DM - Statistique Bayesienne

Jeremy Bron

Mai 2019

- 0 - Description et visualisation des données
  - 0.1 - Structure des données et des couples Matières / Établissements
  - 0.2 - Distibution des variables
- 1 - Régression linéaire
  - 1.0 - modèle linéaire gaussien
  - 1.1 Régression linéaire bayésienne
    - 1.1.1 Approche par prior de Zelnner et échantillonnage de Gibbs
    - 1.1.2 - Bayesian avec le package BAS
  - 1.2 Choix des covariables significatives et compare à une analyse fréquentiste
    - 1.2.1 Choix des covariables significatives
    - 1.2.2 - Comparaison avec le modèle fréquentiste
  - 1.3 Analyse des mutations en mathématiques et en anglais
    - 1.3.1 - Bayésien sur toutes les variables
    - 1.3.2 - Bayésien sur les variables sélectionnées précédemment
- 2 - Loi de Pareto
  - 2.4 - Réalisation d'une loi de Pareto et impact du paramètre α
  - 2.5 & 2.6 - Choix de loi à priori et calcul des posteriors
    - Inférence sur la borne supérieure de Barre
    - Inférence sur le paramètre m
    - Inférence sur le paramètre α
  - 2.7 - Tirage de la loi à postériori
  - 2.8 - Tirages sur les données des matières Mathématiques et Anglais uniquement

Liste des packages utilisés:

```
setwd("C:/Users/jb/Google Drive/BIG DATA DAUPHINE/Bayésien")
mut = read.csv("mutations2.csv")
require(zoo)
library(caret)
library(tidyverse)
library(yarrr)
require(stats)
library(knitr)
require(ggpubr)
library(BAS)
library(monomvn)
library(glmnet)
library(reshape2)
library(VGAM)
library(ggdistribute)
library(HDInterval)
library(gridExtra)
library(goftest)
library(RVAideMemoire)

Load_data = T #mettre FALSE pour recalculer les procédures, sinon les résultats seront chargé
s
```

0 - Description et visualisation des données

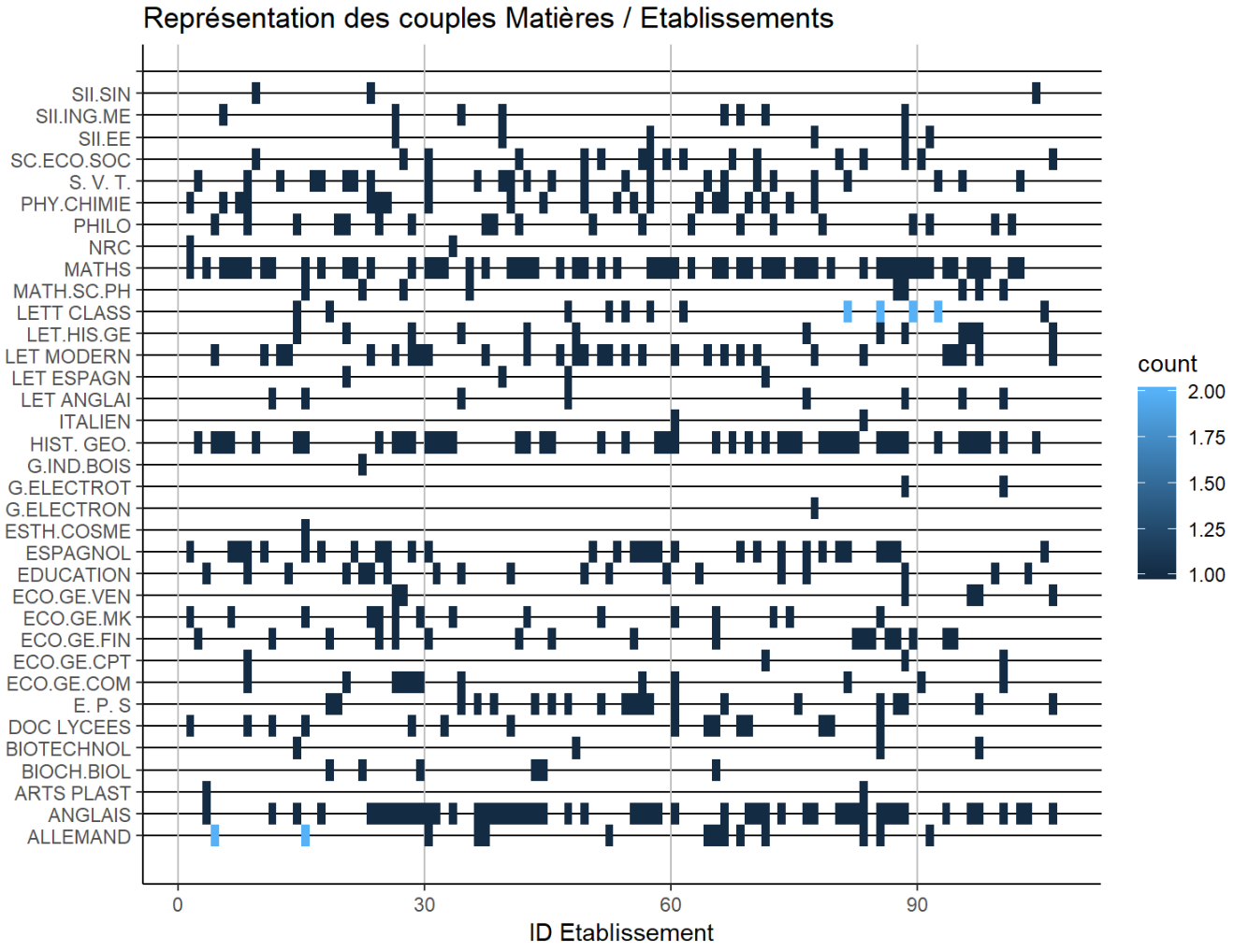
0.1 - Structure des données et des couples Matières / Établissements

Le fichier `mutation2.csv` contient 23 variables et 516 individus. Les 5 premières variables sont qualitatives et renseignent l'établissement, les matières concernées et d'autres paramètres identifiants les établissements (ville, code, etc.). À noter que la variable `établissement` ne renseigne pas les établissements de façon unique, car plusieurs noms sont similaires pour des lycées de différentes communes. Dans la suite de l'étude, la variable `code_etablissement` sera utilisée comme référence.

Les 18 variables suivantes sont numériques en renseignent les caractéristiques de chaque couple (établissement, matière), notamment les effectifs, taux de réussite et d'accès aux fillaires et la variable réponse `Barre` qui indique le nombre de points nécessaire pour une mutation dans un couple donné.

Notre analyse porte sur 35 matières réparties dans 107 établissements. Ainsi les données ne contiennent pas toutes les matières pour chaque établissement.

Le graphique ci-dessous illustre la répartition des couples Matières / Établissements :



On constate les éléments suivants qui seront ensuite approfondis par la suite:

- Tous d'abord la présence de doublons dans les données est mise en évidence, car chaque couple devrait être unique, ce qui n'est pas le cas pour les matières `ALLEMAND` et `LETT CLASS` .
- Les différentes matières sont représentées de façon assez déséquilibrée: certaine une seule fois, d'autres présents pour plus de la moitié des établissements

- Globalement les données qui auraient pu être exhaustives dans tous les lycées pour les matières du tronc commun sont très éparses. La statistique Bayésienne bien adaptée à ce type de données.

Traitement des doublons

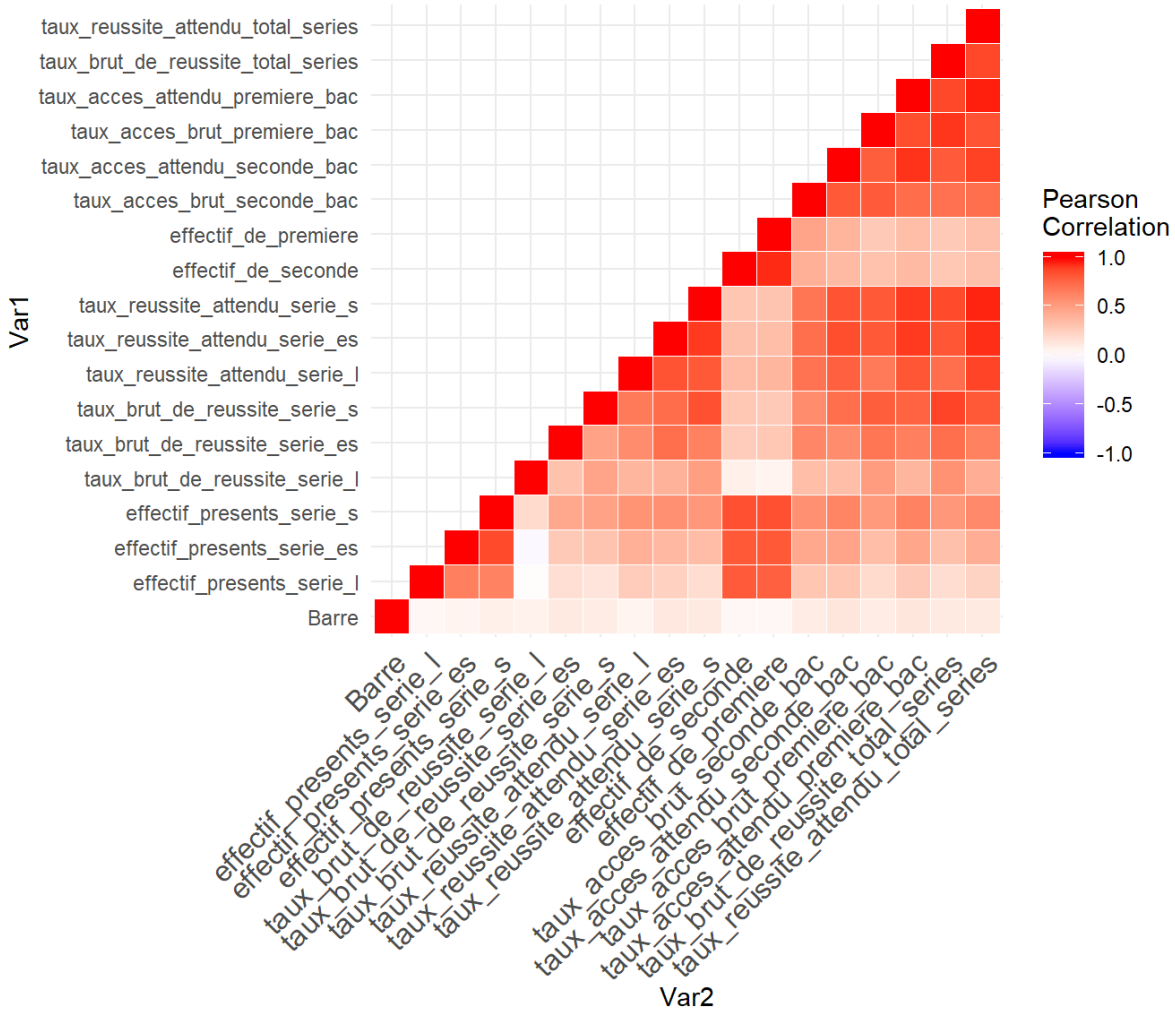
6 doublons ont été trouvés dans le dataset :

	code_etablissement	ville	etablissement	commune	Matiere	Barre
14	0781512V	MONTIGNY LE BRETONNEUX	LYCEE DESCARTES	78423	ALLEMAND	2019.2
16	0781512V	MONTIGNY LE BRETONNEUX	LYCEE DESCARTES	78423	ALLEMAND	2019.2
60	0781951X	MAGNANVILLE	LYCEE LEOPOLD SEDAR SENGHOR (GENERAL ET TECHNO.)	78354	ALLEMAND	338.2
61	0781951X	MAGNANVILLE	LYCEE LEOPOLD SEDAR SENGHOR (GENERAL ET TECHNO.)	78354	ALLEMAND	338.2
393	0950646L	GONESSE	LYCEE RENE CASSIN	95277	LETT CLASS	130.0
395	0950646L	GONESSE	LYCEE RENE CASSIN	95277	LETT CLASS	130.0
409	0950650R	SARCELLES	LYCEE JEAN-JACQUES ROUSSEAU (GENERAL ET TECHNO.)	95585	LETT CLASS	241.0
413	0950650R	SARCELLES	LYCEE JEAN-JACQUES ROUSSEAU (GENERAL ET TECHNO.)	95585	LETT CLASS	241.0
446	0951147F	L ISLE ADAM	LYCEE FRAGONARD	95313	LETT CLASS	224.0
450	0951147F	L ISLE ADAM	LYCEE FRAGONARD	95313	LETT CLASS	224.0
459	0951710T	VAUREAL	LYCEE CAMILLE CLAUDEL (GENERAL ET TECHNO.)	95637	LETT CLASS	440.0
461	0951710T	VAUREAL	LYCEE CAMILLE CLAUDEL (GENERAL ET TECHNO.)	95637	LETT CLASS	440.0

Pour la suite de l'étude, nous avons supprimé les doublons, car ils auraient apporté une pondération erronée aux couples concernés. Notre analyse porte donc sur 510 couples matières / établissement.

0.2 - Distibution des variables

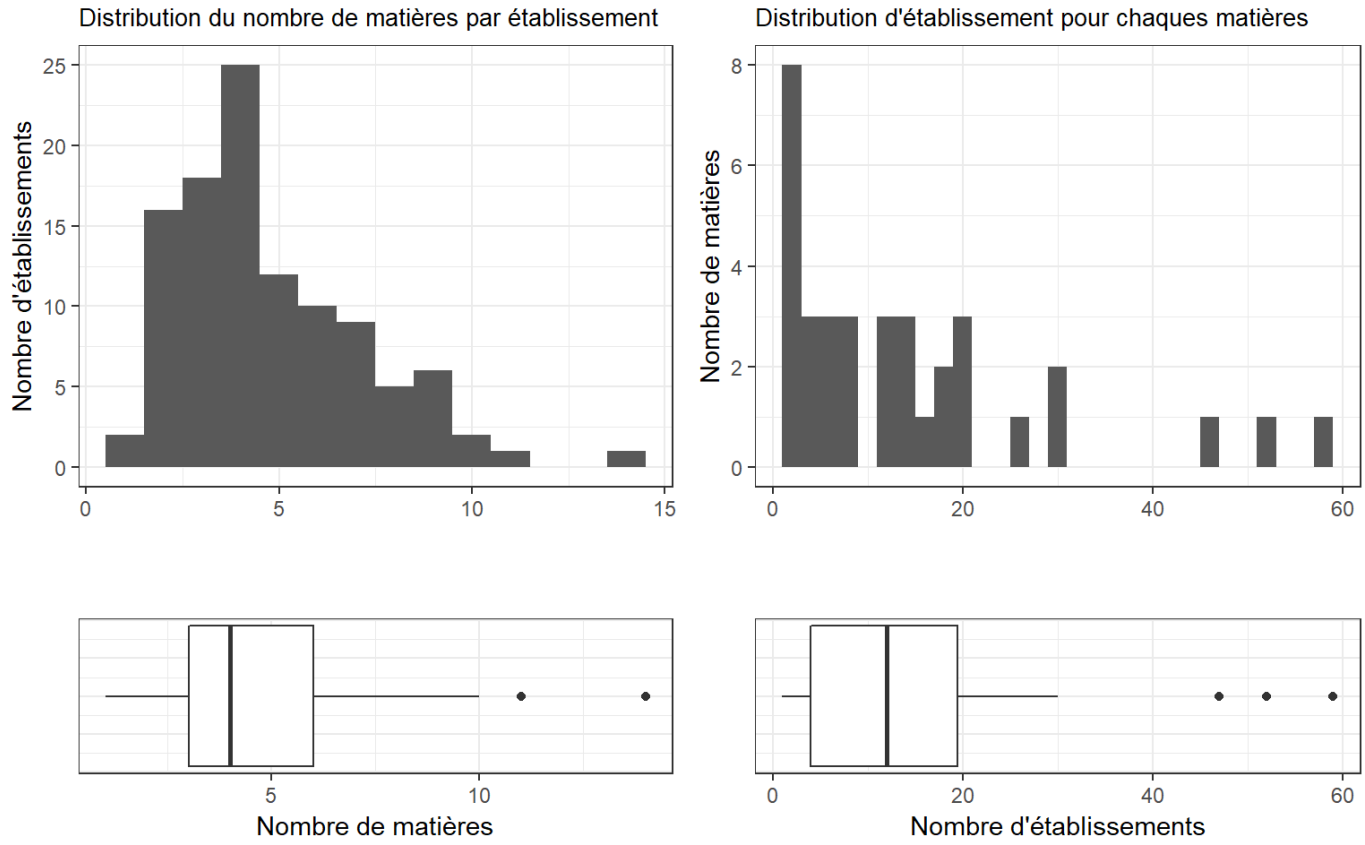
Tous d'abord la corrélation entre les variables est illustrée dans la figure suivante:



Les variables sont globalement toutes très corrélées entre elles, en effet:

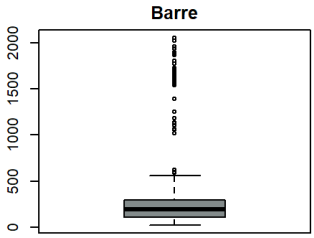
- Les effectifs des classes supérieures des établissements sont naturellement influencés par les effectifs des classes inférieurs
- Les taux de réussites dans une section sont représentatifs du niveau général d'un établissement et forcement assez semblable aux taux de réussite des autres sections

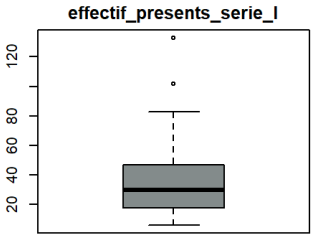
La figure ci-après illustre les distributions des données du nombre de matière par établissement et vice versa.

