STA203: Apprentissage statistique

Etude de classification du genre musical

Anthony Kalaydjian – Mathieu Occhipinti

Avril – Mai 2023

Contents

Introduction

Le monde de la musique est riche et diversifié, avec de nombreux genres différents qui ont évolué au fil du temps. A défault de pouvoir identifier précisemment ce qui distingue un genre d'un autre, des techniques d'apprentissage statistique peuvent être mises en place pour pouvoir déterminer le genre d'un titre à partir de méta-variables qui lui sont associées.

Ces méta-variables sont calculées à partir du signal de la musique, ce sont exclusivement des données récupérées sur l'analyse spectral du morceaux.

On s'intéressera, dans l'étude suivante à la classification binaire d'un morceaux comme étant soit du Jazz soit du Classique.

Nous aborderons entre autre trois méthodes de classification, à savoir la régression logistique, la régression ridge et enfin la classification par méthode des K-plus-proches voisins.

L'objectif sera de déterminer le modèle qui convient le mieux à la structure de nos données.

I Étude des données et régression logistique

1 Analyse descriptive

1.1 Analyse uni et bi-variée

Le dataset contient 191 variables. Pour réaliser l'analyse univariée, il sera nécessaire de choisir quelques variables à étudier.

On peut afficher l'évolution de la forme des boxplots de la variables ASE, en fonction des indices de cette dernière, c'est à dire en fonction de la bande de fréquence sur lesquelles elles mesurent chacune l'ASE.

On remarque d'abord que les valeurs médianes du boxplots pour les genres Classiques et Jazz sont très proches, et évoluent similairement selon les indices, en grandissant jusqu'aux variables d'indices 20 puis en diminuant.

On note aussi que pour les basses et les très hautes fréquences, l'ASE des morceaux de Jazz est un peu plus élevé que l'ASE des titres Classiques. C'est moins clair dans l'entre-deux.

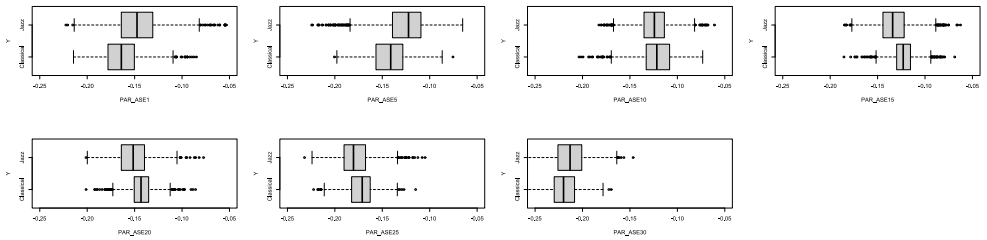

Figure 1: Evolution de la variable ASE en fonction de son indice

L'étude bi-variée sera aussi éffectuée sur des sous-ensembles de variables.

On peut par exemple s'intéresser aux variables SFM d'un côté, et des variables ASE de l'autre.

La matrice de corrélation des variables SFM montre des corrélations fortes au niveau de la diagonale de la matrice, et qui s'éteignent plus elles s'en éloignent. Ceci signifie que les variables SFM d'indices succéssifs seront plus corrélés que des variables d'indices éloignés. Ce n'est pas étonnant, étant donné que deux variables succéssives vont mesurer la même donnée physique sur des fenêtre de fréquence voisines.

On observe le même comportement sur les variables ASE, avec même des anticorrélations lorsque l'on s'éloigne assez de la diagonale.

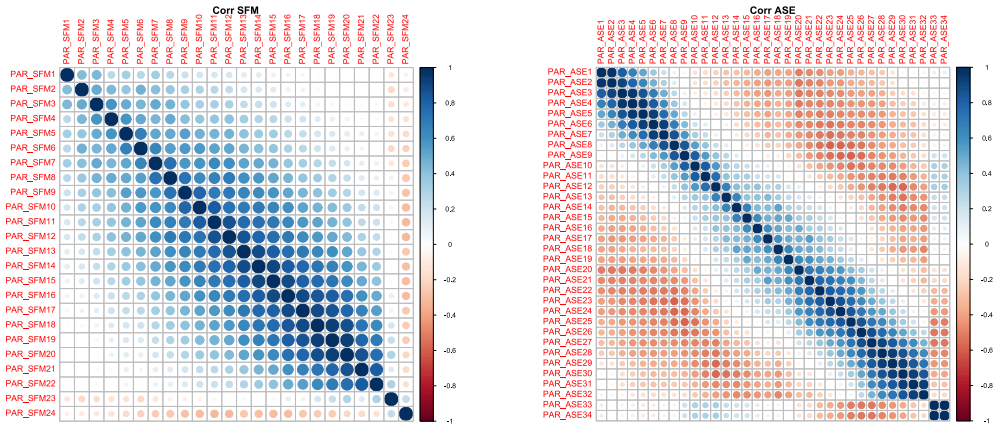

Figure 2: Matrice de corrélation des variables SFM et ASE

On peut aussi tracer la matrice de corrélation des 5 premières variables SFM, ainsi que des 5 premières variables ASE.

