comparaison de la fonction *EM* avec *mixtools* sur des données réelles

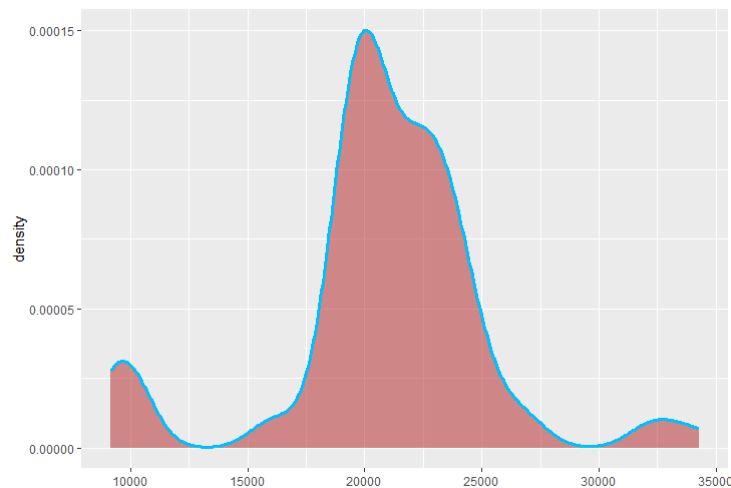
Dans cette section, l'étude sera menée sur des données réelles. Nous utiliserons le jeu de données *galaxies* provenant de la librairie *MASS* de *R*. Dans un premier temps, nous allons estimer les paramètres du mélange associés à ce dataset via l'utilisation de notre implémentation de l'algorithme EM (la fonction *EM*). Nous estimerons ensuite une seconde fois les paramètres de ce même mélange mais cette fois-ci, en utilisant la fonction *normalmixEM* prédéfinie de *R* qui est disponible via le package *mixtools*. Nous avons choisi d'utiliser cette librairie car l'algorithme EM y est implémenté et il est utilisé par la fonction *normalmixEM* pour estimer les paramètres d'un mélange. Le fait de comparer les résultats de notre fonction avec ceux obtenus par celle prédéfinie de *R* nous permettra d'évaluer la performance et la robustesse de notre implémentation sur de vraies données.

Le jeu de données *Galaxies* est un vecteur numérique qui représente les vitesses en km/s (kilomètre par seconde) de 82 galaxies. La figure 5 est un extrait de ce jeu de données.

```
> galaxies
[1] 9172 9350 9483 9558 9775 10227 10406 16084 16170 18419 18552 18600 18927
[14] 19052 19070 19330 19343 19349 19440 19473 19529 19541 19547 19663 19846 19856
[27] 19863 19914 19918 19973 19989 20166 20175 20179 20196 20215 20221 20415 20629
[40] 20795 20821 20846 20875 20986 21137 21492 21701 21814 21921 21960 22185 22209
[53] 22242 22249 22314 22374 22495 22746 22747 22888 22914 23206 23241 23263 23484
[66] 23538 23542 23666 23706 23711 24129 24285 24289 24366 24717 24990 25633 26690
[79] 26995 32065 32789 34279
```

FIGURE 5 – Extrait du jeu de données *galaxies*

Comme il a été mentionné dans la section 2.2, si l'on veut obtenir de bonnes estimations pour les paramètres du mélange, il est primordial de sélectionner correctement les paramètres qui serviront de conditions initiales pour l'algorithme. Nous allons donc détailler la stratégie qui a été mise en place pour sélectionner ces derniers. Étant donné que nous ne connaissons pas la vraie densité associée à ces données, nous avons utilisé dans un premier temps une méthode d'estimation non paramétrique (à noyau) afin d'estimer cette dernière. La figure 6 représente la courbe de densité estimée par la méthode à noyau.

FIGURE 6 – Densité du dataset *galaxies* estimée par une méthode à noyau

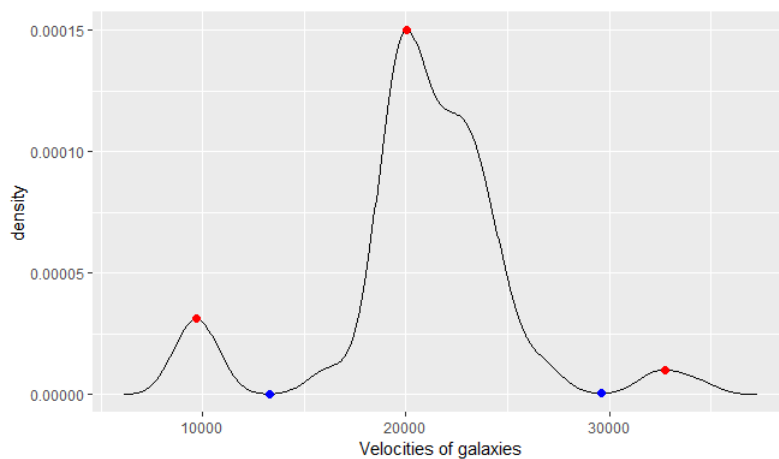
En observant la figure 6, on peut facilement distinguer 3 pics principaux. Un premier vers 9800 sur l'axe des abscisses, un deuxième vers 21000 et un 3ème vers 32000. Il semblerait aussi y avoir une sorte de pic vers 24000 mais celui-ci n'étant pas clairement visible nous ne le considérerons pas. On supposera donc pour la suite de l'étude que nous sommes dans le cas d'un mélange à 3 composantes.

Il nous faudra donc déterminer les paramètres :

- α_1, α_2 et α_3 qui sont les proportions associées aux 3 gaussiennes
- μ_1, μ_2 et μ_3 qui sont les moyennes des 3 gaussiennes
- σ_1, σ_2 et σ_3 qui sont les écarts-types des 3 gaussiennes

Pour commencer, nous dirons que les paramètres $(\alpha_j)_{j \in \{1, \dots, 3\}}$ qui seront utilisés dans les conditions initiales seront tous égaux, c'est à dire que $(\alpha_1)_{init} = (\alpha_2)_{init} = (\alpha_3)_{init} = \frac{1}{3}$ (car il faut que $\sum_{i=1}^3 \alpha_j = 1$).

Pour la recherche des paramètres $(\mu_j)_{j \in \{1, \dots, 3\}}$ et $(\sigma_j)_{j \in \{1, \dots, 3\}}$ initiaux, nous allons nous aider de la figure 7.

FIGURE 7 – Extremums locaux de la densité estimée du dataset *galaxies*

À l'aide de la fonction *find_peaks()* du package *ggpmisc* et de la fonction *local.min.max()* du package *spatialEco*, nous avons pu déterminer les extremums locaux de la courbe de densité estimée des données *galaxies*. Les maximums locaux sont représentés par les points rouges et les minimums locaux par les points bleus sur la figure 7.