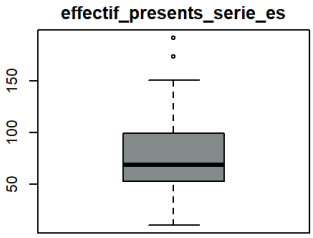


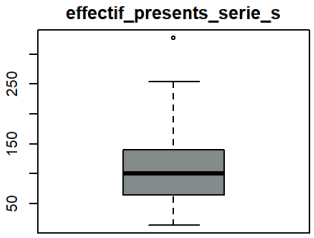
Ainsi on comprend que:

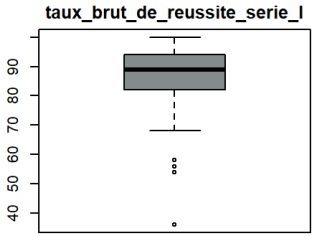
- En moyennes les données renseignent 4 matières par établissement, 75% des établissements ont entre 3 et 6 matières et 2 établissements avec 10 ou plus matières
- En moyennes les données matières sont présentes pour 12 établissement différents, 75% des matières sont présente dans 4 à 20 établissements et quelques matières comme l'anglais, les mathématique et l'histoire géographie sont disponible pour 40 à 60 établissements.

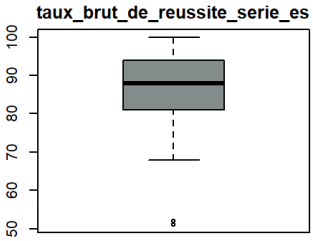

Barre
0
500
1000
1500
2000

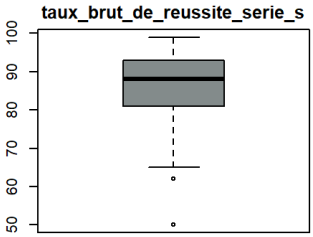

effectif_presents_serie_l
20
40
60
80
120

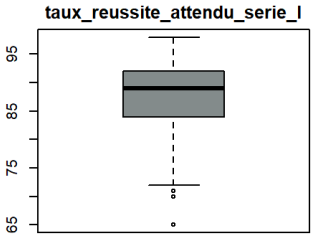

effectif_presents_serie_es
50
100
150

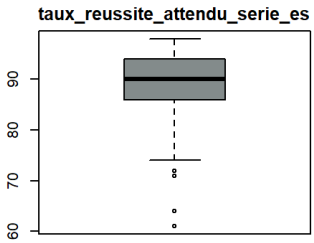

effectif_presents_serie_s
50
150
250

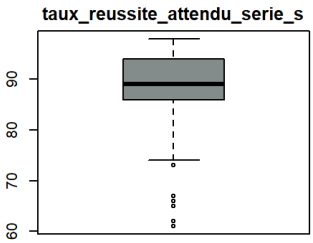

taux_brut_de_reussite_serie_l
40
50
60
70
80
90

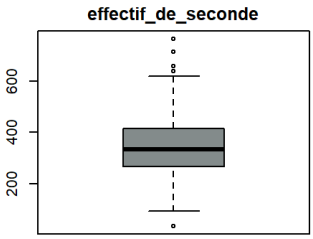

taux_brut_de_reussite_serie_es
50
60
70
80
90
100

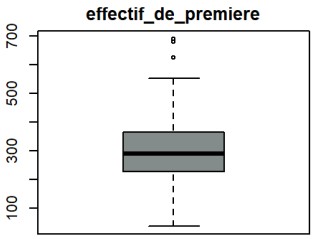

taux_brut_de_reussite_serie_s
50
60
70
80
90
100

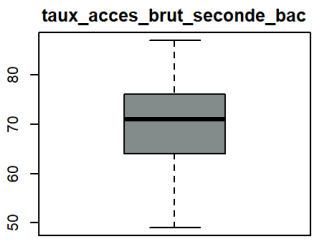

taux_reussite_attendu_serie_l
65
75
85
95

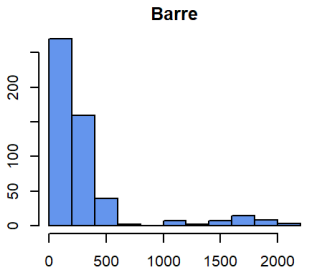

taux_reussite_attendu_serie_es
60
70
80
90

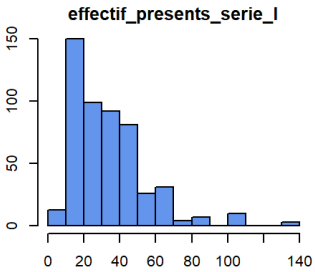

taux_reussite_attendu_serie_s
60
70
80
90

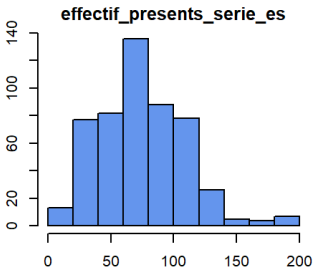

effectif_de_seconde
200
400
600

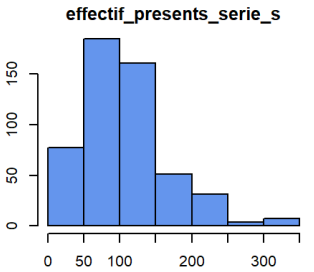

effectif_de_premiere
100
300
500
700

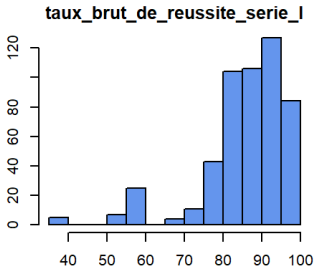

taux_acces_brut_seconde_bac
50
60
70
80

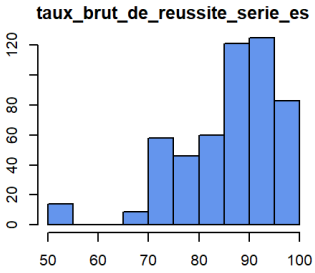

Barre
0
50
100
200
0
500
1000
1500
2000

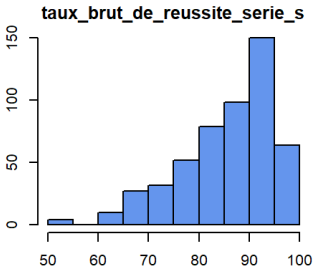

effectif_presents_serie_l
0
50
100
150
0
20
40
60
80
100
140

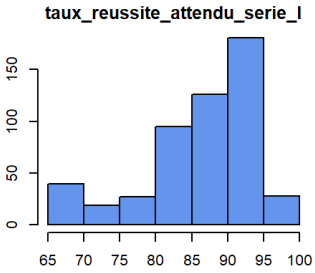

effectif_presents_serie_es
0
20
60
100
140
0
50
100
150
200

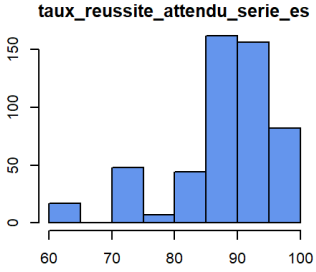

effectif_presents_serie_s
0
50
100
150
0
50
100
200
300

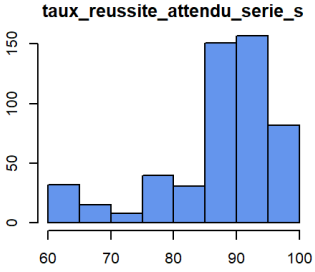

taux_brut_de_reussite_serie_l
0
20
40
60
80
120
40
50
60
70
80
90
100

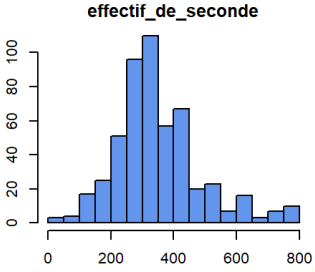

taux_brut_de_reussite_serie_es
0
20
40
60
80
120
50
60
70
80
90
100

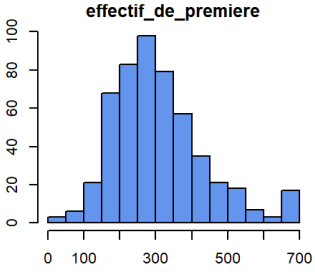

taux_brut_de_reussite_serie_s
0
50
100
150
50
60
70
80
90
100

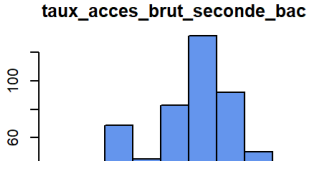

taux_reussite_attendu_serie_l
0
50
100
150
65
70
75
80
85
90
95
100

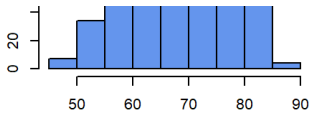

taux_reussite_attendu_serie_es
0
50
100
150
60
70
80
90
100


taux_reussite_attendu_serie_s
0
50
100
150
60
70
80
90
100


effectif_de_seconde
0
20
40
60
80
100
0
200
400
600
800

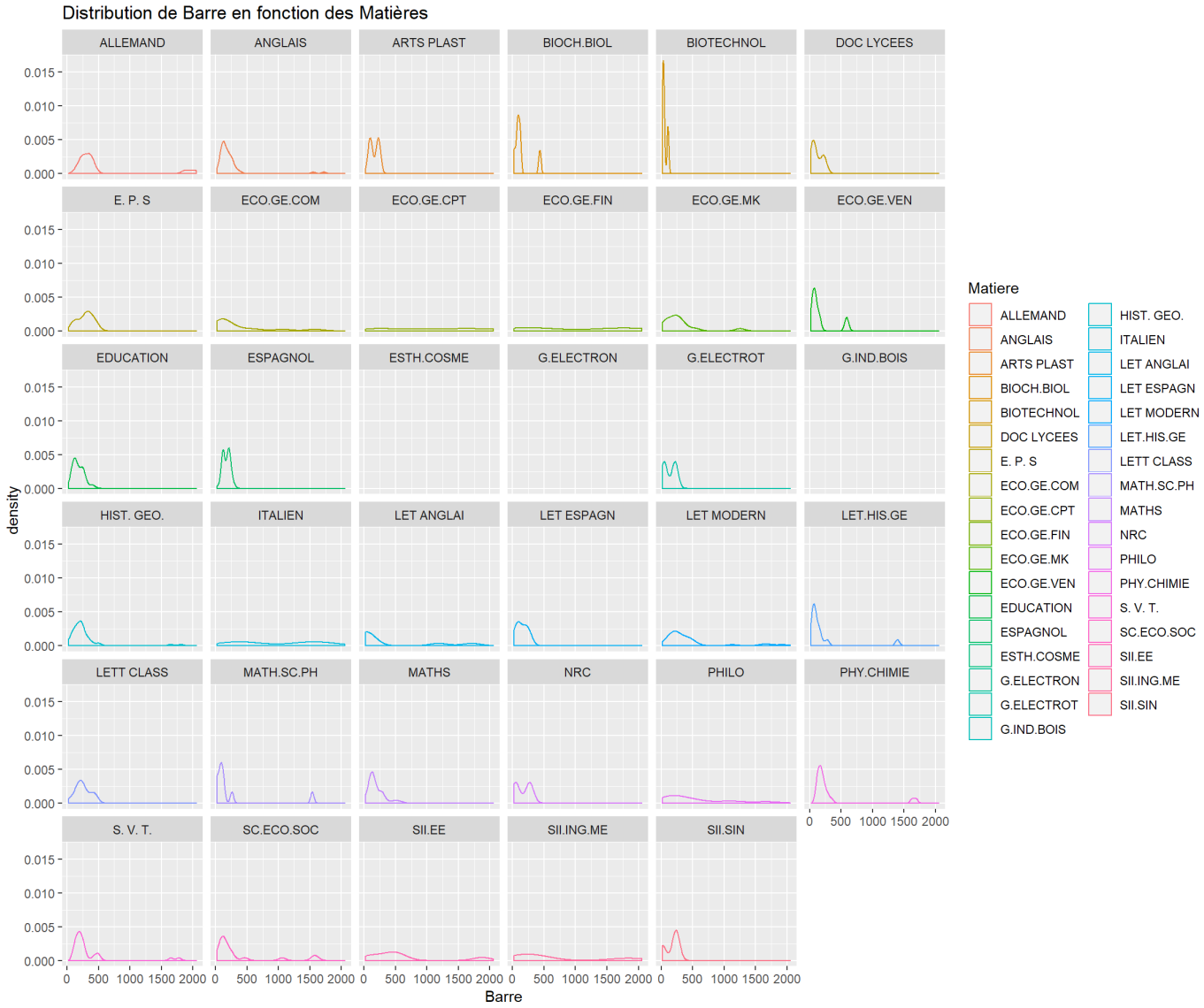

effectif_de_premiere
0
20
40
60
80
100
0
100
300
500
700


taux_acces_brut_seconde_bac
60
100



Les distributions des variables caractérisant les établissements (sauf `Barre`) sont globalement centrées autour d'une valeur moyenne, les variables représentant des taux de réussites (donc max 100) sont faussées vers la gauche. On constate également la présence d'individus sur les queues de distributions.

La variable `Barre` est analysée avec plus de détails dans les figures suivantes. Les Matieres présentes dans 1 seul établissement ou établissement avec une seule matière ne sont pas représentées.