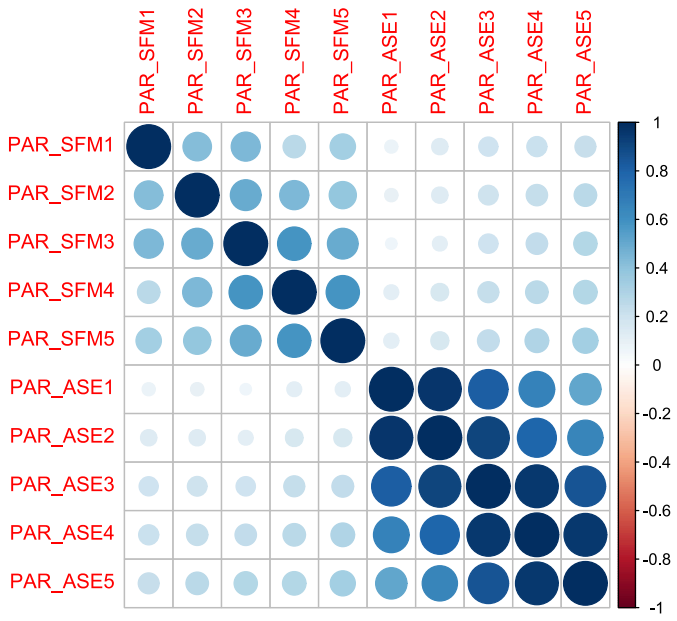

Figure 3: Matrice de corrélation des couples 5 premieres variables ASE, SFM

Cette matrice a la particularité d'être quasi-diagonale par bloc. Ceci signifie que les variables ASE et SFM sont décorrélées.

1.2 Cas des variables PAR_SC_V et PAR_ASC_V

En observant les distributions de ces variables, on remarque que leur distribution ne semblent pas gaussienne au contraire d'autres variables de notre modèle. Dès lors, on décide d'appliquer une transformation log à nos deux variables afin d'obtenir des distributions gaussiennes.