Les paramètres $(\mu_j)_{j \in \{1, \dots, 3\}}$ qui seront utilisés dans les conditions initiales seront donc les abscisses respectives de ces 3 maximums locaux. Nous aurons donc :

- $(\mu_1)_{init} = 9698.471$
- $(\mu_2)_{init} = 20050.850$
- $(\mu_3)_{init} = 32717.292$

Les $(\alpha_j)_{j \in \{1, \dots, 3\}}$ et les $(\mu_j)_{j \in \{1, \dots, 3\}}$ initiaux ayant été déterminés, il ne nous reste plus qu'à trouver les $(\sigma_j)_{j \in \{1, \dots, 3\}}$ initiaux. Nous allons les déterminer en calculant les écarts-types de chacun des 3 pics. Tout d'abord, nous scindons nos données en 3 intervalles car il y a 3 pics. C'est à ce moment là qu'interviennent les minimas locaux. En effet, c'est eux qui vont justement nous permettre de délimiter notre jeu de données en 3 intervalles.

Le premier intervalle sera composé des abscisses allant du début du jeu de données jusqu'à l'abscisse du premier minimum local. Le deuxième intervalle quant à lui sera composé des valeurs des abscisses allant du premier minimum local jusqu'au deuxième minimum local. Enfin le dernier intervalle contiendra les abscisses allant du deuxième minimum local jusqu'à la fin du jeu de données. Dans la liste à puces ci-dessous nous avons expliciter les bornes *inf* et *sup* de chacun des trois intervalles :

- **Première intervalle** : [6166.482; 13291.36]
- **Deuxième intervalle** : [13291.36; 29611.58]
- **Troisième intervalle** : [29611.58; 37284.52]

En appliquant la fonction $sd()$ sur le premier intervalle nous obtiendrons $(\sigma_1)_{init}$. En appliquant la fonction $sd()$ sur le deuxième intervalle nous obtiendrons $(\sigma_2)_{init}$. Nous procéderons de même pour avoir $(\sigma_3)_{init}$. Au final, nous obtenons que :

- $(\sigma_1)_{init} = 2083.124$
- $(\sigma_2)_{init} = 4737.603$
- $(\sigma_3)_{init} = 2241.339$

L'ensemble des paramètres initiaux qui ont été déterminés précédemment sont résumés dans le tableau 5.

	α_{init}	μ_{init}	σ_{init}
Paramètres du 1er mélange	$\frac{1}{3}$	9698.471	2083.124
Paramètres du 2ème mélange	$\frac{1}{3}$	20050.850	4737.603
Paramètres du 3ème mélange	$\frac{1}{3}$	32717.292	2241.339

TABLE 5 – Paramètres initiaux

Maintenant que nous avons toutes nos conditions initiales, nous sommes en mesure d'exécuter notre implémentation de l'algorithme *EM* sur les vraies données *galaxies*. Les paramètres du mélange estimés par notre fonction *EM* sont visibles sur la capture d'écran (figure 8). Nous les avons également résumé dans le tableau 6 pour une meilleure visibilité.



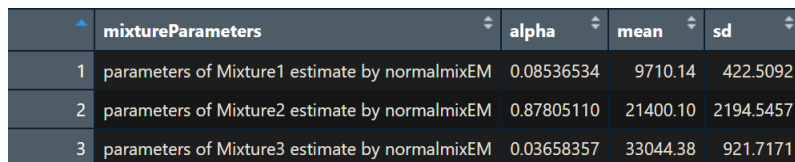
	mixtureParameters	alpha	mu	sigma
1	parameters of Mixture1	0.08536534	9710.14	422.5092
2	parameters of Mixture2	0.87805110	21400.10	2194.5457
3	parameters of Mixture3	0.03658357	33044.38	921.7171

FIGURE 8 – Paramètres estimés par notre fonction *EM*

	α	μ	σ
Paramètres du 1er mélange estimés par <i>EM</i>	0.08536534	9710.14	422.5092
Paramètres du 2ème mélange estimés par <i>EM</i>	0.87805110	21400.10	2194.5457
Paramètres du 3ème mélange estimés par <i>EM</i>	0.03658357	33044.38	921.7171

TABLE 6 – Paramètres estimés par notre fonction *EM*

Maintenant que nous avons les paramètres du mélange, estimés par notre implémentation, comparons les avec ceux obtenus en utilisant la fonction *normalmixEM* de *R* dans laquelle est implémenté l'algorithme EM. La figure 9 est une capture d'écran dans laquelle sont stockés tous les paramètres estimés par la fonction *normalmixEM* de *R*. Nous les avons également résumé dans le tableau 7 pour une meilleure visibilité.



	mixtureParameters	alpha	mean	sd
1	parameters of Mixture1 estimate by normalmixEM	0.08536534	9710.14	422.5092
2	parameters of Mixture2 estimate by normalmixEM	0.87805110	21400.10	2194.5457
3	parameters of Mixture3 estimate by normalmixEM	0.03658357	33044.38	921.7171

FIGURE 9 – Paramètres estimés par *normalmixEM*

	α	μ	σ
Paramètres du 1er mélange estimés par <i>normalmixEM</i>	0.08536534	9710.14	422.5092
Paramètres du 2ème mélange estimés par <i>normalmixEM</i>	0.87805110	21400.10	2194.5457
Paramètres du 3ème mélange estimés par <i>normalmixEM</i>	0.03658357	33044.38	921.7171

TABLE 7 – Paramètres estimés par notre fonction *EM*

En regardant les tableaux 6 et 7, on constate que les valeurs des paramètres estimées avec notre fonction *EM* et celles estimées avec la fonction *normalmixEM* sont exactement les mêmes. Cela nous montre donc que notre implémentation de l'algorithme EM est robuste et qu'elle peut être aussi utilisée sur des données réelles de mélanges. Notre fonction *EM* fournira donc de bonnes estimations aussi bien sur des données simulées que sur des données réelles à partir du moment où les conditions initiales sont correctement choisies.

3 THEME sur des données de Vins

Dans cette section, à l'inverse de la partie précédente, nous nous plaçons dans un cadre non aléatoire avec un modèle à composantes, les données pour l'illustration sont issues de 21 Vins de Loire, ces Vins étant d'écrit par 2 variables qualitatives (Label : Bourgueuil, Chinon, Saumur), (Sol : En1, Env2, Référence, Env4) ainsi que 29 variables quantitatives décrivent des caractéristiques comme l'odeur, le goût des Vins. Précisément on organise les variables quantitatives suivants les blocs de thèmes suivants : odeur, arôme, goût, note de goût, plante et composition (chimique).