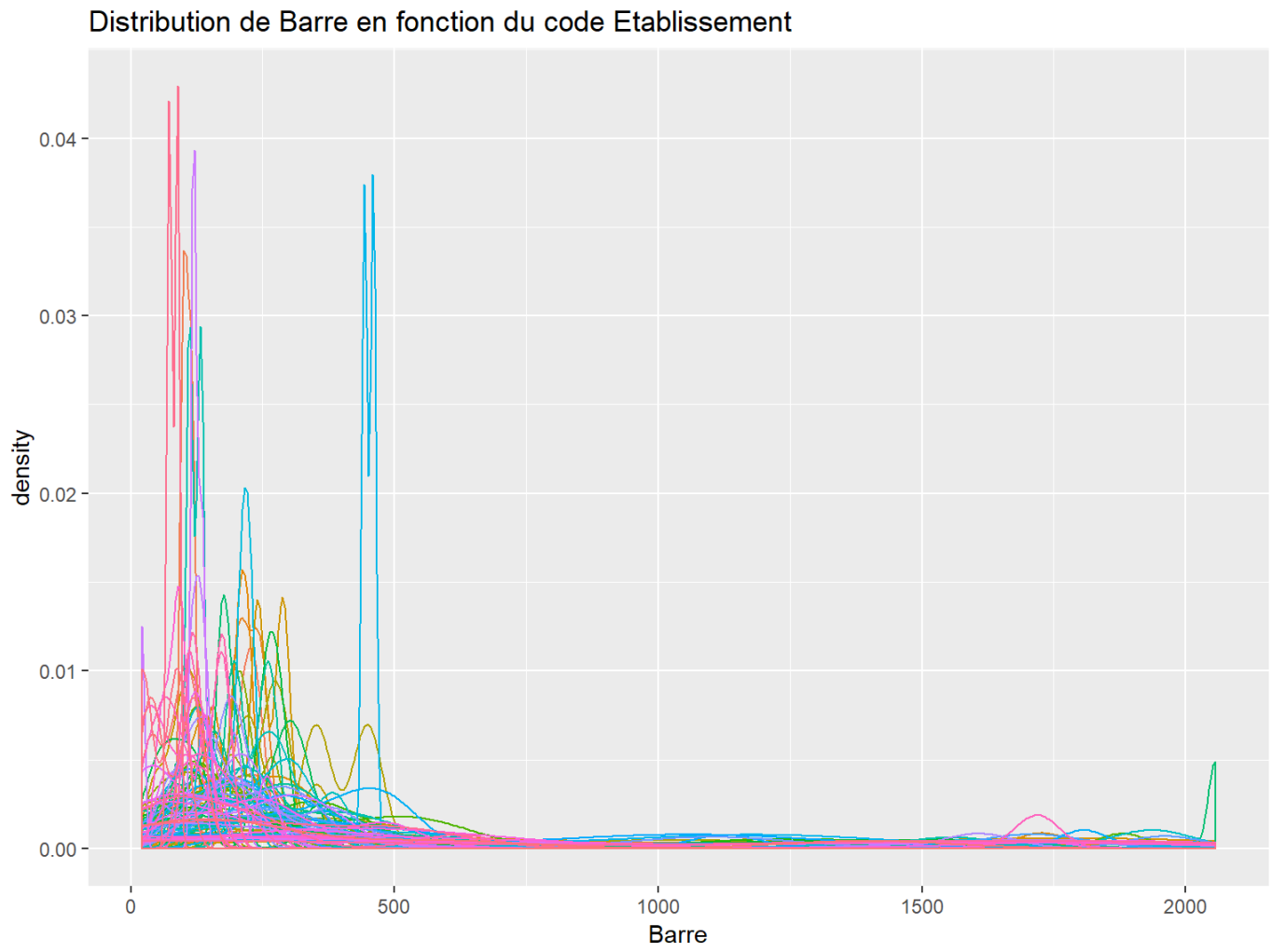


On retrouve visuellement une distribution qui ressemble à l'histogramme de `Barre` (pic entre 0 et 500 et quelques individus après 1500) avec cette fois plus d'information sur l'influence des `Matiere`.

Globalement la `Barre` par `Matiere` se situe entre 0 et 500, pour certaines matières le pic est plus étroit et proche de 0 (eg: biotechnol), pour d'autres on constate la présence d'un 2e pic pour des valeurs hautes de `Barre` (eg: phy.chimie, eco.ge.ven, etc.) .

Certaines `Matiere` (philo, italien) ont des distributions de `Barre` relativement plate.

Ainsi on comprend que la matière en elle-même à une certaine influence sur les barres de mutations nécéssaire mais qu'il y a un autre facteur (distribution à plusieurs pics ou aucuns pic), l'établissement, responsable des différences de `Barre`



Nous n'avons pas séparé les établissements, car certains ont très peu de matières et donc cela ne permet pas de conclure sur l'influence de la matière ou de l'établissement. Globalement les nombreux pics par établissement (représentés avec des couleurs différentes) renseignent d'une certaine influence des établissements sur la `Barre` .

1 - Régression linéaire

1.0 - modèle linéaire gaussien

Avant d'effectuer la régression linéaire bayésienne, nous effectuons une régression linéaire standard pour expliquer la variable `Barre` :

- La régression est faite sur toutes les variables numériques
- Les variables `Matiere` et `code_etablissement` sont testés dans la régression, les autres variables sont redondantes ou n'apporte pas d'informations à ce stade.

Cette première approche nous permet de voir que:

- Les modèles incluant `code_etablissement` ne convergent pas pour les variables numériques. Ce qui est normal, car toutes les variables numériques renseignent de paramètres d'effectif et de réussite propre à chaque établissement et donc complètement corrélés (toutes égales pour chaque établissement)
- Cependant certaines classes de `code_etablissement` (certains établissements) ont des p-valeur significatifs:
    - Lorsque la régression est effectuée seulement sur `code_etablissement` : environ 10% des établissements ont une influence significative sur la variable réponse et certaine avec une p-

valeur inférieur à 0.001.
    - Lorsque la régression est effectuée sur `code_etablissement` et `Matiere` les p-valeur significatifs des établissements sont globalement moins nombreux et moins élevés
- Le modèle avec la variable `Matiere` converge sur les variables numériques, en effet les matières sont présentes une fois par établissement et dans plusieurs établissements et donc cela résout les problèmes de corrélation.
    - Le modèle avec seulement la variable `Matiere` faut ressortir que plusieurs matières ont un impact significatif sur la variable réponse. L'intercepté est également très significative
- Dans le modèle avec les variables numériques en plus, le type de matière est moins un peu souvent significatif, mais reste globalement plutôt similaire. Le modèle avec seulement les variables numérique cherche donc à expliquer la variable réponse uniquement avec les caractéristiques des établissements. Seulement 4 variables sont un peu significatives (p-valeur entre 5 et 10%).
    - On comprend bien qu'étant donnée l'influence de certaines matières sur la variable réponse, une régression seulement sur les caractéristiques numériques des établissements ne permet pas d'obtenir de modèle satisfaisant. La variabilité de `Barre` est trop importante entre les matières.
    - Cela se vérifie dans les scores Adjusted R-squared proche pour le modèle avec `Matiere` seulement (0.1477) et le modèle avec `Matiere` et variables numériques (0.1574); le modèle avec variable numérique seulement est bien plus mauvais (0.004397).

Pour la suite on commencera donc à l'intéresser à la distribution de l'estimateur de la variable `Barre` grâce à la statistique Bayésienne. Dans un premier temps de façon générale, seulement grâce aux variables numériques afin d'obtenir plus d'information sur la dispersion de la variable lorsque l'ont le considère pas de matières ou établissement en particulier. Ensuite sur certaines matières en particulier afin de mieux comprendre les subtilités et l'influence des covaraibles numérique.

1.1 Régression linéaire bayésienne

1.1.1 Approche par prior de Zelnner et échantillonnage de Gibbs

On souhaite réaliser la régression linéaire dans le cadre bayésien, pour cela une première approche peu être effectué selon la méthodologie suivante:

choix de la prior

- Les données à notre disposition et notre connaissance de la variable barre ne nous permettent pas d'inférer sur la "vrai" loi suivie pour calculer l'attractivité. À nous donc de faire un choix éclairé de la prior.
- Il n'y a donc pas de raison spécifique de choisir une loi conjuguée particulière pour notre prior.

Ainsi nous faisons le choix ici d'une prior non informative pour limiter l'impact de l'inférence initiale sur la prior. La prior de Zelnner est bien adapté à cette situation et pour la régression linéaire.

Description de l'algorithme de sélection de variables MCMC avec échantillonnage de Gibbs

- On tire au sort un vecteur binaire γ pour décider quelles covariables inclure dans le modèle
- Pour chaque composante γ_i on calcule la vraisemblance du modèle (avec ou sans la covariable i)
    - La fonction `marglkd1` est utilisée calcul la log-vraisemblance marginale pour deux modèles (l'intercepte est toujours gardé dans le modèle)
    - On tire au sort la nouvelle composante composantes γ_j selon une loi de Bernouilli donc la probabilité corresponds au ratio des vraisemblances marginales des modèles testés
- Le nouveau modèle et sa vraisemblance sont calculés et sauvegardés
- L'opération est répétée assez de fois pour que l'hypothèse de stationnarité de la chaine de Markov nous permette d'atteindre la convergence.

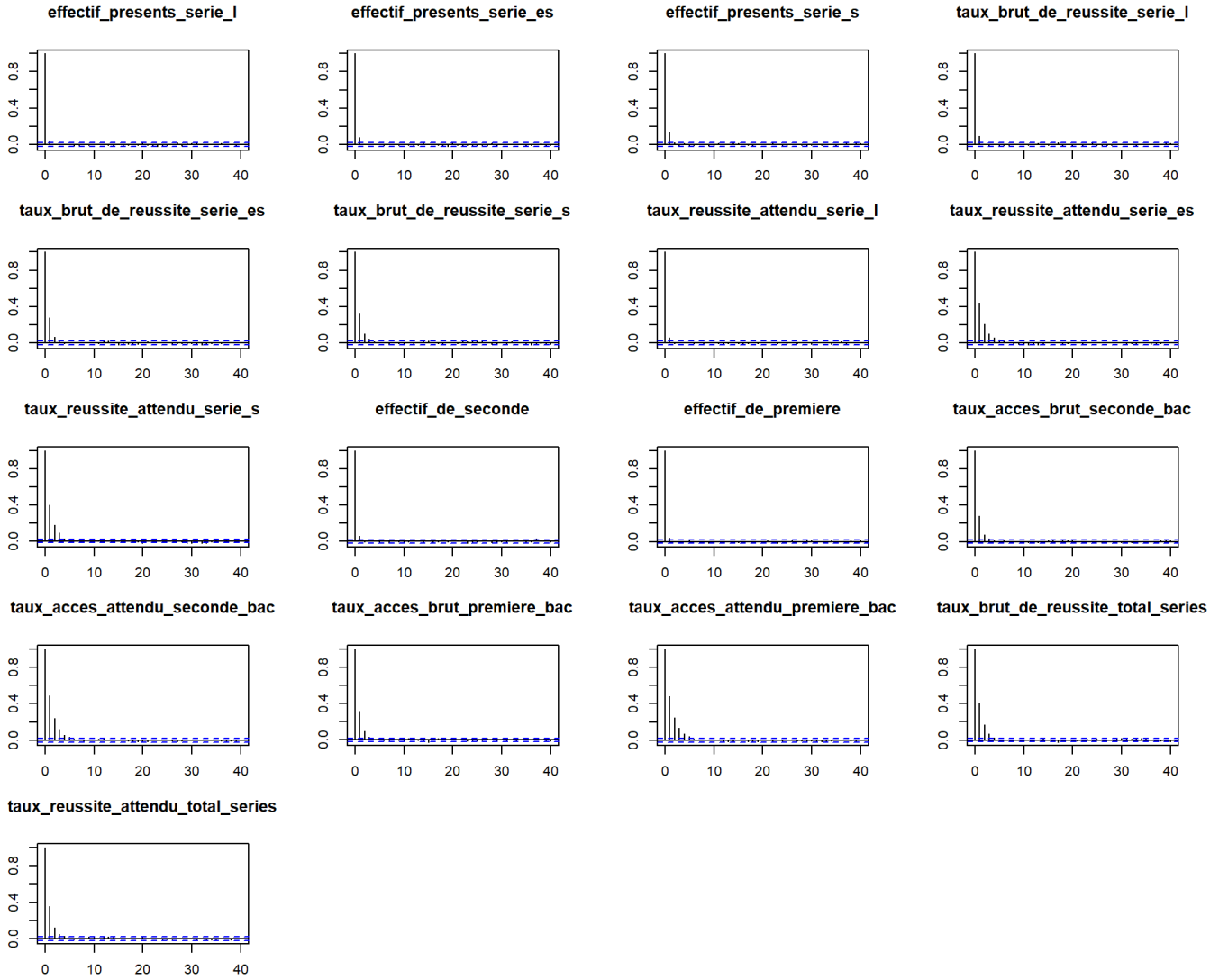
- le choix du paramètre g est fixé à $g = n = 510$ (on donne le même poids à la loi a priori qu'à une observation)

L'algorithme a été testé plusieurs fois avec 10000 itérations et une fois avec 100000 itérations, les résultats sont stables et comparables. À noter que nous avons 2^17 = 131072 combinaisons différentes pour le vecteur γ.

Verification de la convergence

Autocorrelation de la chaine des modèles

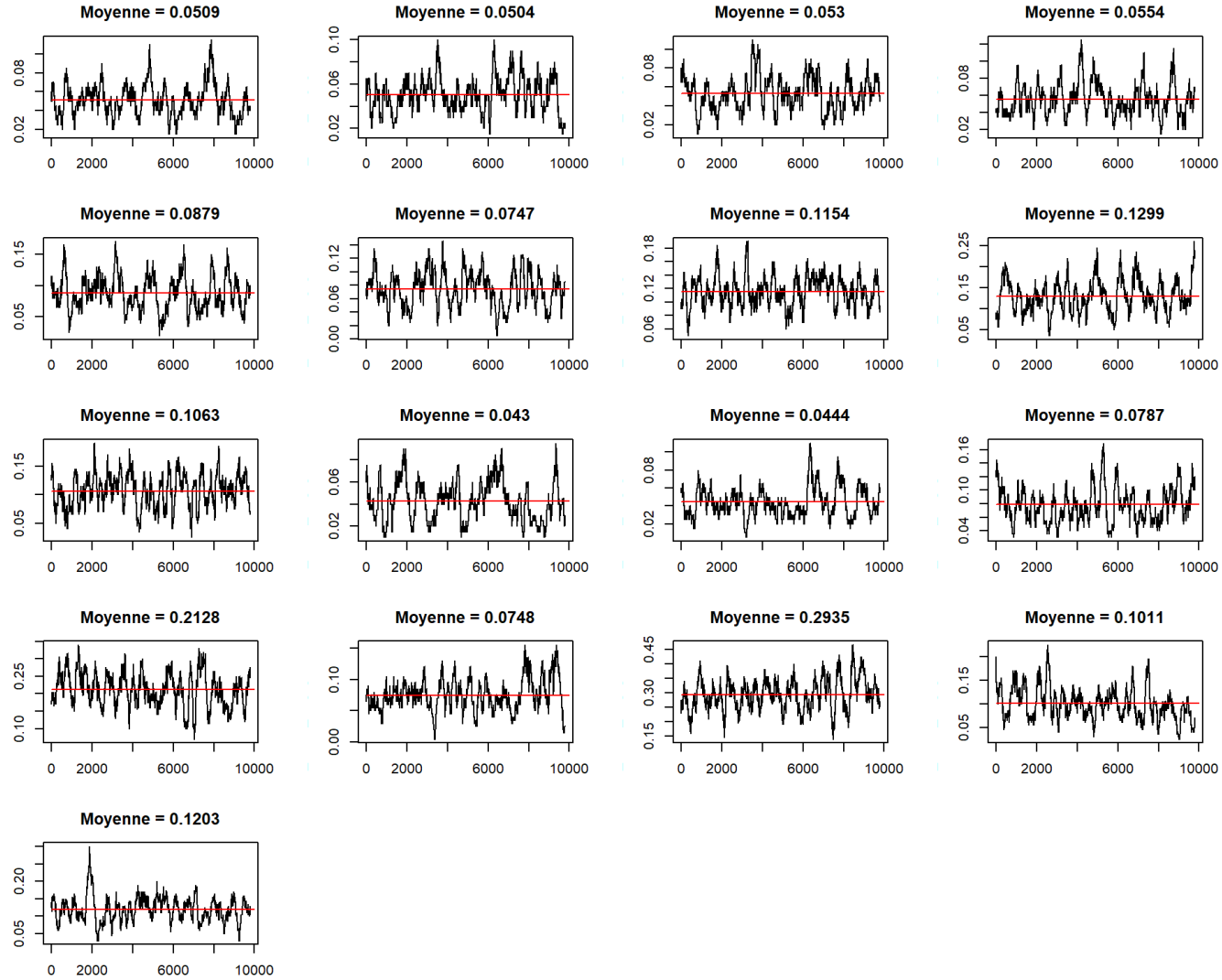
On vérifie tout d'abord la qualité du mélange de la chaine de Markov grâce l'autocorrélation de l'exploitation des modèles de γ.



L'autocorrélation décroit presque instantanément ou très rapidement pour toutes les variables. La chaine de Markov ne met pas beaucoup d'itération pour explorer les lois sur les covariables.