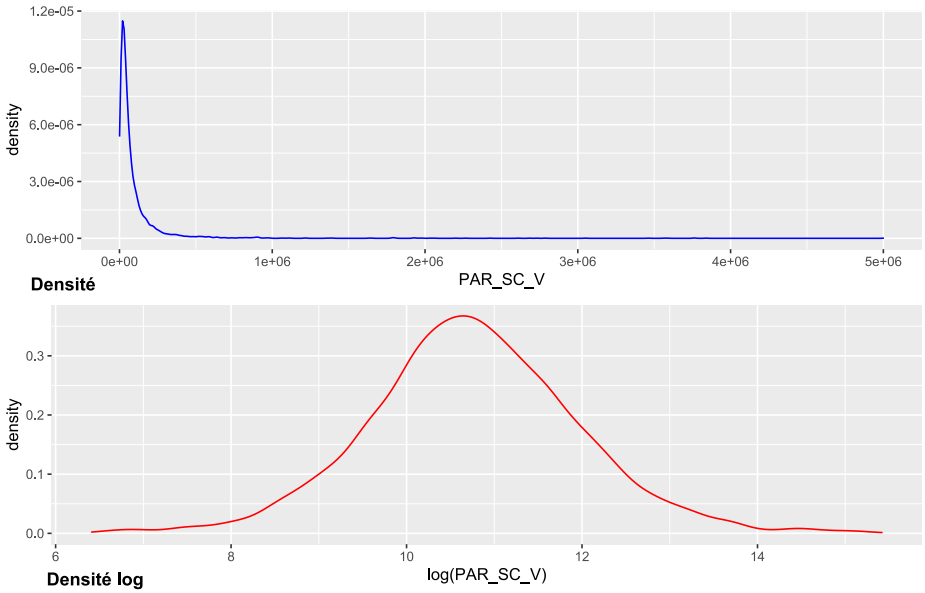

Figure 4: Densité de la variables SC_V et de sa log-transformée

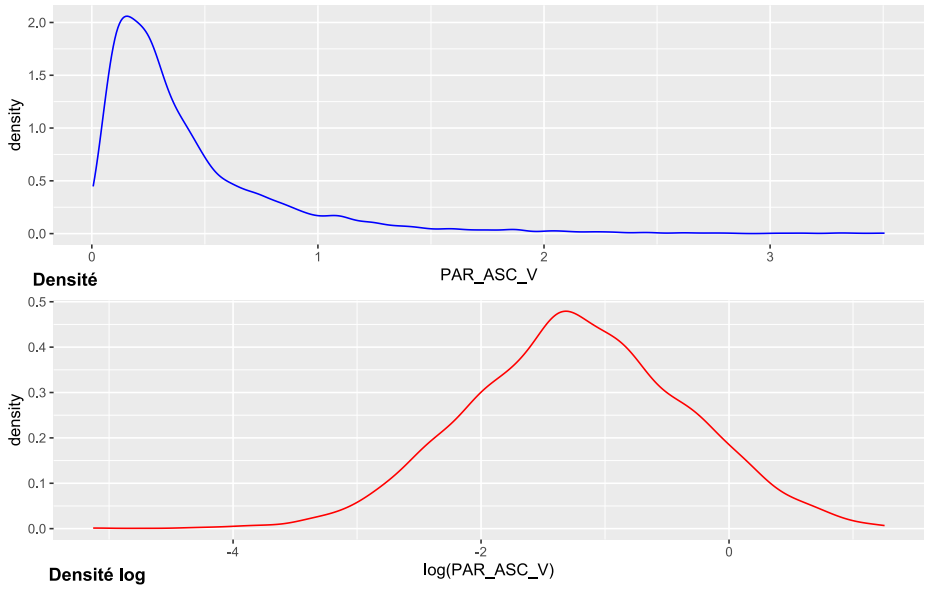

Figure 5: Densité de la variables ASC_V et de sa log-transformée

On remarque bien que les distributions n'étaient pas gaussiennes avant passage au logarithme et qu'elles se rapprochent fortement de distributions gaussiennes après.
On appliquera donc la transformation dans la suite de notre étude.

1.3 Cas des variables dupliquées

La description du dataset nous indique que les variables 148 à 167 sont les mêmes que les variables 128 à 147.

Etant donné qu'elles n'apportent pas d'information complémentaire, il sera judicieux de les éliminer lors de l'étude pour pouvoir simplifier le modèle, sans perte de performance.

1.4 Cas des variables très corrélées

Le cas des variables très corrélées est problématique puisqu'en régression linéaire, chaque coefficient est interprété comme la variation moyenne de la variable réponse par rapport à la variable associée à ce coefficient, en supposant toutes les autres variables constantes. Cela signifie que nous supposons que nous sommes capables de modifier les valeurs d'une variable prédictive donnée sans changer les valeurs des autres variables prédictives.

Cependant, lorsque deux variables prédictives ou plus sont fortement corrélées, il devient difficile de modifier une variable sans en changer une autre. Il est donc difficile pour le modèle de régression d'estimer la relation entre chaque variable prédictive et la variable de réponse indépendamment, car les variables prédictives ont tendance à changer à l'unisson.

Ce problème reste valable dans le cas de la classification binaire où l'on fait correspondre la probabilité de classer un individu dans une certaine classe à un régresseur linéaire que l'on fait passer dans une fonction lien.

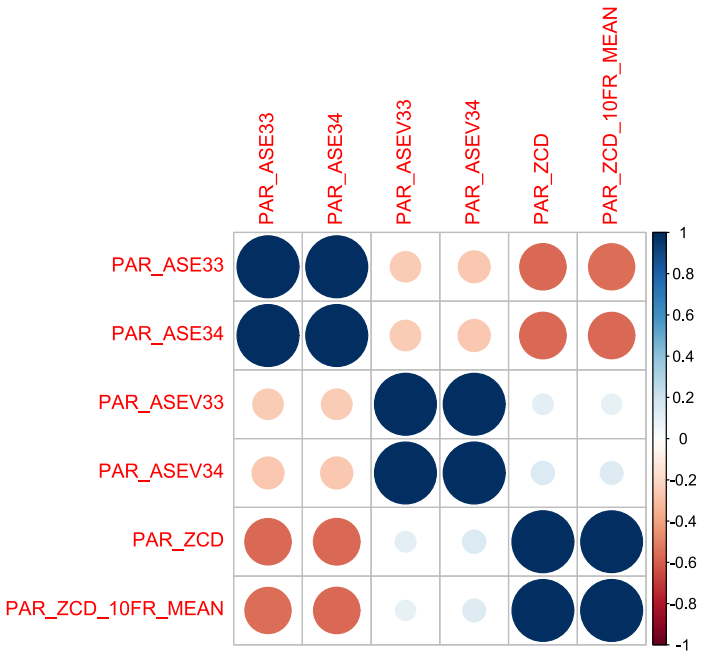

Figure 6: Matrice de corrélation des variables fortement corrélées

On remarque dans la figure 6 que les deux premiers couples de variables très corrélées sont en fait les deux dernières mesures associées respectivement aux variables PAR_ASE et PAR_ASEV.