- *Odeur*, qui comprend les variables *Odor.Intensity.before.shaking*, *Odor.Intensity*, *Quality.of.odour* décrivant les caractéristiques olfactives des Vins.
- *Arome*, qui renvoie aux variables *Aroma.quality.before.shaking*, *Aroma.intensity*, *Aroma.persistency*, *Aroma.quality* faisant référence aux caractéristiques sensorielles des Vins telles que l'intensité, la qualité, la présence des arômes.
- *Goût*, comprenant les variables renvoyant au goût fuité, amère, acide ou épicé des Vins : *Fruity.before.shaking*, *Fruity.dity*, *Spice.before.shaking*, *Spice*, *Bitterness*
- *Note_goût*, renvoyant aux notes / caractéristiques gustatives des Vins, le thème comprend les variables comme l'équilibre des saveurs *Balance*, *Smooth*, *Attack.intensity*, *Intensity*, *Harmony*, *Typical*
- *Plante*, ce thème correspond à des caractères de plantes utilisées dans la fermentation des Vins : *Flower.before.shaking*, *Visual.intensity*, *Nuance*, *Surface.feeling*
- *Composition*, ce dernier theme renvoie aux compositions chimiques des fuits utilisées, telles la teneur alcool ou l'astringence : *Astringency*, *Alcohol*, *Phenolic*, *Plante*, *Flower*

On s'intéresse, pour ces variables décrivant les Vins, à prédire les variables caractéristiques de l'odeur des Vins ainsi que d'en tirer des composantes explicatives. Puisque ces caractéristiques d'odeur sont à la fois liées aux variables d'arôme et de goût, elle-même dépendante des variables de composition, de note de gout et de caractéristiques de plantes utilisées dans la fermentation des Vins, l'utilisation de THEME est dans ce cas adéquate à cette recherche. Nous obtenons les deux équations thématiques suivantes :

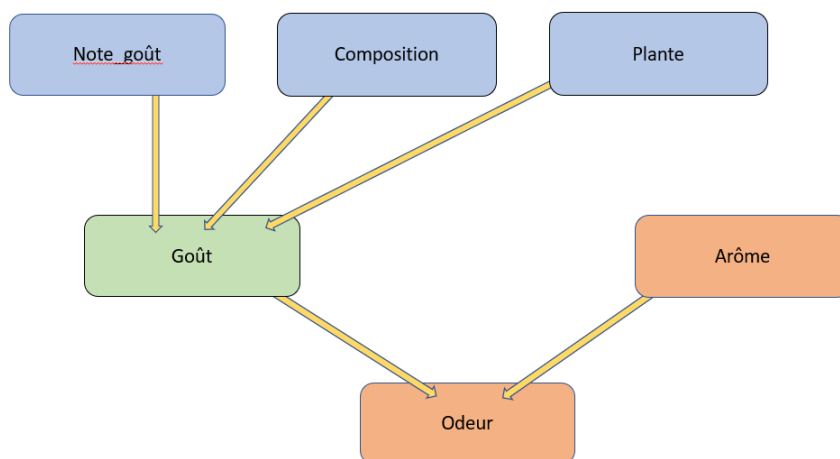


FIGURE 10 – Equations thématiques

Dans un premier temps, nous avons décidé d'appliquer THEME sans *cross-validation backward* en partant du modèle avec 2 composantes pour le thème *Odeur*, 2 composantes pour le thème *Arôme*, 3 composantes pour le thème *Goût*, 3 composantes pour le thème *Note_goût*, 2 composantes pour le thème *Plante* et 2 composantes pour le thème *Composition*. Puisque que l'on cherche à analyser les caractéristiques des odeurs des Vins, on ne prend pas en compte les variables qualitatives de *Sol* et *Env*, de plus la variable *Overall.quality* ne rentrant dans aucune thématique.

Voici les resultats d'ajustement des prédictions aux données que nous obtenons pour les deux équations :

	Fruity.before.shaking	Spice.before.shaking	Fruity	Spice	Acidity
1	0,511697479856819	0,752138687341394	0,691005137250852	0,70787953425451	0,34319068712367

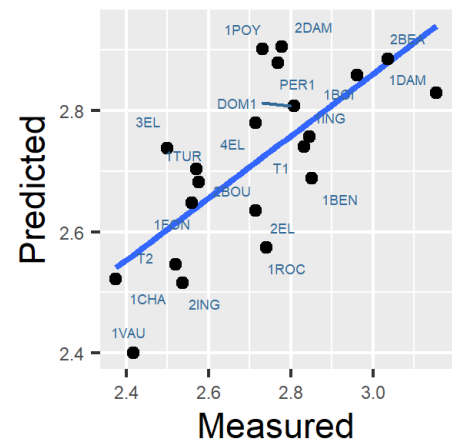
(a) R2 de regression de l'équation 1

	Odor.Intensity.before.shaking	Odor.Intensity	Quality.of.odour
1	0,837752921086672	0,757978208109395	0,80214187926411

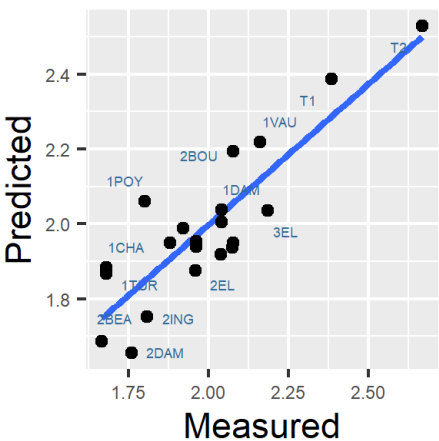
(b) R2 de regression de l'équation 1

FIGURE 11 – R2 de regression linéaire des thèmes Odeur et Goût

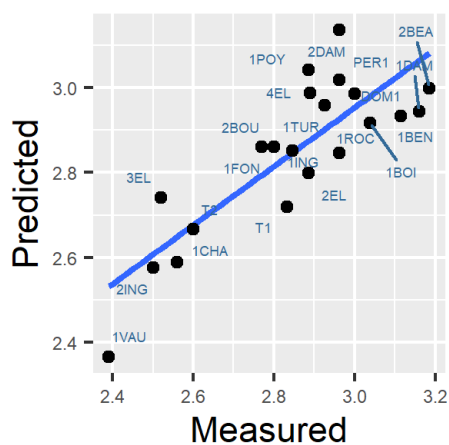
On peut constater pour la régression (equation 1) des variables du thème *Goût* sur les composantes des thèmes *Note_goût*, *Plante*, *Composition* sont relativement bonnes hormis pour la regression des variables *Fruity.before.shaking* et notamment *Acidity*. Si l'on représente les variables prédites avec celles mesurées on peut se rendre compte que les points sont assez bien répartis autour de la bissectrice mais que certains d'entre eux en sont également assez éloignés. En particulier on observe certains points extrêmes et des points assez éloignés de la bissectrice pour les deux variables *Fruity.before.shaking*, *Acidity* :



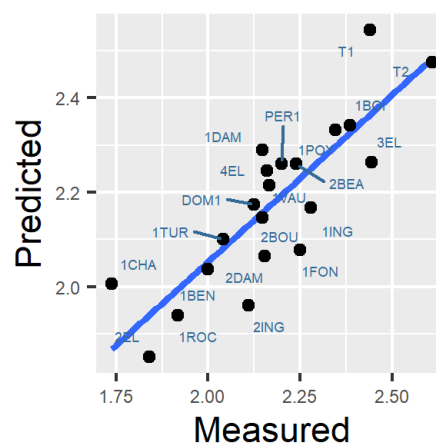
(a) Regression de Fruity.before.shaking, équation 1