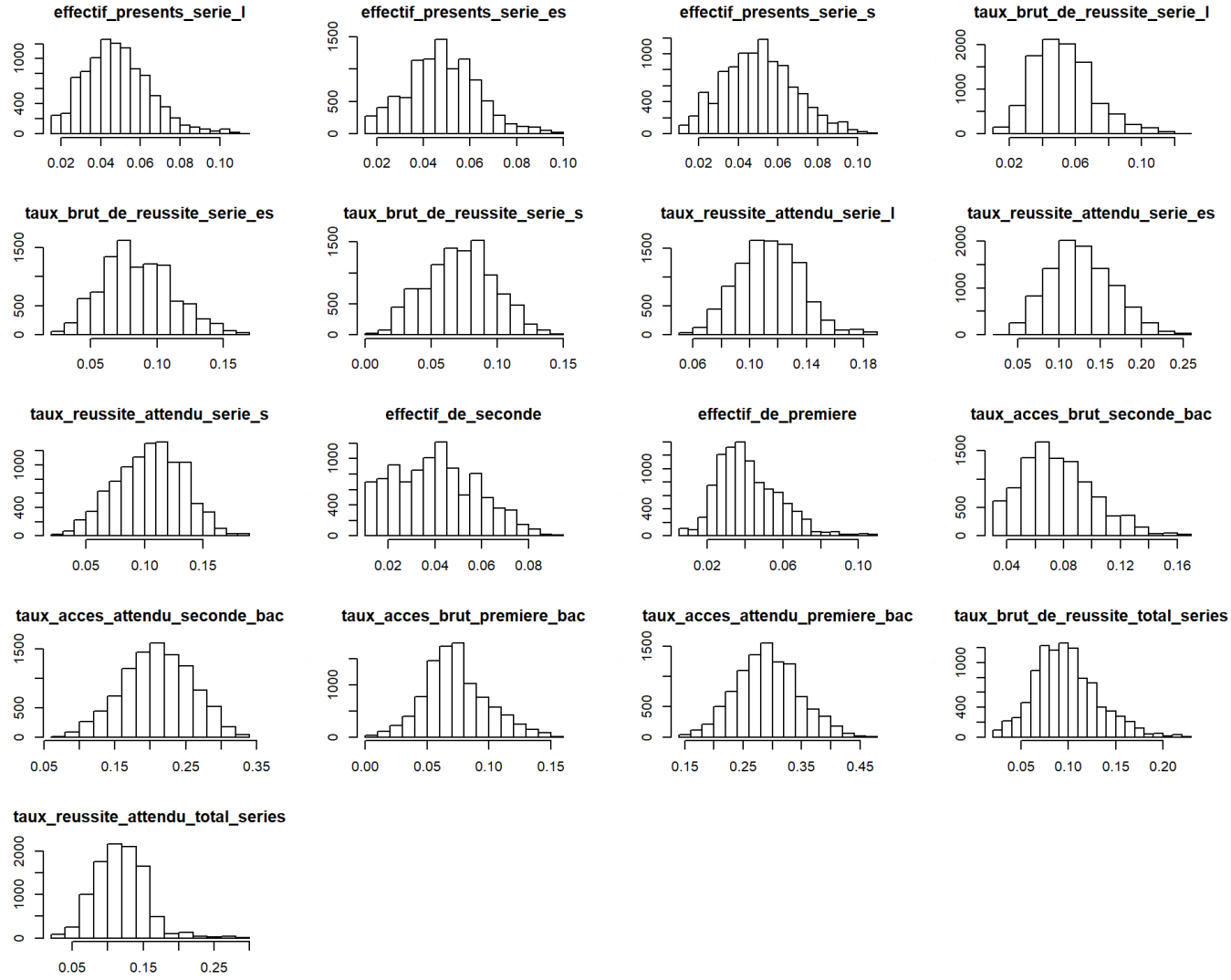
Valeurs et distribution des γ_i

On peux également afficher la moyenne mobile (ici par blocs de 200) de la valeur des coefficients binaire de γ. C'est à dire le pourcentage moyen de fois que la chaine a sélectionné ces variables:



Pour chaque variable le pourcentage de fois qu'elle a été incluse dans le modèle est représenté en rouge. On constate l'algorithme fonctionne bien et rapidement les différentes variables sont sélectionnées de façon relativement stable autour d'une valeur moyenne.

On peut également représenter la moyenne mobile des $ _{j} $ par leurs histogrammes.



On voit bien que certaines covariables sont plus vraisemblables que d'autres pour expliquer la variable réponse et qu'il existe également une certaine incertitude sur les modèles même si la chaine converge. Certaines distributions sont proches de 0 et permet de penser que le β_i de ces covariables est probablement nul.

Récurences des modèles

L'algorithme MCMC va revenir de nombreuses fois sur les modèles les plus probables. On s'interroge ici sur le nombre de modèles différents explorés en fonction du nombre d'itérations. On enlève systématiquement les 2000 premières interactions de burnin :

- En repérant 10 l'algorithme avec 10000 itérations le nombre de modèles différents après burnin est compris entre 800 et 860
- Sur l'agrégation des 10 précédents : environ 2400 modèles différents
- En effectuant 100000 itérations environ 2600 modèles différents
- Sur l'agrégation des 10 fois 10000 itérations et une fois 100000 itérations: environ 3400 modèles différents.

Le nombre de modèles exploités par l'algorithme croit beaucoup plus lentement que le nombre d'intégration, ce qui implique bien la récurrence d'apparition des modèles les plus probables. La nature aléatoire faite cependant apparaitre des modèles unique pour chaque fois que l'on repete l'algorithme.

```
## [[1]]
## [1] 823
##
## [[2]]
## [1] 839
##
## [[3]]
## [1] 805
##
## [[4]]
## [1] 856
##
## [[5]]
## [1] 832
##
## [[6]]
## [1] 834
##
## [[7]]
## [1] 815
##
## [[8]]
## [1] 815
##
## [[9]]
## [1] 842
##
## [[10]]
## [1] 861
```

```
##     nb
## 1 2412
```

```
##     nb
## 1 2623
```

```
##     nb
## 1 3369
```

Meilleurs modèles

On regarde la fréquence des modèles avec 2000 de burnin (qui ne sont pas comptés) pour 10000 et 100000 itérations.

```
##       probtop20_1e4iter
##  [1,]          0.117875 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
##  [2,]          0.095750 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
##  [3,]          0.053000 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
##  [4,]          0.041125 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
##  [5,]          0.038875 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
##  [6,]          0.036500 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
##  [7,]          0.024000 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0
##  [8,]          0.023375 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
##  [9,]          0.022625 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
## [10,]          0.020875 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
## [11,]          0.020375 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
## [12,]          0.011375 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1
## [13,]          0.009250 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0
## [14,]          0.007375 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [15,]          0.007000 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0
## [16,]          0.007000 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0
## [17,]          0.006625 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1
## [18,]          0.006500 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0
## [19,]          0.006250 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0
## [20,]          0.006125 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
```

```
## [1] "1e5 itérations:"
```

```
##       probtop20_1e5iter
##  [1,]          0.119571 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
##  [2,]          0.085235 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
##  [3,]          0.048480 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
##  [4,]          0.042888 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
##  [5,]          0.041612 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
##  [6,]          0.035796 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
##  [7,]          0.025541 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
##  [8,]          0.024796 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
##  [9,]          0.024102 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0
## [10,]          0.023184 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
## [11,]          0.021673 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
## [12,]          0.012184 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1
## [13,]          0.008102 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0
## [14,]          0.007724 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [15,]          0.006071 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1
## [16,]          0.006051 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0
## [17,]          0.005949 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0
## [18,]          0.005673 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0
## [19,]          0.005622 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
## [20,]          0.005490 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0
```

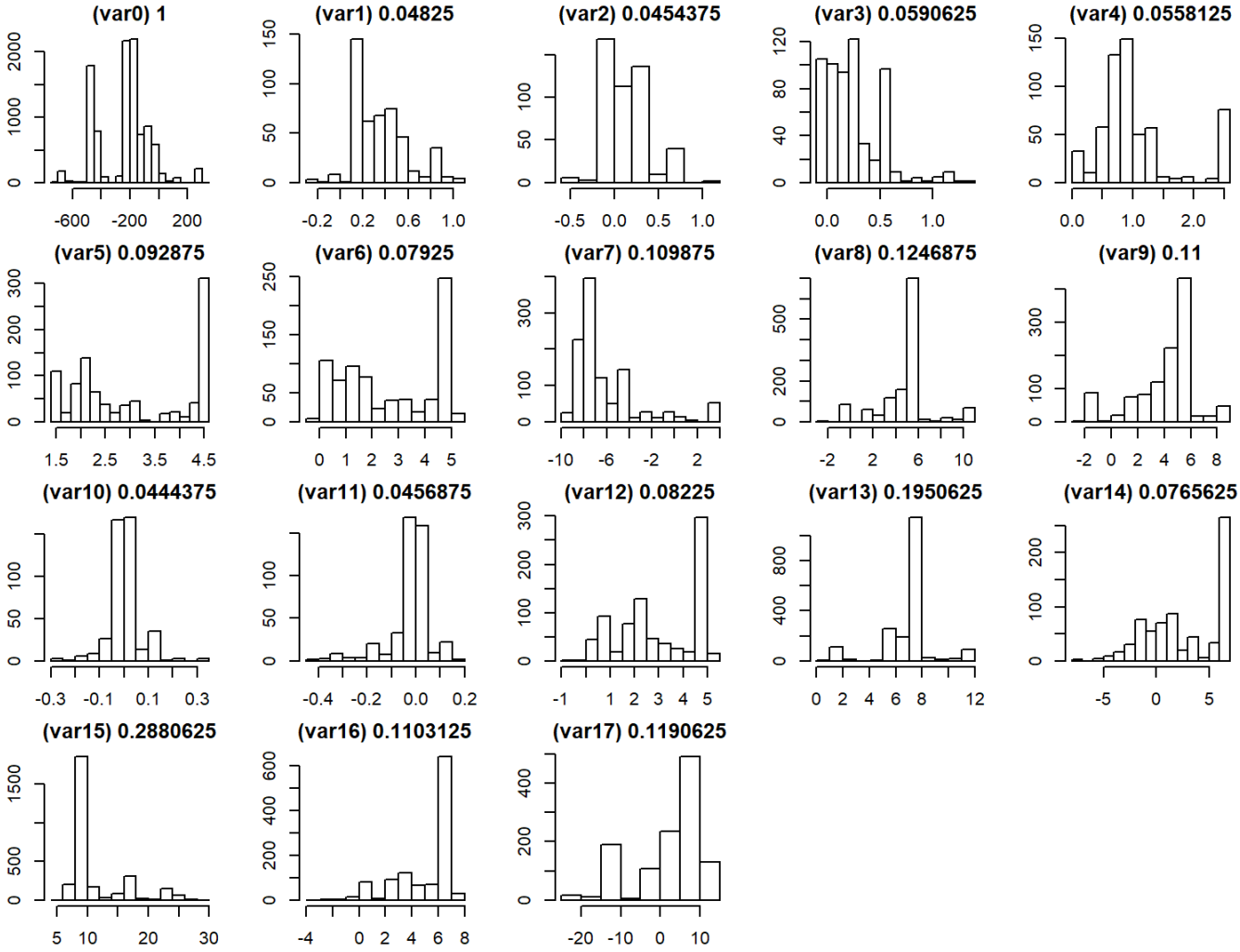
Tous d'abords la comparaison entre 10000 et 100000 itérations montre des résultats assez similaires et valide les points précédents.

- le meilleur modèle apparait avec une probabilité de 12%
- L'ordre des modèles entre 10000 et 100000 itérations est légèrement différent, par exemple le 4e plus probable avec 1e4 itérations est le 6e plus probable avec 100000 interaction.
- Les 11 premiers modèles le plus probables ne font intervenir d'une seule covariables.
- Le premier modèle avec 2 covariable est le 12e avec une probabilité assez faible de 1,2%

Pour s'assurer d'un bon rapport convergence / temps de calcul, on prendra 20000 itérations avec 4000 de burnin pour la suite de nos analyses

Estimation des coefficients de β

Les coefficients de β sont calculés en faisant des tirages aléatoires des modèles selon leurs probabilités. Les β sont ensuite calculés usuelement par régression linéaire, grâce à la fonction `lm`. Ainsi on obtient une distribution des β.



La figure ci-après illustre la distribution des β et la probabilité de tirage de chaque coefficient (l'intercepte est toujours incluse). On constate que:

- De façon générale, les coefficients de β qui sont issus de modèles avec de petites probabilités (moins de 10%) ont une distribution assez dispersée, sauf pour β_{10} et β_{11}
- Les valeurs moyennes des coefficients ne sont pas vraiment informatives, car les coefficients de β varient énormément selon le nombre de β_i inclus dans les modèles.
- Selon le modèle choisi l'intercepte peut être négative ou positive et les β_i également. Cependant lorsque l'un des β_i est sélectionné avec une forte probabilité, comme pour β_{15} qui est présent dans 29% des modèles, la distribution du coefficient est peu dispersée.

À noter quand dans la procédure utilisée ici tous les modèles tirés au sort sont concernés. Cette procédure pourrait être améliorée et raffinée, par exemple, en calculant le facteur de Bayes à chaque tirage pour ne garder que les modèles le plus informatifs.

Une implémentation basée sur ce principe est disponible dans le package `BAS`. Nous comparons les résultats calculés précédemment grâce à celui-ci dans la section suivante.

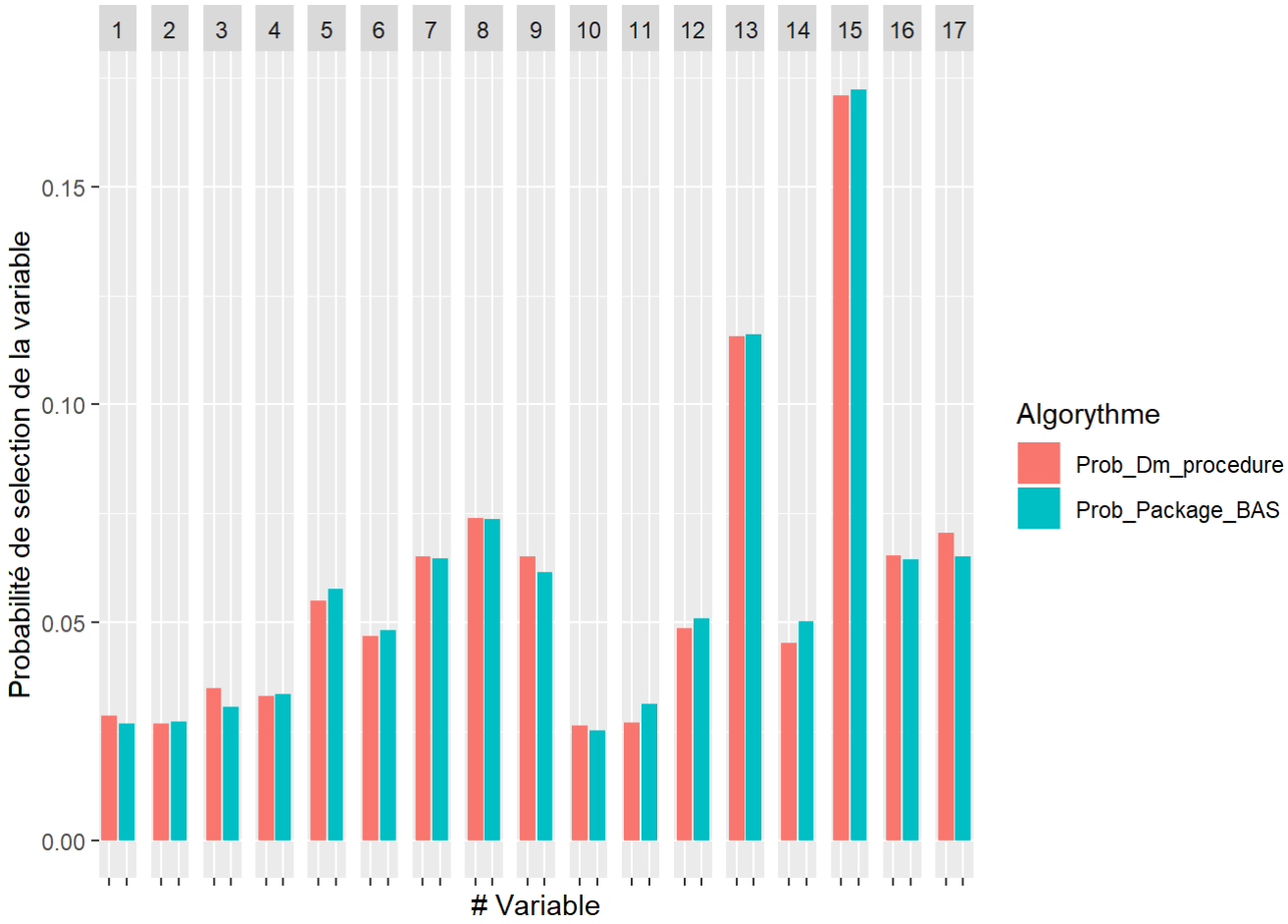
1.1.2 - Bayesian avec le package BAS

Pour reproduire l'algorithme utilisé précédemment

On utilise le package avec les paramètres suivants:

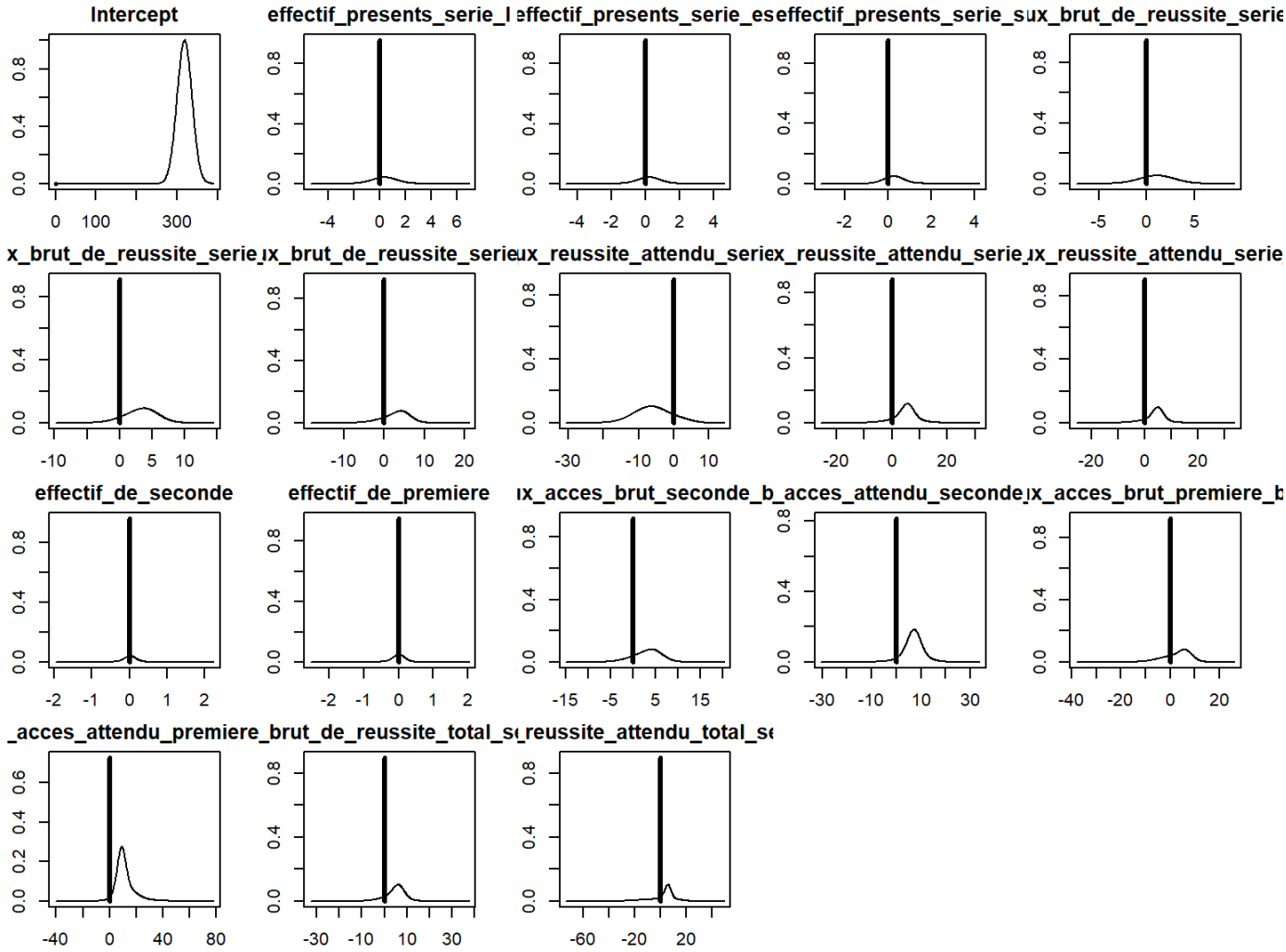
- `method="MCMC"` : méthode similaire au MCMC précédent avec en plus une sélection des modèles pertinents grâce au facteur de Bayes (algorithme MC3)
- `prior = "g-prior"` : prior de Zellner (comme précédemment)
- on prend le même nombre d'itérations (20000)

La moyenne des γ_i est comparé aux probabilités de sélection de chaque variable obtenue dans le package.



Les résultats pour la sélection de variables sont identiques entre le package et l'algorithme utilisé précédemment. Les petites variation dans les coefficients sont dues à la nature aléatoire de la chaine.

On compare maintenant avec la distribution des β_i



On constate clairement que la distribution des β_i et de l'intercepte est très différente des résultats obtenus avec notre algorithme. plusieurs explications sont possibles:

- L'algorithme que nous avons construit est erroné, cause que nous espérons peu probable
- Le package utilise une ou des étapes supplémentaires pour le calcul des β_i. Notamment grâce à un calcul de facteur de Bayes. Cela nous semble être la cause explicative.

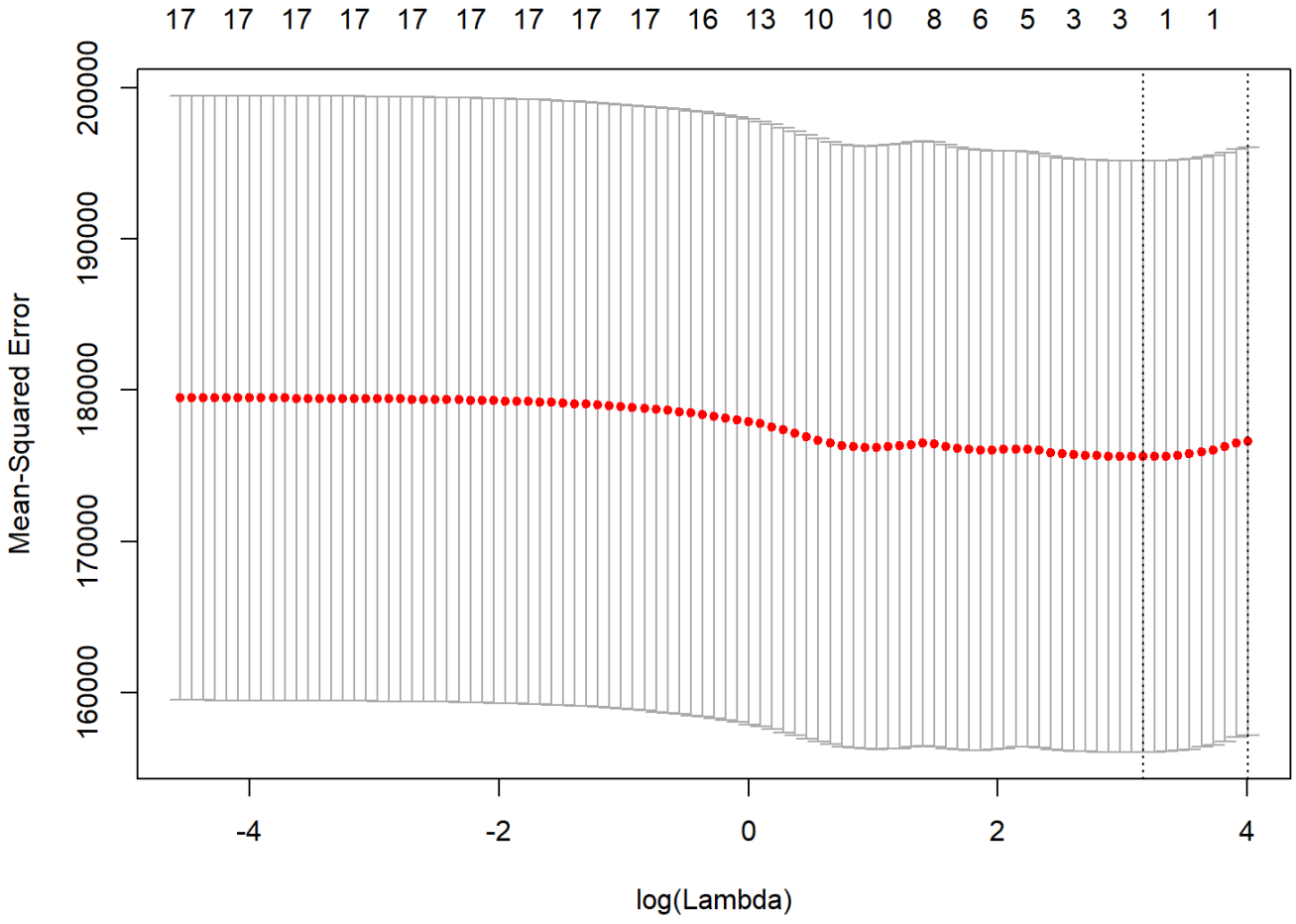
De plus on constate qu'avec les paramètres utilisés, l'hypothèse nulle est la plus probable pour toutes les variables (sauf l'intercepté). Cela se voit grâce à la hauteur de la barre verticale en 0 qui est systématiquement supérieure au maximum des distributions des β_i. Ainsi pour les questions suivantes, avant de comparer à l'analyse fréquente on souhaite trouver un modèle permettant de conserver au moins une variable. Pour cela on essayer par exemple: de faire une sélection de variables, de rajouter les variables catégorielles ou de modifier les paramètres (prior, calculé de vraisemblance, etc.)

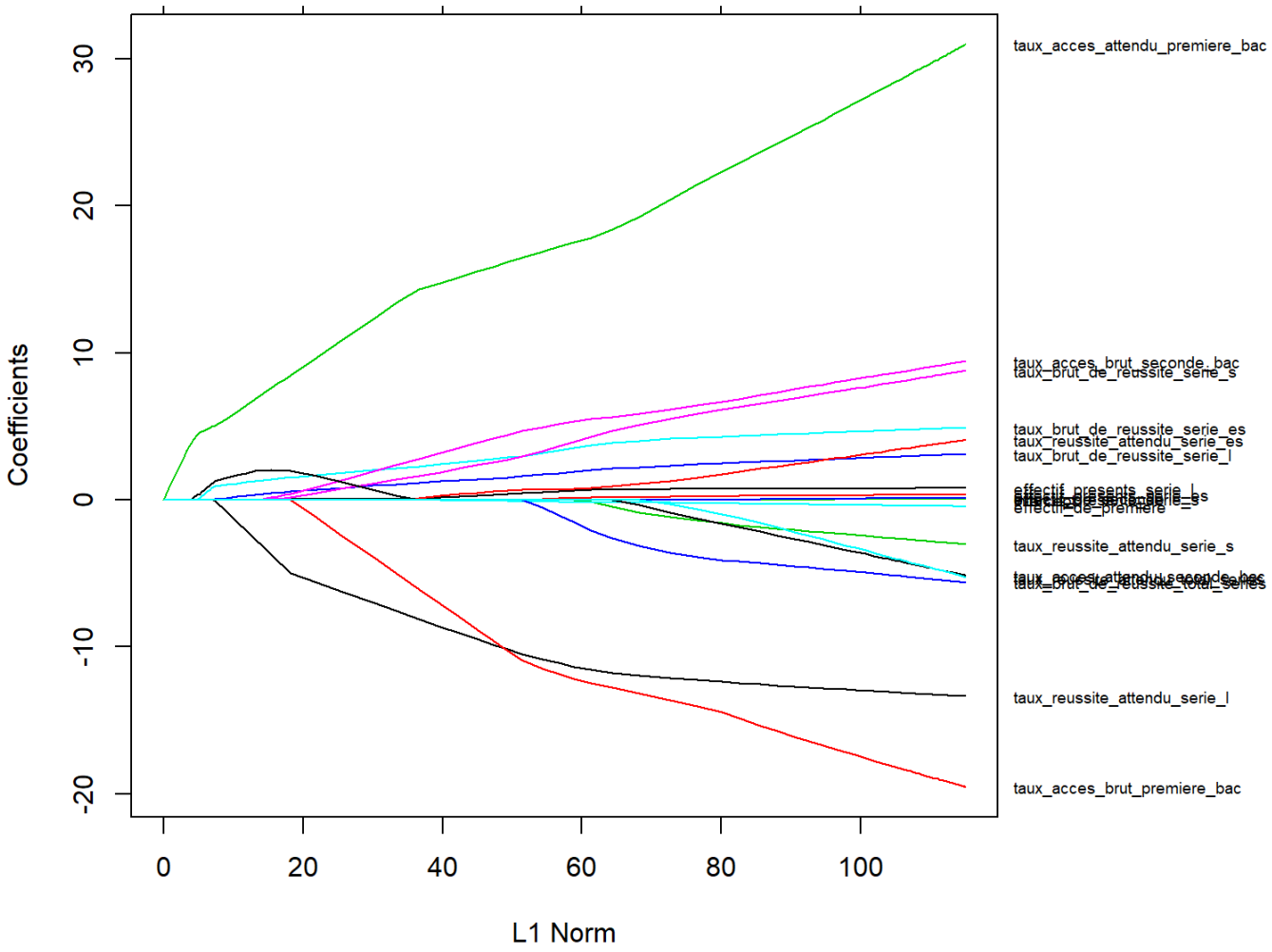
1.2 Choix des covariables significatives et compare à une analyse fréquentiste

1.2.1 Choix des covariables significatives

Lasso / ridge

Par validation croisée sur λ, la sélection de variables est effectuée, le paramètre `alpha` ne change pas les variables sélectionnées pour le λ optimal.


17 17 17 17 17 17 17 17 17 16 13 10 10 8 6 5 3 3 1 1
Mean-Squared Error
200000
190000
180000
170000
160000
-4
-2
0
2
4
log(Lambda)


Coefficients
30
20
10
0
-10
-20
0
20
40
60
80
100
L1 Norm
taux_acces_attendu_premiere_bac
taux_acces_brut_seconde_bac
taux_brut_de_reussite_serie_s
taux_brut_de_reussite_serie_es
taux_reussite_attendu_serie_es
taux_brut_de_reussite_serie_l
taux_reussite_attendu_serie_s
taux_reussite_attendu_serie_l
taux_acces_brut_premiere_bac

Les variables sélectionnées en plus de l'intercepte sont:

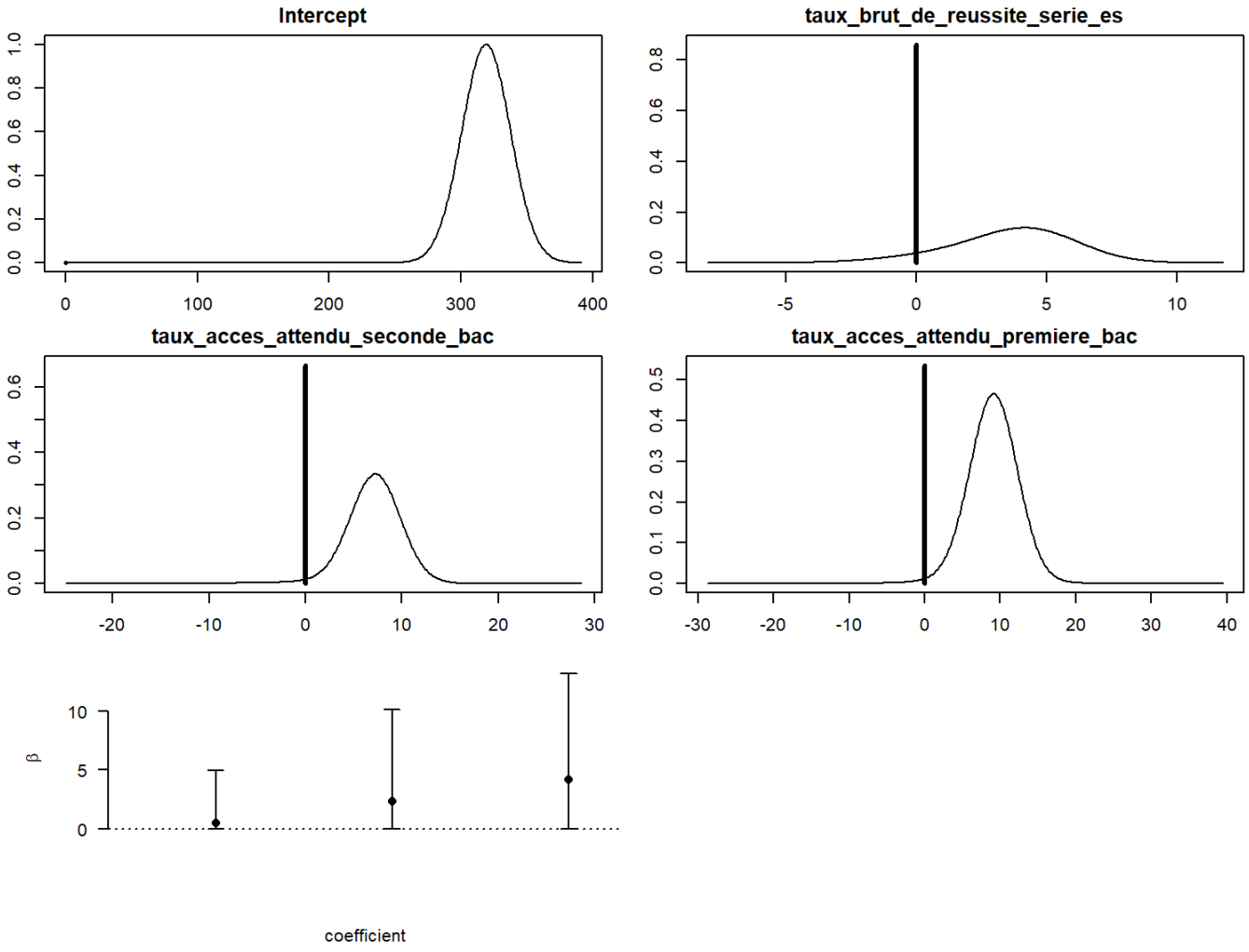
- `taux_brut_de_reussite_serie_es`
- `taux_acces_attendu_seconde_bac`
- `taux_acces_attendu_premiere_bac`

On remarque que ces 3 variables étaient également celles avec la probabilité marginale d'inclusion la plus forte dans le modèle bayésien calculé précédemment.