Le dernier couple de corrélation très élevée montre que la variable PAR_ZCD est très corrélée avec PAR_ZCD_10FR_MEAN dont le nom semble indiquer qu'il s'agit d'une moyenne des PAR_ZCD.

On veillera à bien retirer à chaque fois l'une des deux variables très corrélées, en effet les garder augmenterait la dimmension et la complexité du modèle, sans pour autant apporter de l'information utile.

On retirera par exemple les variables PAR_ASE34, PAR_ASEV34 et PAR_ZCD_10FR_MEAN.

1.5 Cas des variables PAR_ASE_M, PAR_ASE_MV, PAR_SFM_M et PAR_SFM_MV

Les variables PAR_ASE_M, PAR_ASE_MV, PAR_SFM_M et PAR_SFM_MV ont toutes pour suffixes M, ce qui pourrait nous faire penser qu'elles ne représentent chacune que la moyenne de l'ensemble des variables qu'elles désignent.

Ainsi, la variable PAR_ASE_M représenterait par exemple la moyenne des PAR_ASE_i pour $i \in [1, 34]$. Il en serait de même pour les autres variables.

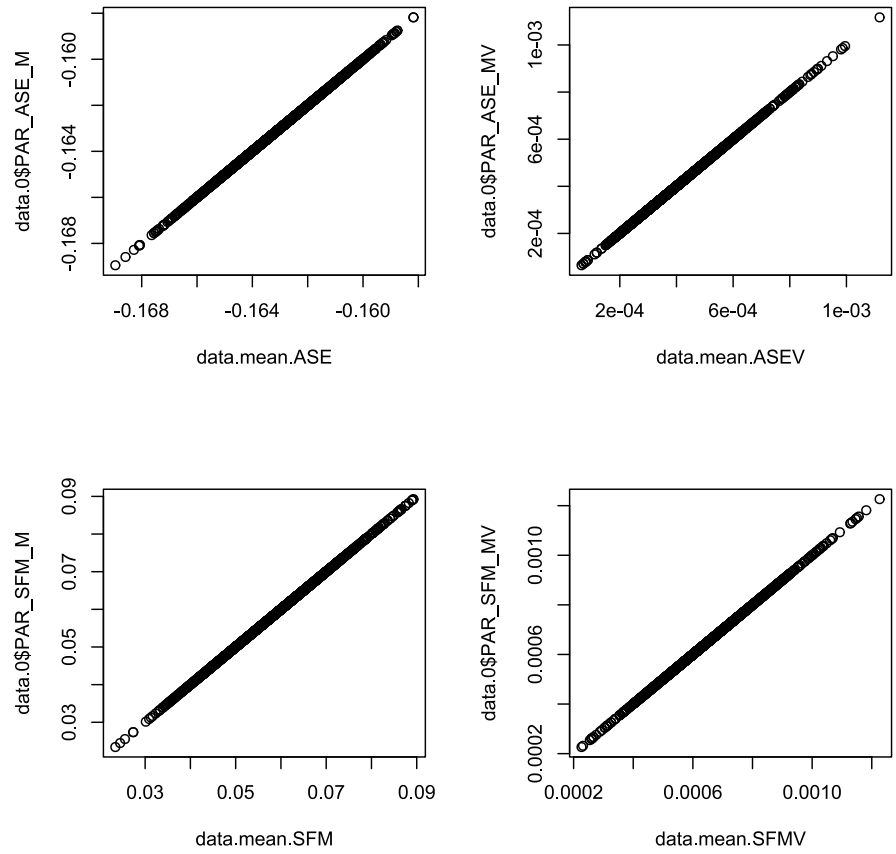

Figure 7: Analyse des variables PAR_ASE_M, PAR_ASE_MV, PAR_SFM_M et PAR_SFM_MV

Ceci est en effet vérifié, la figure 7 montre une parfaite correspondance entre les 4 variables suffixées par M et les moyennes calculées sur l'ensemble des variables qu'elles désignent.

Ainsi, ces variables n'apportent pas d'information et l'on pourra les supprimer de nos données.