(b) Regression de Spice.before.shaking, équation 1



(a) Regression de Fruity, équation 1



(b) Regression de Spice, équation 1

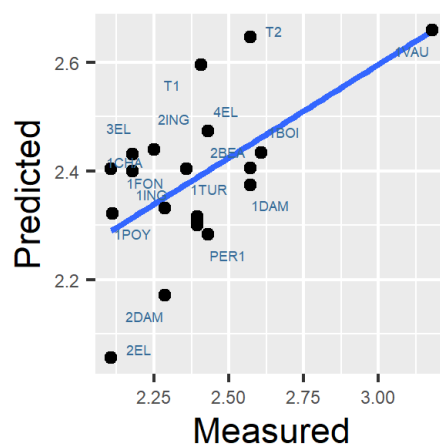
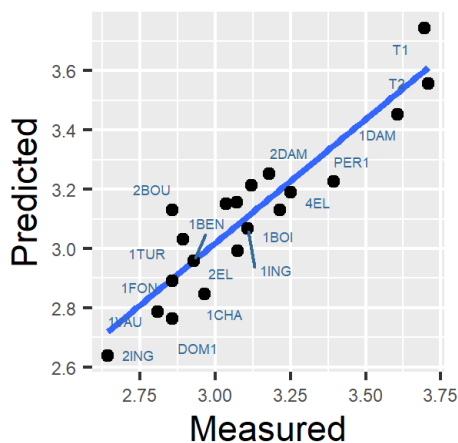
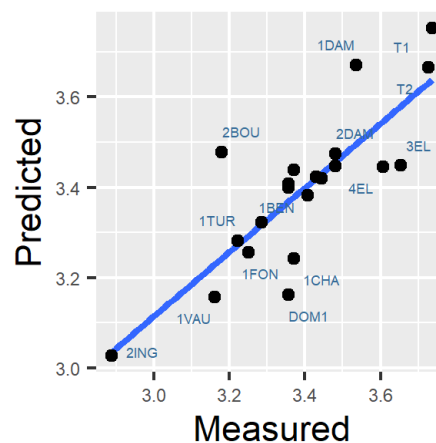


FIGURE 12 – Regression de Acidity, équation 1

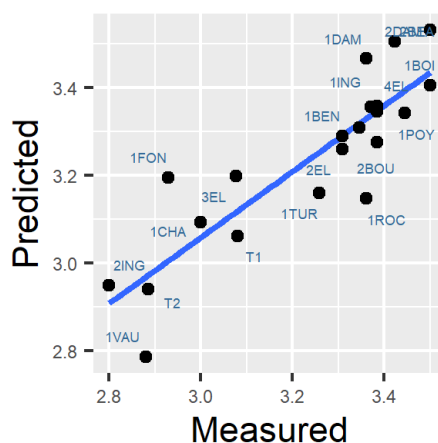
Concernant la seconde équation, on se rend compte que les variables *Odor.Intensity.before.shaking*, *Odor.Intensity*, *Quality.of.odour* sont très bien prédites, on peut l'observer sur l'image 11 et les graphiques suivants :



(a) Regression de *Odor.Intensity.before.shaking*, équation 2



(b) Regression de *Odor.Intensity*, équation 2



(c) Regression de *Quality.of.odour*, équation 2

FIGURE 13 – Regression linéaire des variables *Odeur*, équation 2

En complémentaire de l'ajustement de régression des variables des thèmes *Odeur* et *Goût*, on s'intéresse également à la qualité de prédiction que nous étudierons avec le modèle avec *cross-validation backward*. Néanmoins on peut donner les deux erreurs de prédictions pour les deux équations, sur l'ensemble des données :

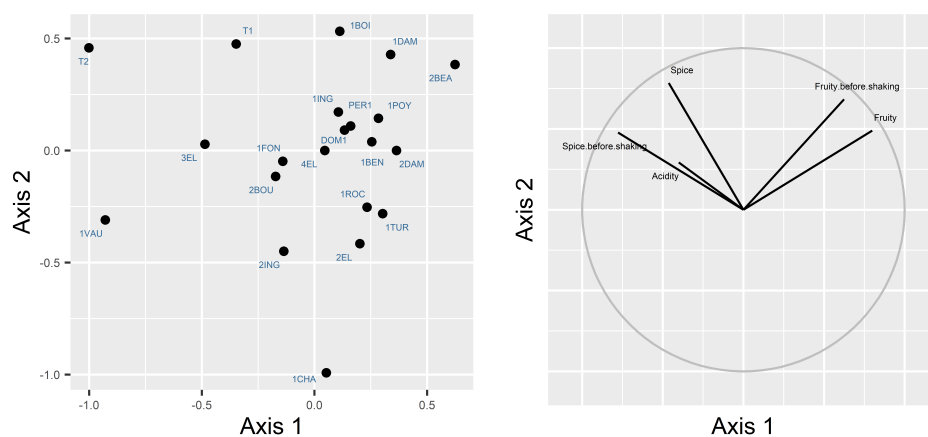
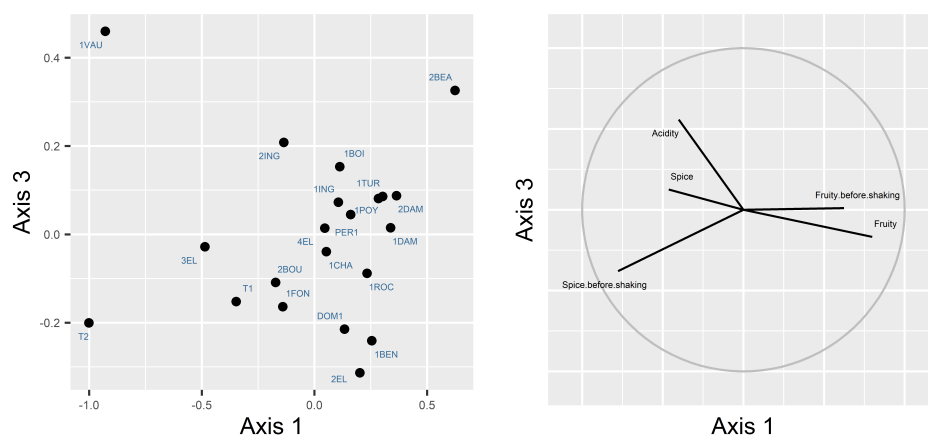
```
> PRESS_eq1
Fruity.before.shaking Spice.before.shaking Fruity Spice Acidity
1 0.3946544 0.2847325 0.3034634 0.2487501 0.7559759
```