Bayesien sur les variables sélectionnées

On constate que même avec moins de variables, l'intervalle de confiance des coefficients recouvre le 0 . La probabilité marginale d'inclusion des 3 variables (sauf l'intercepte) est plus faible que l'hypothèse nulle. Le modèle n'est toujours pas satisfaisant, on regarde si paramètres du package / lois utilisés ne peuvent pas être modifiés pour trouver un meilleur modèle.

```
## NULL
```

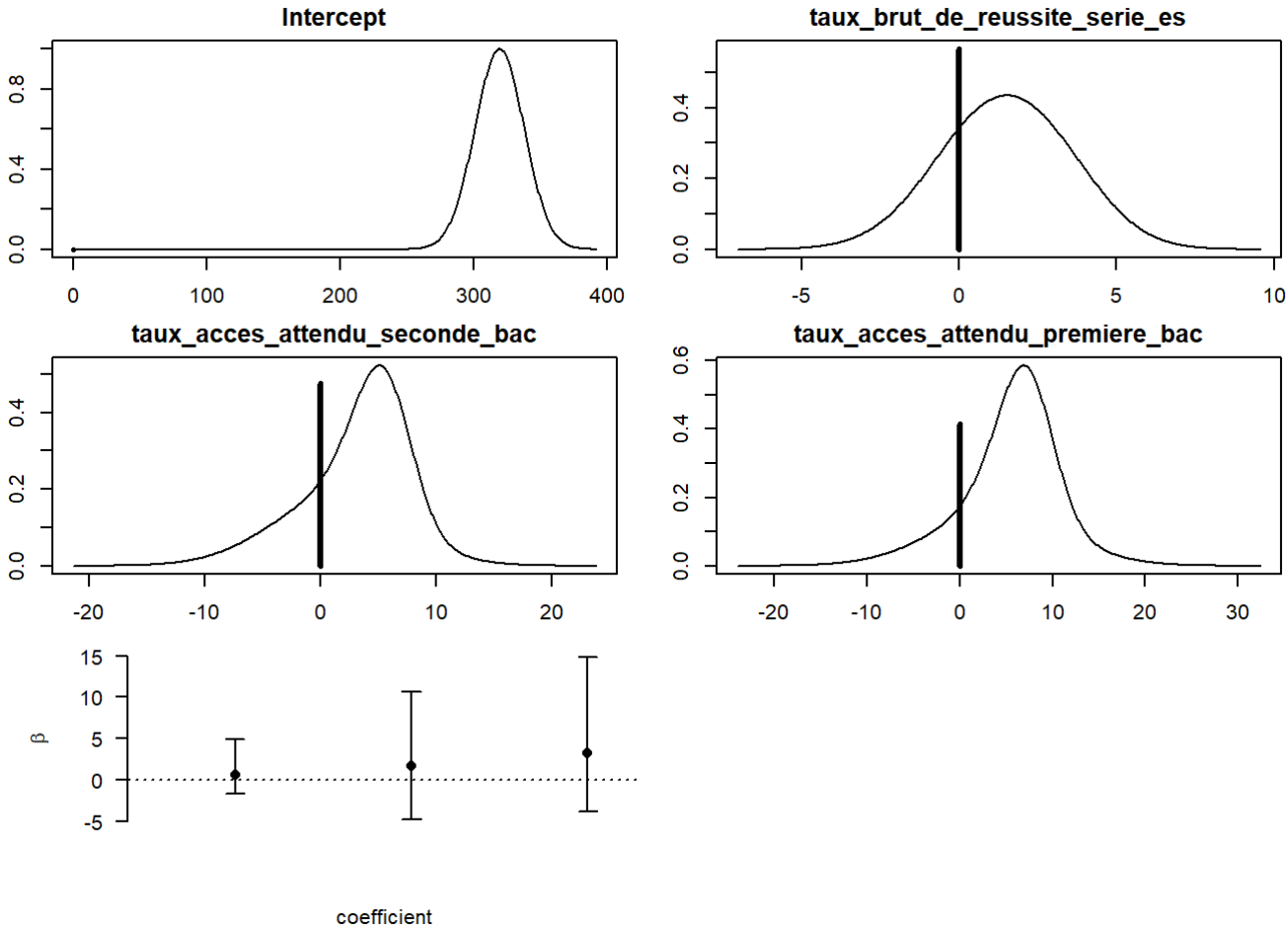


Recherche d'un meilleur modèle sur les variables sélectionnées

Paramètres utilisés:

- `method="MCMC+BAS"` : méthode MC3 + échantillonnage préférentiel
- `modelprior = beta.binomial(1,1)` , correspond à une loi uniforme discrète, on considère que chaque valeur de Barre à la même probabilité de réalisation, a prairie, mais avec un caractère plus discret, pouvant probablement se rapproche plus de ce qui est observé dans les données.
- `prior = hyper-g-laplace` : pour calculer la vraisemblance

```
## NULL
```



Bien que la probabilité postérieure des variables 2 et 3 est plus importante que l'hypothèse nulle, l'intervalle de confiance recouvre toujours le 0. De plus les coefficients de β sont exactement les mêmes que ceux calculés au point précédent. Finalement on a juste changé la manière dont la vraisemblance est calculée pour arriver au même résultat.

1.2.2 - Comparaison avec le modèle fréquentiste

On effectue une sélection de variables par la fonction step et le critère de l'AIC:

- Lorsque l'ont inclus toutes les variables, le step sélection 3 variables: taux_acces_attendu_premiere_bac, taux_reussite_attendu_serie_l et Matiere
- Lorsque l'on retire les variables catégorielles, seules taux_acces_attendu_premiere_bac et taux_reussite_attendu_serie_l sont concervés. En détail :

Lorsque l'on regarde les coefficients, on constate que:

- L'intercepte est négative on pars donc d'une `Barre` négative, cependant le coefficient de `taux_acces_attendu_premiere_bac` est tellement importante que toutes les prédictions restent positives.
- Le `coeficient de taux_reussite_attendu_serie_l` est négatif. Ce qui laisse penser que la `Barre` d'accès diminue losque l'établissement à un meilleur taux de réussite pour la section L. Cela parait étonnant, car ce sont généralement les établissements avec les meilleurs taux de réussite qui sont les plus convoités et donc ceux avec le Barre la plus haute.

Finalement on constate que `taux_acces_attendu_premiere_bac` qui est la variable donc le coefficient et les plus vraisemblablement différents de 0 avec l'approche Bayésienne est également celui avec la p-value la plus faible (0.002) pour la sélection de variables du modèle linéaire avec le step et l'AIC.

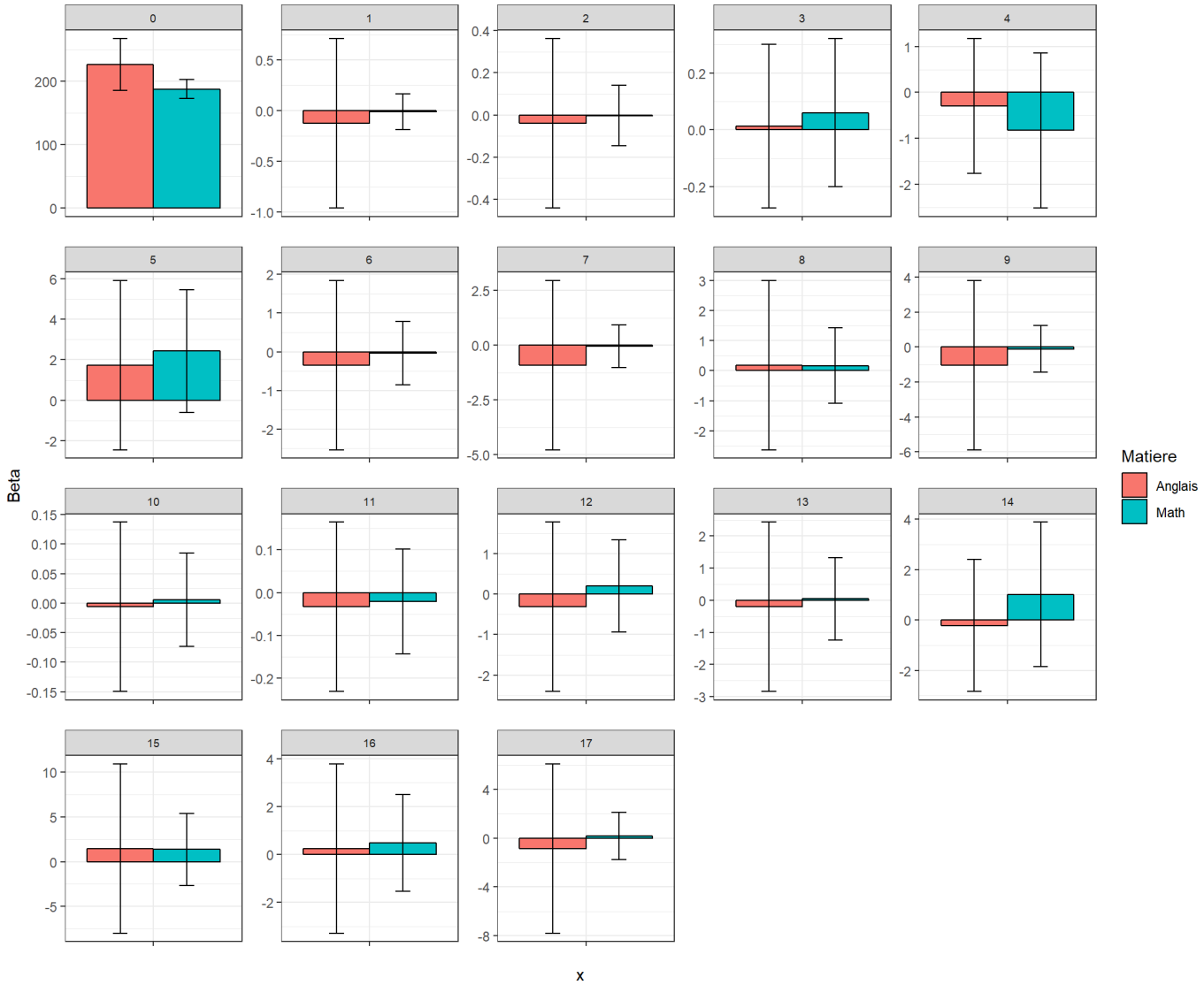
Afin de comparer les deux approches, on peut procéder à une validation croisée :

1.3 Analyse des mutations en mathématiques et en anglais

Ces matières sont parmi celles les plus représentés dans le jeu de données avec respectivement 52 et 59 individus (établissements) pour l'Anglais et la Mathématiques. Parmi ceux-ci, 31 Matiere sont issues du même établissement.

1.3.1 - Bayésien sur toutes les variables

On effectue de nouveau la régression Bayesienne, avec les paramètres optimisés, sur 2 jeux de données avec seulement les matières Anglais et Mathématiques. Tous d'abord, nous avons conservé toutes les variables. La figure suivante indique les coefficients ainsi que leurs intervalles de confiances. L'intercepte correspond à la variable 0 dans les figures suivantes.



En comparant les coefficients trouvés, on constate que:

- Certaines variables ont des coefficients relativement proches avec des valeurs proches et le même signe
- Certaines variables sont du même signe, mais avec des valeurs de coefficients assez éloignés
- Certaines variables ont des coefficients de signes opposés

Les intervalles de confiances ne sont pas commentés, car forcément globalement plus petit pour la matière avec le plus d'individus (Mathématiques)

Ainsi on peut conclure que les covariables agissent de manière différente pour ces deux disciplines. Surtout pour les variables avec des coefficients de signes opposés.

Le détail des variables et modèles les plus fréquents est illustré ci-dessous.

On constate que les modèles et variables les plus fréquents sont très différents selon la matière.

1.3.2 - Bayésien sur les variables sélectionnées précédemment

On recommence l'analyse avec un nombre de variables réduites. Pour cela on réutilise les variables sélectionnées par le lasso sur le jeu de données complet effectué précédemment.

On constate que lorsque l'on se place dans un modèle plus simple avec seuelement 3 variables + l'intercepte:

- Les coéficients n'ont pas tous le même signe
- les valeur des coéficients sont relativement diférentes

De plus, dans les figures suivante, on constate que les modèles choisie sont encore une fois différents :

- Les deux premiers modèles les plus probables pour les mathématiques font intervenir l'intercepte et une variables
- le Modèles le plus probable pour l'anglais n'as que l'intercepte.

Au vu des constatations des points précédents, on conclut que les covariables agissement de manières différentes pour ces deux disciplinent.

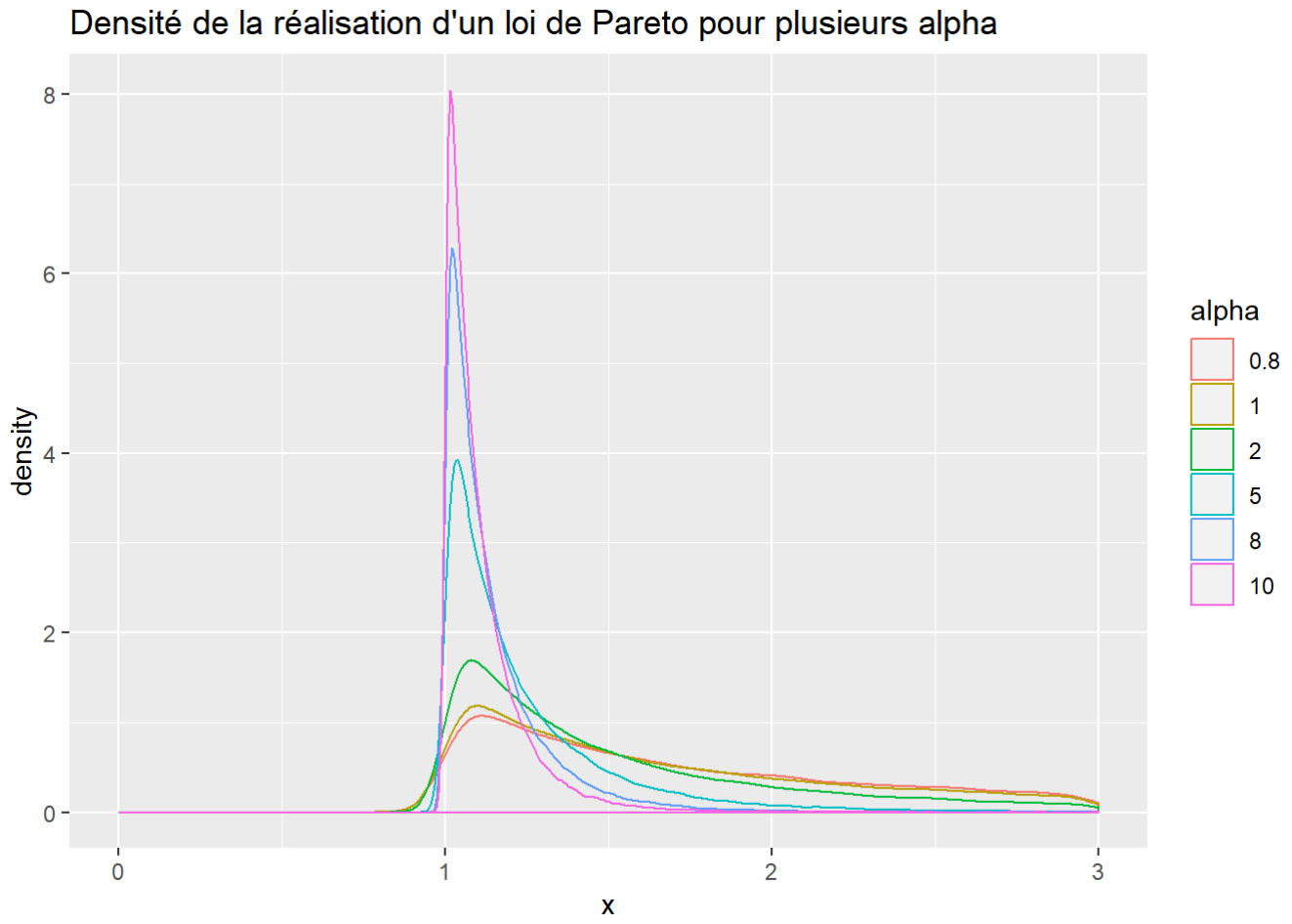
2 - Loi de Pareto

2.4 - Réalisation d'une loi de Pareto et impact du paramètre α

On utilise le package `VGAM` pour tirer les réalisations de loi de Pareto dont la densité s'écrit:

$$f_Z(z; m; \alpha) = \alpha \frac{m^\alpha}{z^{\alpha+1}} \mathbb{I}_{\{z \geq m\}}$$