2 Estimation des modèles

Maintenant que nous avons analysé et étudié les données, nous allons estimer différents modèles de régression logistiques, se basant chacun sur un jeu de variables différents. L'objectif sera de comparer ces différents modèles et ainsi choisir le "meilleur".

Voici donc les 5 modèles retenus:

- Mod0 formé des variables PAR_TC, PAR_SC, PAR_SC_V, PAR_ASE_M, PAR_ASE_MV, PAR_SFM_M, PAR_SFM_MV.
- ModT contenant toutes les variables retenues dans l'analyse précédente.
- Mod1 formé par toutes les variables significatives au niveau 5% dans ModT.
- Mod1 formé par toutes les variables significatives au niveau 20% dans ModT.
- ModAIC obtenu par sélection de variables stepwise sur critère AIC à partir d'un modèle initial ModT.

Chaque modèle est entraîné sur un jeu de donnée d'apprentissage puis ces performances seront testés par la suite sur un jeu de donnée test. Les deux jeux de donnée (apprentissage et test) sont obtenus à partir d'un tirage aléatoire sur le jeu de donnée fourni.

Nous avons gardé deux tiers des données pour le jeu d'apprentissage et le reste pour le test.

3 Courbes ROC

Une fois nos modèles entraînés, nous pouvons analyser nos modèles à l'aide des courbes ROC ainsi que de l'aire sous la courbe ROC (AUC) de chacun de nos modèles.

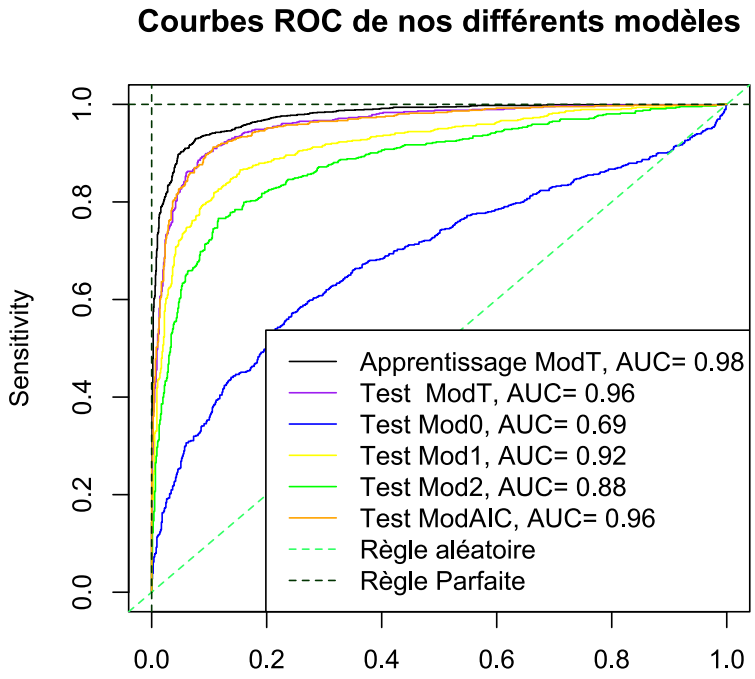

Figure 8: Courbes ROC de nos différents modèles

D'après notre figure nous constatons que les modèles les plus performants sont ModT et ModAIC. En effet, leur performance sur l'échantillon test est presque similaire tant il est difficile de différencier les deux courbes. On constate également que leur AUC arrondies aux dizièmes sont identiques. En outre, on observe logiquement que les performances de ModT sur l'échantillon d'apprentissage sont très proches de la règle parfaite.

Par ailleurs, on constate que les performances de Mod0 sont très faibles et se rapproche d'une règle de décision aléatoire. Cela est dû au faible nombre de variables prises par le modèle mais aussi la véracité de ces dernières comme étudié avant.

Enfin, Mod1 et Mod2 ont des performances correctes mais plus faibles que ModT et ModAIC. Toutefois, ils sont moins complexes que ModT par exemple car ils prennent moins de variables que ce dernier au vue de leur définition.

Pour départager définitivement nos modèles nous allons observer leur taux d'erreur sur les échantillons d'apprentissage et de test.

4 Performance des modèles : analyse des erreurs

Nous avons calculer les erreurs de nos modèles sur les échantillons de test et d'apprentissage à savoir le taux de mauvais classement, voici les résultats contenus dans un tableau.