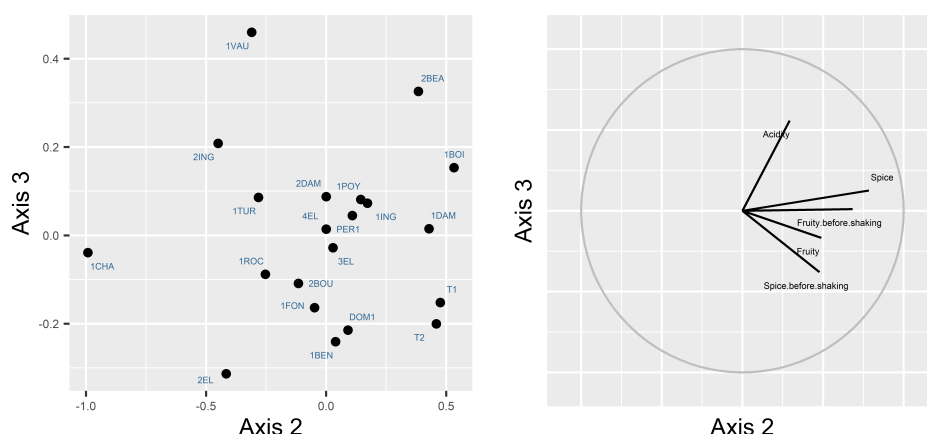
(a) PRESS statistique de l'équation 1

```
> PRESS_eq2
Odor.Intensity.before.shaking Odor.Intensity Quality.of.odour
1 0.2557405 0.2704233 0.2410205
```

(b) PRESS statistique de l'équation 2

A présent et afin d'interpréter les prédictions suivant les variables du modèle, on se propose de représenter les graphiques directs et duals des composantes des équations 1 et 2.

(a) Représentation direct et dual du thème *Goût* plan 1-2(b) Représentation direct et dual du thème *Goût* plan 1-3

FIGURE 14 – Représentation directe et dual du thème *Goût* plan 2-3

On peut constater pour les deux premiers axes, que les variables sont assez bien représentées sans pour autant être directement corrélées ou anti-corrélées à l'un des axes. Toutefois on peut observer que les variables *Fruity.before.shaking*, *Fruity* et *Spice.before.shaking*, *Spice*, *Acidity* sont décorréliées sur le plan. L'analyse indépendamment de l'axe 1 et 2 avec l'axe 3 permet d'observer si les variables sont décorréliées sur tous les plans. Ici on s'aperçoit sur le plan 1 et 3 que les variables *Spice*, *Fruity* sont anti-corrélées à l'inverse du plan 2 et 3. Les représentations avec l'axe 3 ne sont pas des meilleures, mais elles représentent principalement les informations manquantes sur l'axe 1 et 2, qui sont les points communs et différents des variables. Seule la variable *Acidity* n'a pas l'air d'être mieux représentée par une composante que par une autre, il faut les 3. On pouvait se douter de ce dernier résultat avec la valeur du R^2 de régression.

On peut se pencher sur la part que les variables des thèmes explicatifs jouent dans la prédiction des deux équations, celles-ci nous permettront de déterminer les variables les plus pertinentes dans l'analyse :

code	Fruity.before.shaking	Spice.before.shaking	Fruity	Spice	Acidity
csf.	2,284352167	1,933148784	2,16026466	1,99986568	2,94150482
Attack.intensity	-0,024479176	0,354495889	-0,1600716	0,30406329	0,23277735
Balance	0,165624171	-0,379032984	0,31470638	-0,0976057	-0,100562
Smooth	0,080893904	-0,225247733	0,16377212	-0,0901019	-0,0830973
Bitterness	-0,168127253	0,007379511	0,17481959	-0,0313557	0,1427474
Intensity	0,077041979	0,274857541	-0,0486836	0,37666247	0,26150193
Harmony	0,088231738	-0,071184338	0,10299295	0,06483781	0,02842664
Typical	0,090859555	-0,371716669	0,25545032	-0,1987024	-0,1562683
Flower.before.shaking	0,376874762	-0,752641374	0,04771815	0,29919738	0,0153673
Visual.intensity	-0,05632121	0,184452695	-0,0422684	-0,068563	-0,1260921
Nuance	-0,030327639	0,100949028	-0,0235541	-0,0374582	-0,0706936
Surface.feeling	-0,080948552	0,285715208	-0,0708112	-0,1053719	-0,2166721
Flower	-0,868011876	0,867844271	-0,7196034	-0,8071462	-0,3392439
Plante	0,071198161	-0,058111867	0,13102501	0,06440169	-0,0170752
Phenolic	0,126306146	-0,139668904	0,03097824	0,11929687	0,09534609
Astringency	0,089876085	-0,105724064	-0,0128724	0,08576331	0,08962023
Alcohol	0,163773027	-0,182731891	0,03117757	0,1549098	0,12923552

FIGURE 15 – Coefficients de régression linéaire des variables explicatives, équation 1

Dans la table ci-dessus, nous retrouvons les coefficients de la régression multiple expliquant le plus les variables du thème *Odeur* en fonction des variables des thèmes *Note_goût*, *Plante*, *Composition*. Nous avons surligné en jaune les coefficients en valeur absolue les plus élevés.

- Pour le thème *Note_goût*, ce sont les variables *Balance*, *Bitterness* qui ont le plus d'impact sur les variables de Fruit.
- Pour le thème *Plante*, c'est la variable *Flower.before.shaking* qui a le plus d'impact, et surtout les caractères épicés du thème *Goût* cette variable renseigne sur les plantes utilisées dans la fer-

mentation des Vins, il doit donc s'agir d'épices qui ont été ajoutées dans les Vins. Le caractère d'intensité va également marqué les variables *Spice.before.shaking*, *Spice*.

- Pour le thème *Composition*, l'Alcool va jouer un rôle dans la régression des variables de fruits et d'acidité.

On retrouve avec l'analyse des coefficients les points distincts sur le plan 1-2 entre les variables *Fruity.before.shaking*, *Fruity* expliquées par les variables *Balance*, *Bitterness*, et les variables *Spice.before.shaking*, *Spice*, *Acidity* expliquées par *Flower.before.shaking*, *Intensity*. On retrouve également un caractère commun à l'ensemble des variables, comme on a pu le constater sur les plans 1-3 et 2-3, qui est ici marqué par la variables *Flower* et dans une moindre mesure *Flower.before.shaking*.