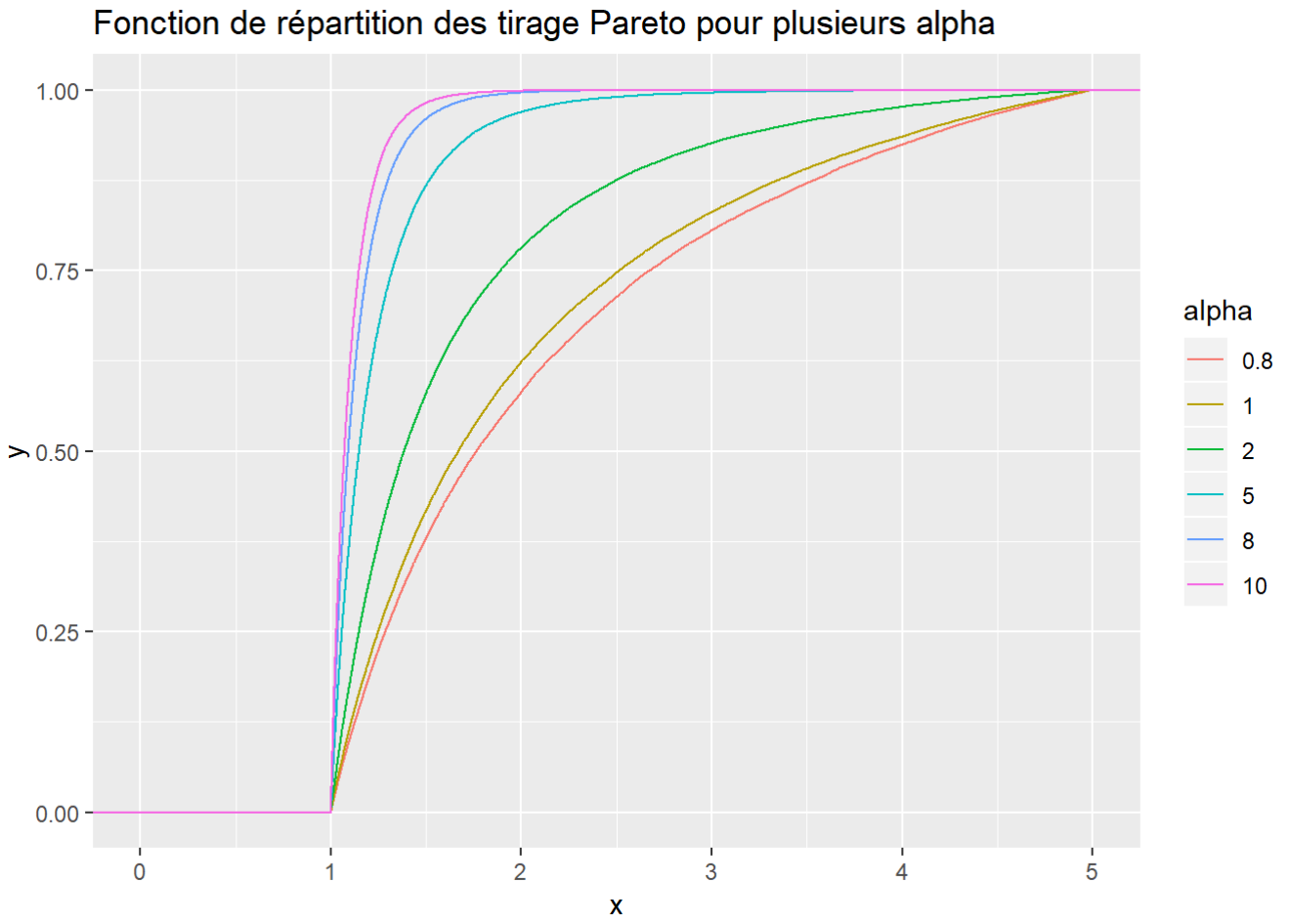
La figure suivante illustre la densité de ces tirages pour plusieurs valeurs de α variant entre 0.8 et 10 et $m = 1$.



On relève principalement deux impactes du paramètre α:

- Conformément à la formule de la densité, la valeur de α qui représente la forme de la distribution resserrée la distribution pour des valeurs élèves
- Pour α est petit et proche de 0, plus la densité est étalée: a queue de distribution est plus lourde, la densité s'aplatit et le maximum de densité se déplace vers la droite.
- Lorsque α est très grand les valeurs des tirages sont très proches de 0 la densité tends vers l'infini. Cela se rapproche d'une fonction de Dirac

Afin de mieux comprendre le comportement on trace la fonction de répartition pour les différents α :



On voit bien la convergence plus rapide de la distribution cumulative pour une loi de Pareto lorsque α est grand, et inversement pour α petit.

2.5 & 2.6 - Choix de loi à priori et calcul des posteriors

Inférence sur la borne supérieure de Barre

Plusieurs choix s'offrent à nous quant à l'interprétation de cette question et de comment utiliser la croyance initiale que la variable aléatoire Barre ($n = 510$, le nombre de données de Barre) suit une loi de Pareto de paramètre $\alpha, m > 0$ et de densité :

$$f_Z(z; m; \alpha) = \alpha \frac{m^\alpha}{z^{\alpha+1}} \mathbb{I}_{z \geq m}$$

La loi uniforme peut s'appliquer sur les données et intervient sur notre croyance du modèle des données pour calculer la borne maximale de Barre. Dans ce cas on choisit comme prior une loi de Pareto et suppose que les valeurs de barre (> 0) sont réparties de façon uniforme. Prior Pareto:

$$\pi(w) = \alpha \frac{m^\alpha}{w^{\alpha+1}}$$

Vraisemblance : loi uniforme avec $w > max(z_i)$ pour couvrir le jeu de données

$$l(w|z) \propto \frac{1}{w^n}, w > max(z_i)$$

Ce qui donne une postérieur :

$$\pi(w|z) \propto \frac{(\alpha+n).max(m,z_i)^{\alpha+n}}{w^{\alpha+n+1}}$$

De cette façon et avec les paramètres α, m de connus on peut inférer sur une valeur maximale de Barre. Ce qui peut être utile pour prévoir sa mutation professionnelle !

Inférence sur le paramètre m

m est le paramètre de localisation de la distribution, il régit la position de la densité de probabilité. on voit bien cela directement dans la formule de la densité avec $\mathbb{I}_{z \geq m}$.

Dans les calculs suivants on utilise $M = 1/m$ qui est le paramètre de précision.

Dans le cas ou α serait connus et M inconnu on peut utiliser un modèle de données suivant une loi inverse Pareto (fonction puissance) et une Prior de Pareto de paramètres (a, b).

Prior Pareto:

$$\pi(M) = a\frac{b^{\alpha.a}}{M^{a+1}}, M > b$$

Vraisemblance :

$$l(M|z) \propto \ M^{\alpha.n}, M < min(z_i)$$

Ce qui donne une postérieure:

$$\pi(M|z) \propto \frac{(a-\alpha n).b^{(a-\alpha n)\alpha}}{M^{a-\alpha n+1}}; a > \alpha n, M > b$$

On retrouve une loi de conjugué de Pareto. Cette approche peut être utilisée pour inférer, par exemple, sur la localisation du minimum de la distribution de Barre, notamment si l'ont se place pour une matière particulière.

Inférence sur le paramètre α

Finalement paramètre de forme α qui décrit la dispersion de la loi de Pareto peu être estimé à l'aide en se plaçant dans un cas particulier d'un modèle exponentiel. En effet, si Z suit une distribution de Pareto, alors $Y = log(X/m)$ suit une distribution exponentielle. On à bien $\lambda > 0$, car $\alpha > 0$. Dans se cas mieux connus nous pouvons utiliser la loi conjuguée Gamma de paramètre (a, b) en prior.

Note: on fait me choix de se placer dans le cas d'une loi gamma $\Gamma(a, b)$ aussi noté $\Gamma(k, \theta)$, car le paramètre θ représente l'échelle pour la loi de Pareto alors que dans $\Gamma(\alpha, \beta)$, β représente un paramètre de taux ($\beta = 1/\theta$).

Prior $\Gamma(a, b)$:

$$\pi(\alpha) = \frac{\alpha^{a-1}e^{-\alpha/b}}{b^a\Gamma(a)}, \alpha > 0$$

Vraisemblance :

$$\begin{aligned} l(\alpha|z) &\propto \prod_{i=1}^{n} \alpha\frac{m^\alpha}{(z_i)^{\alpha+1}} \\ &\propto \frac{\alpha^n m^{n\alpha}}{(\prod_{i=1}^{n} z_i)^{\alpha+1}} \end{aligned}$$

Ce qui donne une postérieure :

$$\begin{aligned}\pi(\alpha|z) &\propto \pi(\alpha).\, l(\alpha|z) \\ &\propto \frac{\alpha^{a+n+1} e^{-\alpha/b'}}{b'^{a+n}\Gamma(a+n)}, \\ b' &= \frac{1}{\frac{1}{b} + log(\prod_{i=1}^{n} z_i)) - nlog(m)} > 0\end{aligned}$$

De plus $b' > 0$ car $log - (\prod_{i=1}^{n} z_i) - nlog(m) = 2682 - 1576 > 0$ pour $m = 21$. Dans la littérature on retrouve la notation x_m comme étant le minimum sur les données et qui est égal à m dans la notation du devoir. Avec cette approche on peut donc penser que le calcul de la portérior fonctionne encore pour des m plus grand. Il serait intéressant de se servir de l'approche détaillée précédemment pour estimer le paramètre m et voir si l'ont retrouve une valeur plus grande, qui annulerait peut-être $log(\prod_{i=1}^{n} z_i)) - nlog(m)$

Finalement on revient à prendre une loi à postériori plus facile à implémenter, si $\pi(\alpha) = \Gamma(a, b)$, avec b implémenté comme un paramètre d'échelle, nous trouvons le postérieur suivante:

$$\pi(\alpha|z) \propto \Gamma(a+n, b + \sum_{i=1}^{n} ln\frac{z_i}{m})$$

Ce qui donne une logpostérior:

$$(a+n).\, log(b + \sum_{i=1}^{n} ln\frac{z_i}{m}) + (a+n-1).\, log(z_i) - [b + \sum_{i=1}^{n} ln\frac{z_i}{m}].\, z_i - \Gamma(a+n)$$

2.7 - Tirage de la loi à postériori

Nous simulons la loi à postériorie à l'aide de l'algorithme de Metropolis-Hastings. Pour l'exploration de la chaine on choisi une loi de proposition normal $\mathcal{N}(\alpha_t, \tau^2)$ avec τ la variance de la proposition qui permet de contrôler la vitesse l'exploitation de la loi.

Tous d'abord nous cherchons à optimiser le paramètre $ $. On choisie une loi à prairie de $\Gamma(1,1)$ et $\alpha_0 = 0.1$ pour l'initialisation.

Nous analysons la proportion d'acceptation de l'algorithme pour différentes valeurs de $\tau = (0.01, 0.5, 0.1, 0.275)$. On choisit de faire 20000 itérations afin d'être certain de bien avoir exploré tous les modes de la loi.

```
## --------------------------------Proportions d'acceptations--------------------------------
--
```

```
## ---Proportions d'acceptation pour tau= 0.01
## [1] 0.5145
```

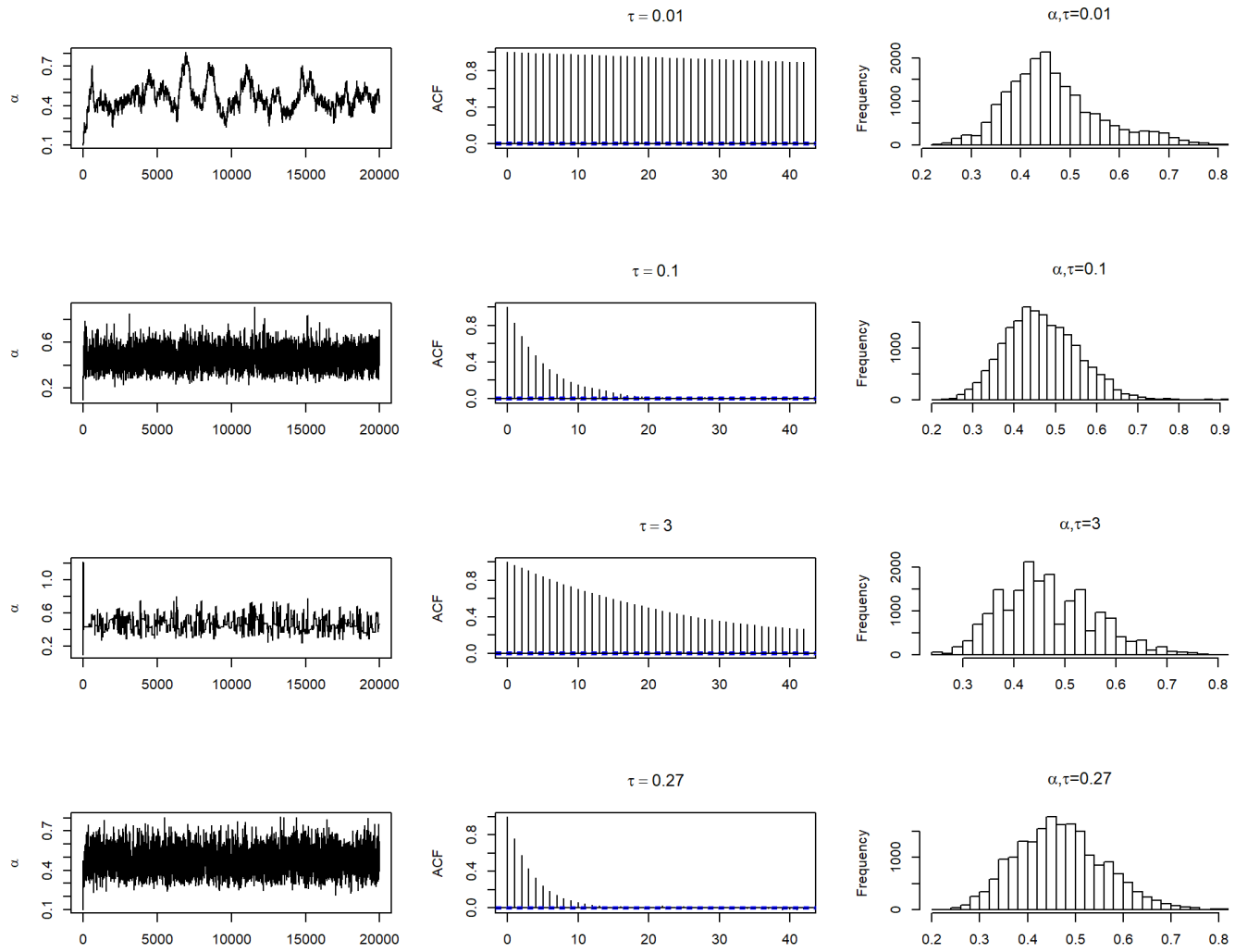
```
## ---Proportions d'acceptation pour tau= 0.1
## [1] 0.4088
```

```
## ---Proportions d'acceptation pour tau= 3
## [1] 0.02505
```

```
## ---Proportions d'acceptation pour tau= 0.275
## [1] 0.2283
```

Pour $\tau = 0.275$ on la proportion d'acceptation est proche de l'ordre de 23%, ce qu est proche de l'optimum trouvé par Robert, Gelman & Gilks (1997) dans un cas plus simple et également une bonne vitesse de mélange qui permet d'explorer correctement la loi.