Modèle	Apprentissage	Test
ModT	0.074	0.097
Mod0	0.355	0.344
Mod1	0.123	0.144
Mod2	0.166	0.178
ModAIC	0.079	0.101

Table 1: Performance des modèles de régression logistique

On retrouve donc les mêmes résultats globalement qu'avec les courbes ROC. Cependant, ici on est capable de distinguer qui est le plus performant entre ModT et ModAIC.

Ainsi, ModT est notre modèle le plus performant car il a un taux d'erreur sur l'échantillon test légèrement inférieur à celui de ModAIC.

Par ailleurs, on remarque assez logiquement que pour tous nos modèles excepté Mod0 l'erreur sur l'échantillon test est inférieure à l'erreur sur l'échantillon d'apprentissage. Cela renforce l'idée que les variables choisies pour Mod0 ne sont pas bonnes.

II Régression ridge

La régression ridge propose l'estimateur linéaire suivant: $\hat{\theta}_{ridge}(\lambda) := (X'X + \lambda Id_p)^{-1} X'Y$

Il s'agit en fait du régresseur linéaire classique, auquel on a ajouté dans le terme inverse λId_p. Le coefficient λ est appelé coefficient pénalité.

Plus celui-ci sera grand, plus le régresseur linéaire aura ses valeurs proches de 0. A l'inverse, un coefficient de pénalité proche de 0 nous renverra le régresseur classique.

C'est donc un paramètre qui pêrmet de régulariser le modèle, en donnant moins de poids aux données.

La régression ridge est plus efficace que la régression linéaire ou logistique simple dans le cas de la grande dimmension (plus d'individus que de variables). Elle peut aussi être très efficace lorsque les variables explicatives sont très corrélées, ce qui est généralement le cas lorsque le nombre de variables explicatives est suffisamment grand devant l'échantillon d'individus.

Dans notre étude, les données possèdent un nombre conséquent de variables explicatives. De plus, même en ayant retiré les variables de la partie 1, on observe que 108 couples variables ont un coefficient de corrélation supérieur à 75% dont 21 avec un coefficient de corrélation supérieur à 90%.

La régression ridge semble donc justifiée.

1 Paramètre de régularisation

Le calcul des estimateurs de $\hat{\theta}_{ridge}$ en fonction d'une grille de paramètres λ de 10^{-2} à 10^5 peut être effectué. On peut ensuite afficher l'évolution des coefficients de l'estimateur en fonction de plusieurs paramètres.

En affichant l'évolution des coefficients en fonction de la norme l_1 sur le graphique gauche de la figure 9 de $\hat{\theta}_{ridge}$, on remarque qu'il y a globalement deux coefficients qui portent beaucoup de poids, et peut-être même un troisème plus discret.

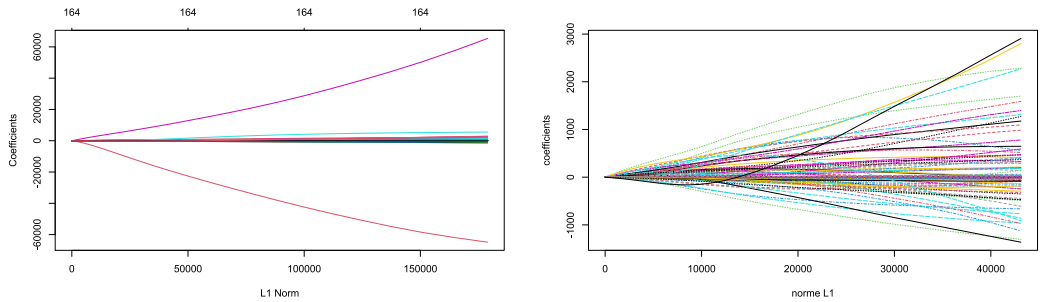

Figure 9: Evolution des coefficients en fonction de la norme l_1 de $\hat{\theta}_{ridge}$

Les variables associés aux coefficients divergents ont été identifiées :

- PAR_SFMV24 la 126ème variable du dataset.
- PAR_SFMV2 la 104ème variable du dataset.
- PAR_THR_3RMS_10FR_VAR la 176ème variable du dataset.

Le deuxième graphique de la figure 9, on peut retracer le graphique précédent:

On voit mieux le comportement des différents coefficients, qui semblent avoir pour la plupart des tendances linéaires.

On peut aussi tracer l'évolution des paramètres en fonction du coefficient de pénalité.