On fait de même pour la régression linéaire du thème *Odeur* :

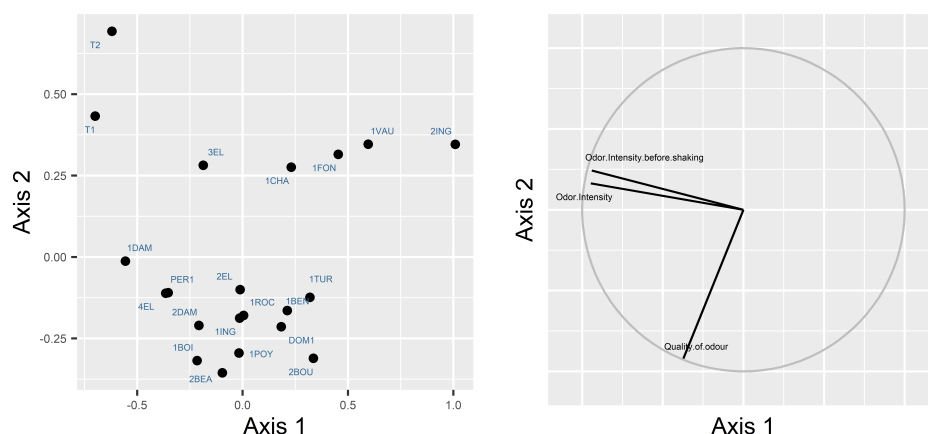


FIGURE 16 – Représentation directe et dual du thème *Odeur* plan 1-2

Dans la projection des variables sur le plan dual 1-2, on s'aperçoit que les variables sont extrêmement bien représentées, de plus quelles ont l'air fortement liées aux axes 1 et 2. Plus précisément les variables *Odor.Intensity.before.shaking*, *Odor.Intensity* sont très liées à l'axe 2, et la variable *Quality.of.odour* à l'axe 1. Toute fois ces variables sont quelque peu expliquées par les axes 1 et 2 respectivement. Leurs représentations nous informe que les caractères de qualité et d'intensité sont décorélées sur le plan, ce qui peut se comprendre. On peut donc donner un sens aux deux composantes, l'axe 1 représenterait une variable de qualité du Vins, et l'axe 2 représenterait un caractère d'intensité aromatique.

Leurs représentations sont bonnes, car les variables *Odor.Intensity.before.shaking*, *Odor.Intensity* sont relativement identiques.

Afin de mieux comprendre cette composantes, l'analyse des coefficients de régression linéaire sur les composantes explicatives nous permettra de déterminer les variables à les plus impactantes dans les prédictions.

code	Odor.Intensity.before.shaking	Odor.Intensity	Quality.of.odour
cst.	-0,165584014	1,252228947	1,188952079
Aroma.quality.before.shaking	0,484254227	0,352006646	0,105890616
Aroma.intensity	0,628726562	0,423335768	0,063477732
Aroma.persistency	0,372122368	0,281035509	0,104520169
Aroma.quality	-0,58306709	-0,250423906	0,253436539
Fruity.before.shaking	-0,093411699	-0,174395326	0,150014094
Spice.before.shaking	0,318803503	0,310486791	-0,23889274
Fruity	-0,017486887	-0,094253837	0,13703763
Spice	0,057771703	-0,09253877	0,074412879
Acidity	0,004491617	-0,012005918	-0,020509651

FIGURE 17 – Coefficients de régression linéaire des variables explicatives, équation 2

On s'aperçoit que ce sont les mêmes variables explicatives qui rentrent en jeu dans l'analyse de *Odor.Intensity.before.shaking* principalement liées aux variables du thème *Arôme*. Concernant la variable *Quality.of.odour* elle est expliquée, comme on peut l'attendre, par les variables du thème *Goût*, mais aussi par la variable de qualité et de présence du thème *Arôme*. On peut alors expliquer distinctement les composantes obtenues avec ces deux thèmes.

Comme entendu, nous allons effectuer les mêmes démarches d'analyse que pour la sous-section précédente en effectuant désormais une *Cross-validation backward* en partant du même modèle, ce qui enlève des composantes au fur et à mesure en testant quelle est l'erreur de prédiction pour chaque modèle testé. Nous pouvons ainsi déterminer le meilleur des ces modèles. Par raison de temps de calcul nous avons décidé de modifier les équations thématiques, ce qui nous permettra également de comparer le modèle précédent avec le modèle à une équation, dont toutes les variables sont explicatives du thème *Odeur*. Nous conservons le même nombre de composantes pour chacun des thèmes, et nous obtenons par *Cross-validation backward* les résultats suivants :

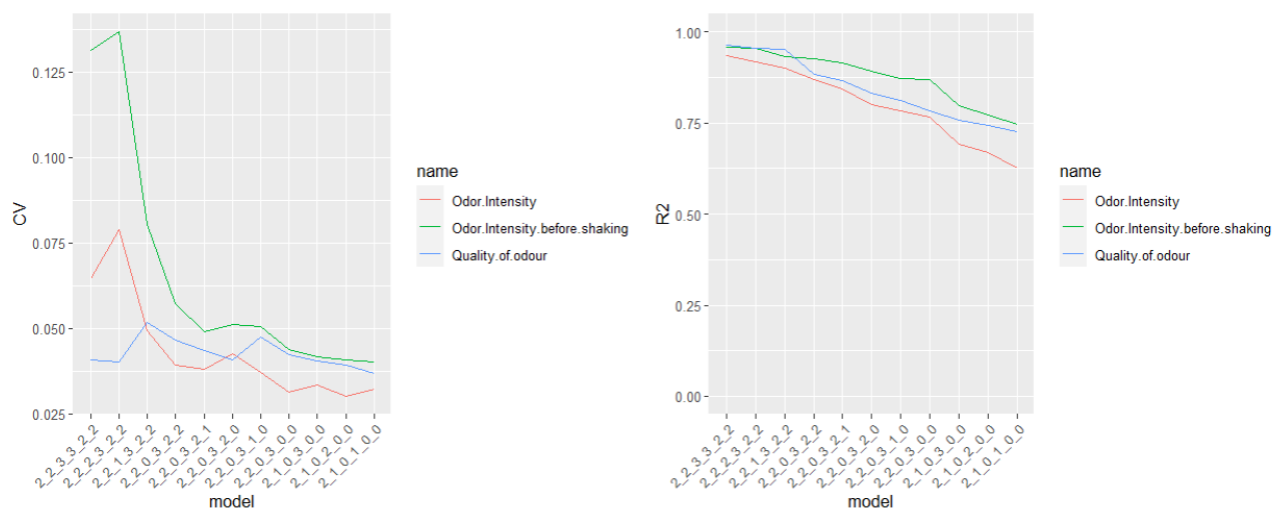


FIGURE 18 – Cross-validation backward et R^2 de chaque modèle, pour chaque variable du thème *Odeur*

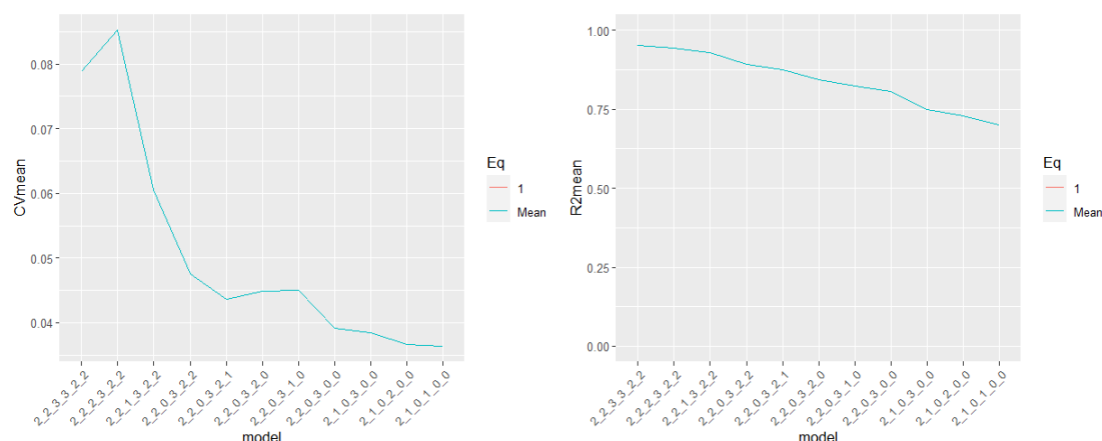


FIGURE 19 – Moyenne des Cross-validation backward et R^2 de chaque modèle du thème *Odeur*

A noter que la *Cross-validation backward* ne test pas tous les modèles, mais réalise une sélection arrière des composantes sélectionnées par Cross-validation, à partir du modèle définie.