Graphique nous regardons (après 1000 itérations de burnin) le tracé de chaine , l'autocorrélation et la distribution des valeurs de α.




```
## --------------------------------Tailles d'échantillons effectives ESS----------------------
-------------
```

```
## ESS Tau=0.01
##  249.6
```

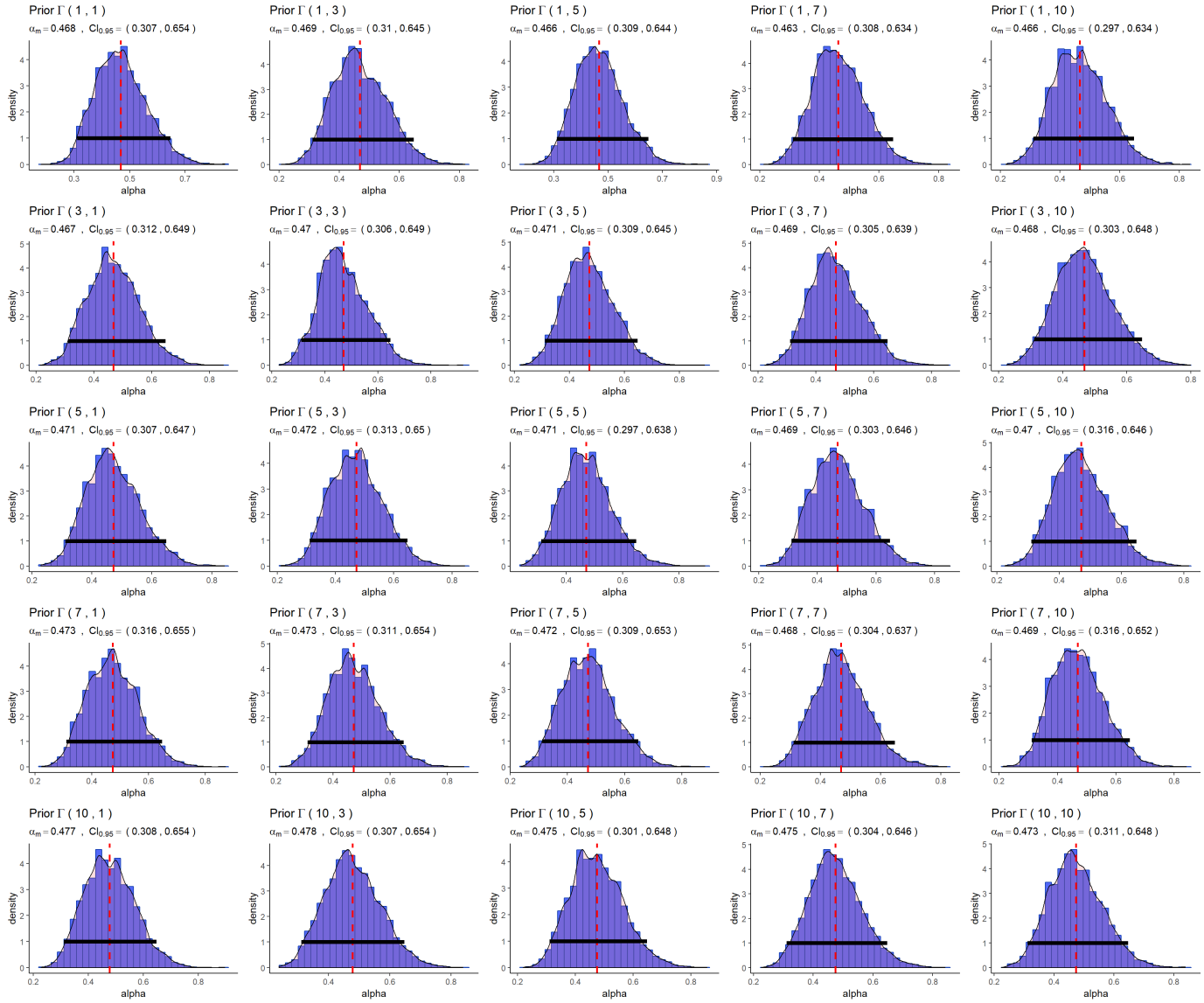
```
## ESS Tau=0.1
##  1923
```

```
## ESS Tau=3
##  446.7
```

```
## ESS Tau=0.275
##  2694
```

- Pour $\tau = 0.01$: la variance de la proposition est trop faible, la chaine met trop de temps à explorer la loi et l'autocorrélation décroît très lentement
- Pour $\tau = 0.1$: la proportion d'acceptation est de 40% sont seulement 10% de moins que pour le $précédent. Cependant l'autocorrélation décroit rapidement et la trace de la chaine explore bien la loi
- Pour $\tau = 3$: seulement 3% des propositions sont acceptés. Graphiquement :la chaine reste bloquée à la même valeur de nombreuses fois, l'autocorrélation décroit lentement et l'histogramme des valeurs est de moins bonne qualité que pour les autres valeurs de τ même s'il reste relativement proche
- Pour $\tau = 0.275$, la trace de la chaine mélange rapidement, l'autocorrélation décroît très rapidement. L'histogramme est très proche de celui pour $\tau = 0.1$, cependant la taille de l'ESS est beaucoup plus importante avec respectivement 2900 contre 1900.

Pour la suite nous utiliserons $\tau = 0.275$. La figure suivante illustre la distribution de tirage de loi de α à postériorie pour différents paramètres de la loi à priori Γ:



Nous constatons que les distributions sont toutes similaires pour différentes valeurs de paramètres de la prior. Ce qui est attendu et permet de se rassurer sur la validité des algorithmes implémentés.

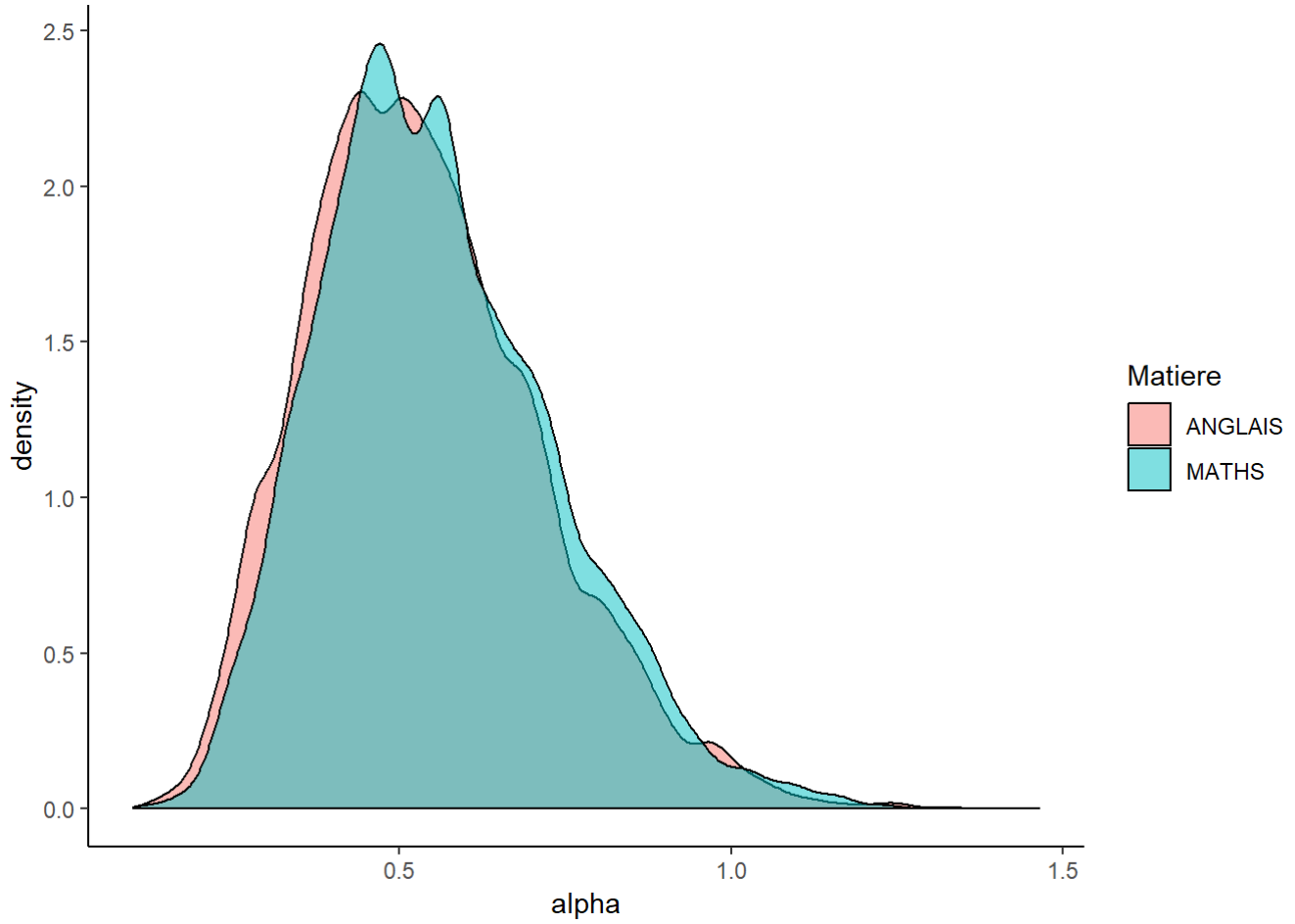
Intervales de crédibilité à 95% :

- Pour $\Gamma(1,1)$ en prior: $\alpha = 0.4681361; CI_{0.95}(\alpha) = (0.3066301, 0.6535263)$
- Pour tous les 25 tirages avec des paramètres de la prior Γ différents:
  $\alpha = 0.4704288, CI_{0.95}(\alpha) = (0.3070352, 0.6485245)$

2.8 - Tirages sur les données des matières Mathématiques et Anglais uniquement

Nous recommençons la procédure précédente pour les données de `Barre` de Mathématiques et Anglais. le jeu de données étant largement plus petit, les valeurs de $\tau = 0.6$ est réajusté pour obtenir environ 23% de taux d'acceptation. On choisie $\Gamma(1, 1)$ comme prior.

Les densités des tirages de α pour les deux matières, dans la figure ci-dessous, se recouvrent presque parfaitement.



Nous trouvons pour les deux matières:

$\alpha_{maths} = 0.5583876, CI_{0.95}(\alpha) = (0.2413205, 0.8972745))$
$\alpha_{anglais} = 0.5388604, CI_{0.95}(\alpha) = (0.2324066, 0.8941443))$

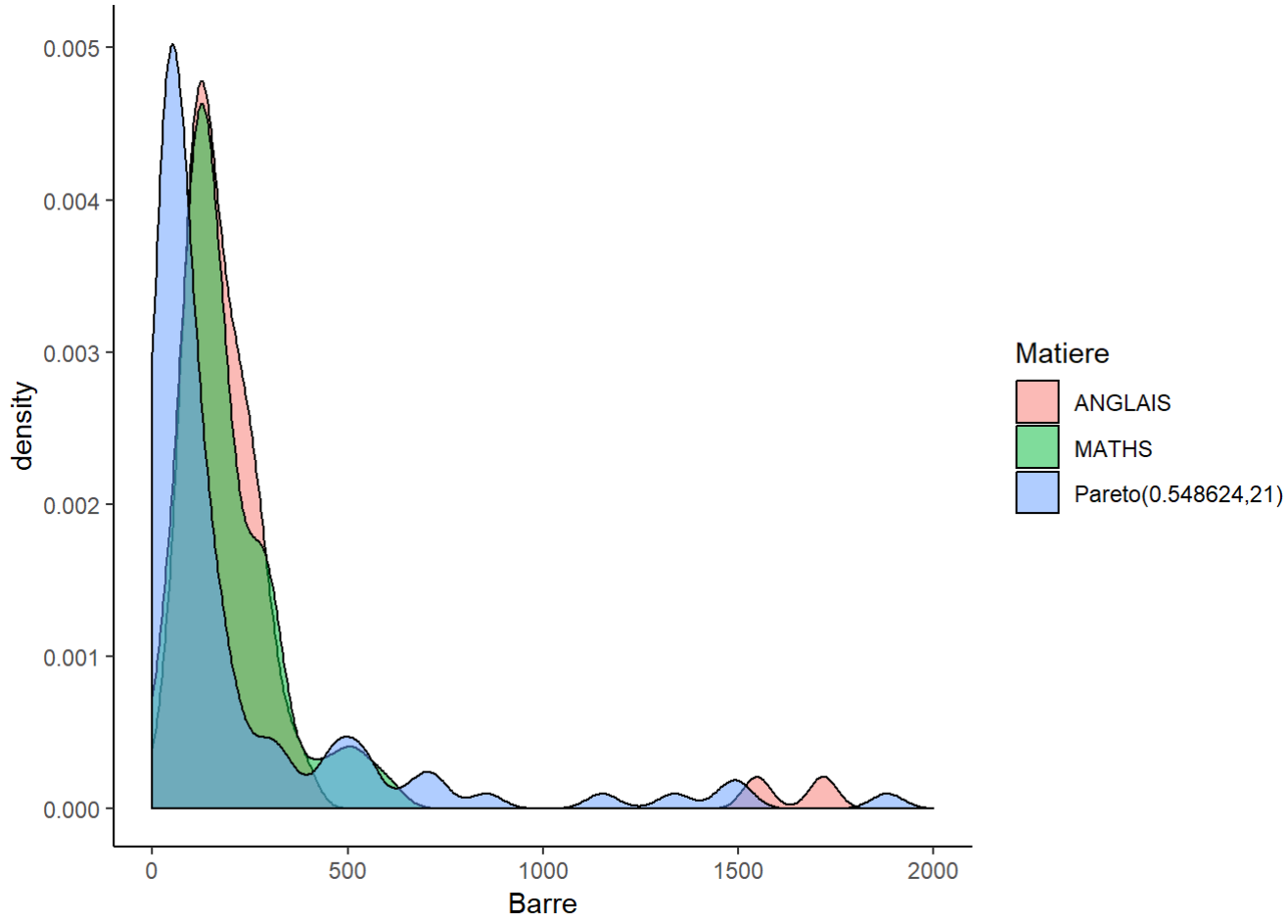
Au vu des valeurs de α sont proches et les grands intervalles de confiances qui se reçoivent également presque parfaitement, nous estimons que $\alpha_{maths} = \alpha_{anglais}$

Nous pouvons également vérifier que les distributions de α sont similaires on peut également utiliser le test de Kolmogorov-Smirnov (car α est continue):

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  alpha_anglais and alpha_math
## D = 0.054, p-value <2e-16
## alternative hypothesis: two-sided
```

Le test indique que les deux distributions pour α_{maths} et $\alpha_{anglais}$ sont identiques.

Nous pouvons également tracer la densité d'un tirage de loi de Pareto avec les paramètres $\alpha_{moyen} = 0.548624$ et $m = 21$ et comparer avec les distributions de `Barre` pour Maths et Anglais:



La forme des trois distributions (donnée pas le paramètre α) est très proche. L'échelle de la distribution de Pareto calculé avec $m = 21$ semble être différents de celle des matières.

Le paramètre m peut s'interpréter comme la "Barre" représentée par le pic de la distribution de Pareto. Dans les faits il dont être au moins supérieur au minimum de la variable `Barre` , car cela représente le minimum des tirages de Pareto.

Pour conclure, il paraît raisonnable de penser que le paramètre m est bien estimé à 21 car s'il était plus grand, bien que le pic de la distribution de Pareto serait plus proche des pics de Maths en Anglais, la distribution ne pourrait pas générer de valeur en dessous du paramètre m alors que les Barre de que le minimum de Barre pour les matières est de 21.

Concernant le paramètre α, il, la convergence de la procédure et la représentation graphique indiquent également qu'il a bien été estimé. Cependant les distributions des matières et de la loi de Pareto généré avec ces paramètres ne coïncident pas: Pareto semble être décalé vers la gauche. On est tout de même satisfait du résultat car on a réussi à bien approché la distribution d'un phénomène découlant d'un mécanisme sociologique par une formule mathématique. La distribution de Pareto, que l'on doit à Vilfredo Pareto, est d'ailleurs utilisé avec succès pour modéliser de nombreux phénomènes sociologiques.