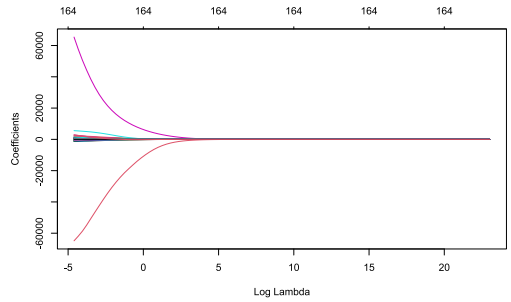

Figure 10: Evolution des coefficients en fonction de λ

2 Choix du paramètre optimal par validation croisée

On s'intéresse maintenant au choix du meilleur λ par validation croisée en 10-folds. Pour ce faire, l'algorithme calcul pour chaque λ l'erreur associée comme suit :

- Il découpe d'abord le jeu de données en 10-folds homogènes.
- Pour chaque fold, il entraîne son modèle sur les 9 folds restants.
- Il calcule la somme des carrés résiduels sur le fold restant.
- Il ajoute cette erreur à l'erreur du modèle et réitère sur le prochain fold.

L'algorithme choisit finalement le λ pour lequel l'erreur calculée est la plus faible.

L'avantage de la validation croisée est qu'elle permet de prendre en compte l'ensemble des données que l'on possède dans le choix du modèle.

Elle permet ainsi de réduire la variance du modèle choisi.

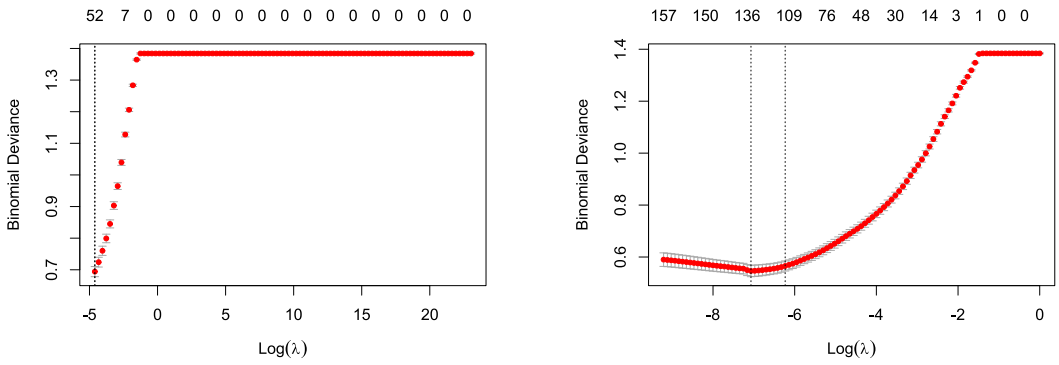

Figure 11: MSE en fonction de λ

On observe sur le premier graphe de la figure 11 que le minimum est atteint sur la frontière, pour $\lambda = 10^{-2}$. On décide alors de changer le domaine de recherche du minimum, qui semble être atteint avant 10^0.

Cette fois-ci, on trouve une valeur de $\lambda_{min} = 0,0008$ qui n'est pas sur la frontière du domaine de recherche mais qui est 1 ordre de grandeur plus petit que celui trouvé précédemment. Le minimum est bien visible sur le second graphe de la figure 11.

3 Performances

Toute la procédure éffectuée précédemment a également été effectuée sur le dataset complet, contenant toutes les variables.

On peut alors comparer les résultats de performance de la régression ridge sur les deux cas.

Dataset	Erreur d'apprentissage	Erreur de généralisation
Dataset réduit	0.08570425	0.09713487
Dataset complet	0.08746048	0.09573725

Table 2: Performances de la régression Ridge sur les deux datasets

On observe que bien que le dataset complet possède plus de variables, ses erreurs d'apprentissage et de généralisation sont équivalentes à celles du dataset réduit.

Les différences marginales que l'on peut constater peuvent tout aussi bien être dues au nouveau choix de germe du générateur de nombres aléatoires, qui induit le choix de folds différents et donc des changements potentiels, même si la validation croisée réduit grandement la variance des modèles.

Finalement, malgré un modèle plus complexe, ses performances sont les mêmes.

III Classification par K-plus-proches voisins

1 $K = 1$

1.1 Performance

Dataset	Erreur d'apprentissage	Erreur de généralisation
Dataset réduit	0	0.0580014
Dataset complet	0	0.05101328

Table 3: Performances KNN, K=1

Le modèle des K-plus-proches voisins avec K=1 ne fournit naturellement pas d'erreur d'apprentissage. Pour ce qui est de l'erreur de généralisation, on remarque que le modèle posédant toutes les variables est plus efficace.