Pour chacune des variables du thème *Odeur*, la valeur de la *PRESS* des modèles à une allure plutôt décroissante lorsque la qualité d'ajustement de la prédiction des modèles diminue. Il faut donc choisir un compris du modèle à choisir, entre qualité d'ajustement et de prédiction.

Ici on peut se rendre compte que la valeur de la *PRESS* de la variable n'a pas la même allure pour chacune des variables du thème. On peut par exemple choisir le modèle *223322* avec le meilleur ajustement R^2 pour la variable *Quality.of.odour*.

Pour tenter d'avoir les meilleurs résultats, on peut choisir un modèle adapter pour chaque variable. Cependant, par soucis de simplicité et de compréhension dans l'analyse des variables du thème *Odeur*, on décide de choisir un même modèle pour les 3 variables. Ce faisant, on regarde la valeur moyenne des Cross-validation et des R^2 (figure 19).

La variation de la valeur moyenne de la *PRESS* étant très minime quant à la variation moyenne du R^2 , on peut se permettre sans grande perte de qualité de prédiction de considérer les modèles entre *223322* et *220321*. Toute fois on perd beaucoup d'information pour les variables *Quality.of.odour* et *Odor.Intensity.before.shaking*, Afin d'être le plus partimonieux on a décidé de choisir le modèle *220322*. On s'aperçoit avec la validation croisée, même si il s'agit de variation très faible de la *PRESS*, que ce sont les composantes du thème *Goût* qui sont supprimées au fur et à mesure, on peut le comprendre puisqu'ici les thèmes *Note_goût*, *Plante* et *Composition* sont directement "liés" au thème *Odeur*. On se rend compte également que l'arôme et très pertinent dans l'odeur des Vins de Loire.

Dans l'idée de comparer le modèle sélectionné avec celui sans *Cross-validation*, *backward*, on peut comparer la valeur de la *PRESS* avec l'erreur de prédiction du modèle précédent, sur l'ensemble des données.

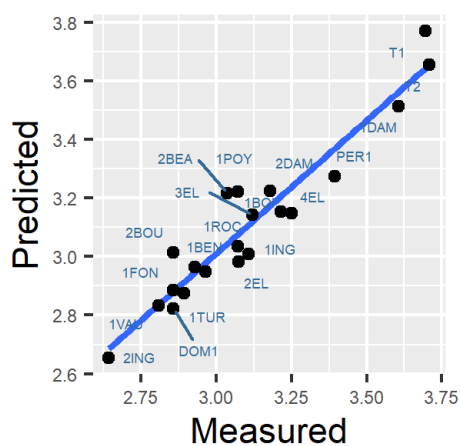
Rem : Pour le modèle *223322* sans cross-validation, il s'agit uniquement de l'erreur de prédiction et non de celle déterminé avec validation croisée. Cette valeur nous permet de nous donner une indication de performance entre les deux modèles.

```
> mean(PRESS_eq2[1,])
[1] 0.2557281
```

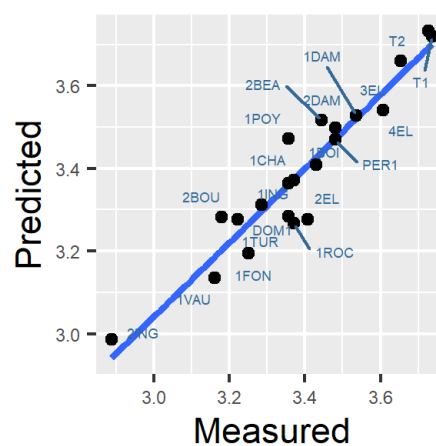
FIGURE 20 – Moyenne de la *PRESS* de chaque variable du thème *Odeur*

On peut observer que le qualité de prédiction est nettement supérieur pour le modèle à 1 équation. Ceci certainement dû au fait qu'on ait choisi un modèle à une équation.

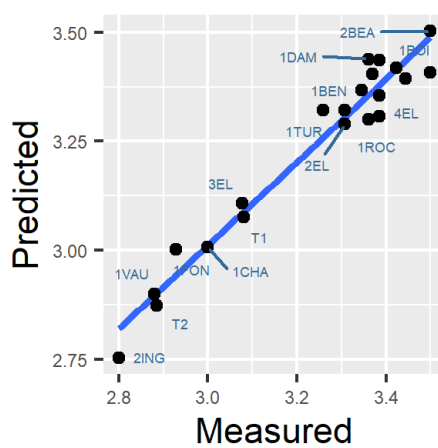
Pour le modèle par *Cross-validation*, *backward* on a ainsi obtenu de meilleurs résultats de prédiction et d'ajustement des données. En voici les représentations des prédictions :



(a) Regression de Odor.Intensity.before.shaking



(b) Regression de Odor.Intensity



(c) Regression de Quality.of.odour

FIGURE 21 – Regression linéaire des variables *Odeur*

On constate que la régression du thème *Odeur* est excellente, les points de chaque graphique longuent bien la bissectrice. On pouvait s'attendre à cette qualité en observant la valeur du R^2 de la figure 19.

Par soucis de lisibilité, on représente seulement le graphique direct et dual du thème *Odeur*, nous permettant dans notre but d'analyser les composantes informatives du thème.