2 K optimal

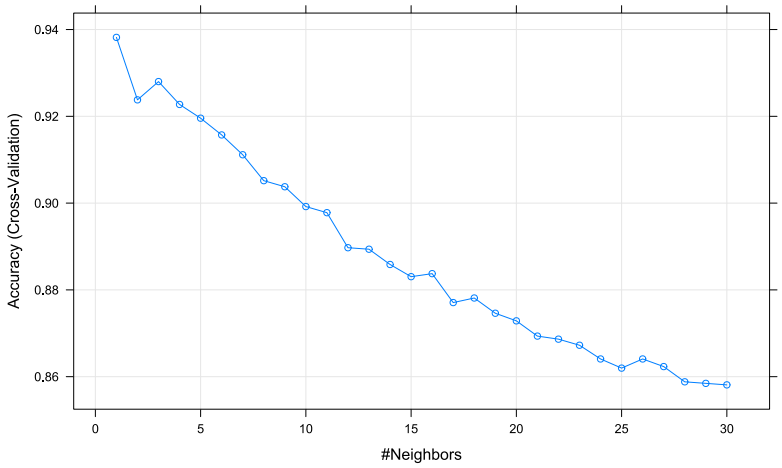

Figure 12: Précision du modèle KNN en fonction de K par validation croisée

La validation croisée nous fournit un K optimal de 1 !

3 Algorithme K plus proches voisins et le fléau de la dimension

Au vu du résultat précédent, on peut se demander si l'algorithme des K-plus proches voisins est pertinent au vue de notre jeu de donnée. La lecture de Cours 2 : Université de Cornell nous démontre que plus un jeux de donnée possèdent de variables explicatives, plus il faudrait pour obtenir de bons résultats augmenter considérablement la taille de notre jeu de donnée d'apprentissage. Ici, nous possédons un nombre de variables explicatives très élevées et notre échantillon d'apprentissage est quand à lui de taille plutôt modérée.

Par ailleurs, le fait que le k optimal soit 1 tend à nous faire penser que l'on se retrouve dans une situation de sur-apprentissage et que ce modèle performerait peut être moins bien sur un jeu de donnée de prédiction.

C'est pourquoi, nous pensons que l'algorithme des K-plus proches voisins n'est pas adapté à ce jeu de donnée.

IV Choix du modèle

Nous avons vu dans la section précédente pourquoi le modèle des K-plus proches voisins n'était pas le modèle idéal dans notre contexte. Il faut donc choisir entre l'un des modèles de régression logistique ou de régression ridge.

Pour ce qui est des modèles de régression logistique, si l'objectif du modèle est uniquement le résultat de la prédiction, alors il sera plus judicieux de choisir le modèle ModT, qui possède le plus de variables et dont l'erreur de généralisation est minimale.

A l'inverse si ce qui est important est de comprendre la structure des données, et de comprendre quels sont les facteurs qui portent le plus de poids dans la classification d'un titre selon l'une des deux classes, alors il sera plus judicieux de choisir le modèle ModAIC.

Du côté de la régression ridge, nous avons vu que l'ajout ou non des variables "nuisibles" ne changeait pas la performance du modèle. Il sera donc judicieux de travailler avec le modèle plus simple dans lequel nous avons enlevé les différents variables identifiées.

Pour conclure, dans le cas de la classification dûre de nouveaux individus en un genre qui leur correspond, nous préconisons le choix du modèle de régression ridge obtenu sur le dataset tronqué des variables identifiées, car ce dernier possède des erreurs d'apprentissage et de généralisation très proches de celles du modèle modT, tout en ayant une variance de modèle faible, puisqu'il a été calculé par validation croisée.

Il sera donc potentiellement plus efficace sur des nouveaux jeux de données.

Conclusion

Nous avons pu voir dans notre étude plusieurs modèles pour établir une classification d'un morceau de musique selon son genre (Jazz ou Classique). L'établissement de ces modèles est d'abord passée par une analyse descriptive des données, qui a permi de réduire le nombre de variables utiles à la régression.

Parmi les modèles que nous avons obtenus, certains étaient adaptés à notre situation, tout en montrant de bonnes performances. Les excellentes performances du modèles du plus proche voisin nous laisse tout de même perplexes. Il est tout à fait possible que le jeu de données qui nous a été fournit ne soit pas représentatif de l'ensemble des différents morceaux Classique ou Jazz, à tel point que ceci puisse expliquer les performances excellentes de ce modèle.

Etant donné que ce modèle est réputé pour suraprendre, il tout aussi possible que cette méthode fonctionne mal pour d'autres jeux de test. C'est pourquoi nous nous sommes plutôt orientés vers les régression logistique et ridge. Nous avons enfin choisi ce dernier modèle, étant donné que la validation croisée réduit sa variance.