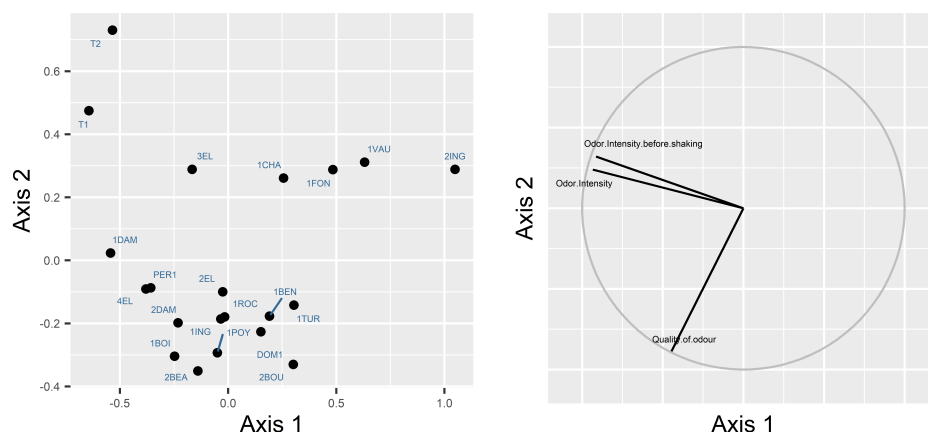


FIGURE 22 – Représentation directs et dual du thème *Odeur* plan 1-2

Et on s'aperçoit qu'il s'agit sans grande surprise d'une représentation direct et dual très similaire à celles vues pour le modèle à 2 équations, les variables *Odor.Intensity.before.shaking* et *Odor.Intensity* étant assez semblables. Il va ainsi s'agir de bien expliquer chacune des variables du thème afin de donner un bon sens aux composantes.

On représente ci-dessous les coefficients de régression linéaire du thème sur l'ensemble des variables des autres thèmes :

code	Odor.Intensity.before.shaking	Odor.Intensity	Quality.of.odour
cst.	0,334788479	0,277049692	0,640645671
Aroma.quality.before.shaking	0,481269226	0,132371439	0,018442057
Aroma.intensity	0,737920662	0,247995664	-0,148132595
Aroma.persistency	0,28855999	0,068128703	0,055083255
Aroma.quality	-0,391425192	-0,229195543	0,461094415
Attack.intensity	0,439706493	0,246652646	-0,241065433
Balance	-0,309737367	-0,077784591	0,206041057
Smooth	-0,122915496	-0,00510615	0,08730648
Bitterness	0,122318921	0,069355413	0,046654723
Intensity	0,197914362	0,153311882	-0,093529385
Harmony	-0,017699614	0,052921329	0,02960145
Typical	-0,202555236	-0,038301624	0,141000472
Flower.before.shaking	0,231694031	0,394485892	-0,177418371
Visual.intensity	0,02895705	-0,045774285	0,010960514
Nuance	0,029295779	-0,050375538	0,012505653
Surface.feeling	0,055349402	-0,078694857	0,017883643
Flower	-0,111243582	-0,24277657	-0,203806906
Plante	0,14796397	0,30794946	0,257559997
Phenolic	-0,190430421	0,191271028	0,199436351
Astringency	-0,233916262	0,10387077	0,126545804
Alcohol	-0,378872778	0,14812226	0,186789529

FIGURE 23 – Coefficients de régression linéaire sur variables explicatives

4 Conclusion

Au vu des résultats et des comparaisons faites entre notre fonction *EM* et celle *normalmixEM* du package *mixtools*, nous pouvons conclure que notre implémentation fournira de bonnes estimations aussi bien sur des données simulées que sur des données réelles de mélanges à condition bien évidemment de choisir correctement les conditions initiales. Une piste d'amélioration possible de notre fonction serait d'y ajouter une fonctionnalité de recherche automatique de paramètres initiaux. D'après la littérature, plusieurs techniques peuvent être envisagées. Parmi celles qui reviennent le plus souvent, on retrouve la méthode des kmeans.

5 Annexes

5.1 Script de la fonction *simulation*

Ci dessous, l'export de code de la fonction *simulation*.

```
simulation = function(dt_param, n=100){
  X = rep(NA,n) #echantillon
  vect_alpha = dt_param[,2]
  vect_mean = dt_param[,3]
  vect_sd = dt_param[,4]
  for(i in 1:n){
    Z = runif(1)
    if (Z <= vect_alpha[1]){
      X[i] = rnorm(1, vect_mean[1], vect_sd[1])
    }else{
      k = 1
      l = 2
      Bool = FALSE
      cumul_alpha = cumsum(vect_alpha)
      while(Bool == FALSE){
        if((cumul_alpha[k]<=Z) & (cumul_alpha[l]>=Z)){
          Bool = TRUE
          param_index = l
        }
        k = k+1
        l = l+1
      }
      X[i] = rnorm(1, vect_mean[param_index], vect_sd[param_index])
    }
  }
  return(X)
}
```

5.2 Script de la fonction *EM*

Ci dessous, l'export de code de la fonction *EM*.

```
EM = function(dt_init, X, K){
  J = dim(dt_init)[1]
  n = length(X)
  data_stateE = param_State_E(n, J)

  for(k in 1:K){
    vect_alpha = dt_init[,2] #de longueur J
    vect_mean = dt_init[,3]
    vect_sd = dt_init[,4]
    # vecteur contenant la somme des des numerateurs de P_thetat(j|X = X_i)
    # pour chaque valeur de l echantillon
    v = rep(0,n)

    # Etape E
    for(j in 1:J){
      # On remplit le tableau param_State_E contenant P_thetat(j|X=X_i)
      data_stateE[,j] = vect_alpha[j]*dnorm(X,vect_mean[j],vect_sd[j])
      v = v+data_stateE[,j]
    }
    for(j in 1:J){
      data_stateE[,j] = data_stateE[,j]/v
    }

    # Etape M
    H = data_stateE
    for(col in 2:4){ #on met a jour le dt_init
      for(ind in 1:J){

        # on met a jour les alpha
        if(col == 2){
          dt_init[,col][ind] = mean(H[,ind])
        }
        # on met a jour les mu
        if(col == 3){
          dt_init[,col][ind] = (sum(X*H[,ind]))/(sum(H[,ind]))
        }
        # on met a jour les sigma
        if(col == 4){
          dt_init[,col][ind] = sqrt((sum( (X-rep(dt_init[,col-1][ind],n))^2
                                           *H[,ind] ))/sum(H[,ind]))
        }
      }
    }
  }
  new_df = dt_init
  colnames(new_df) = c('mixtureParameters', 'alpha', 'mu', 'sigma')
  return(new_df)
}
```


5.3 Script de la fonction *plot_distrib*

Ci dessous, l'export de code de la fonction *plot_distrib*.

Cette fonction permet d'afficher l'histogramme des données ainsi que la courbe de densité estimée associée à ces données. Cette courbe de densité sera superposée à l'histogramme.

```
plot_distrib = function(df, X){  
  data_distrib = data.frame(gaussian_mixture = X)  
  ggplot(data_distrib, aes(x=data_distrib[, 'gaussian_mixture'])) +  
    geom_histogram(aes(y=..density..), colour="black", fill="white", bins = 50) +  
    geom_density(alpha=.5, color = "green", fill="orange", size=1.2) +  
    ggtitle("Distribution du melange gaussien") + xlab("")  
}
```

6 Bibliographie

- [1] Dempster A.P., Laird N. M., Rubin D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, Vol. 39, 1, 1-38
- [2] Frédéric Santos (2015). L'algorithme EM : une courte présentation <https://members.loria.fr/moberger/Enseignement/AVR/Exposes/algo-em.pdf>