ENSAI École nationale de la statistique et de l'analyse de l'information

Travaux Pratiques de Régression Bayesienne
2026-01-22

Contents

Table des matières

Liste des tableaux

Liste des figures

0 Introduction

Dans ce travail, on s'intéresse à l'analyse et à la prédiction du score de satisfaction de clients d'une chaîne câblée à partir de leurs usages des différentes chaînes de télévision. Le jeu de données étudié comporte 150 observations, pour lesquelles un score de satisfaction global est mesuré, ainsi que 160 variables explicatives correspondant aux temps passés et au nombre de visites sur différentes catégories de chaînes. Les variables explicatives ont été préalablement normalisées.

Le nombre de variables explicatives étant supérieur au nombre d'observations, les méthodes classiques de régression linéaire ne sont pas adaptées. On adopte donc une approche de régression bayésienne, qui permet à la fois de régulariser les coefficients, d'améliorer la capacité de prédiction et de sélectionner les variables les plus pertinentes en terme d'influence.

L'échantillon est aléatoirement divisé en deux parties : un jeu d'apprentissage de 100 observations, utilisé pour entraîner les modèles, et un jeu de test de 50 observations, utilisé pour évaluer les performances prédictives de ces modèles afin de les comparer. Afin d'assurer la reproductibilité des résultats, la clé aléatoire utilisée pour ce découpage est fixée à 2026.

Quatre méthodes de régression sont ensuite comparées : la régression aléatoire de type RR-BLUP, la régression bayésienne de type Bayes A, le LASSO bayésien et la méthode de sélection de variables SSVS. Ces approches sont évaluées en termes de qualité de prédiction, de comportement de shrinkage et de capacité à identifier les variables explicatives les plus pertinentes.

1 Importation et exploration des données

Avant la mise en oeuvre des différents modèles de régression bayésienne, il est nécessaire de présenter brièvement le jeu de données et d'examiner ses principales caractéristiques. Cette étape permet de vérifier la cohérence des données, d'identifier d'éventuelles anomalies et de mieux comprendre la structure du problème étudié. Néanmoins, le grand nombre de variables explicatives par rapport au nombre d'observations limite la pertinence de certaines analyses exploratoires classiques, et justifie le recours à des méthodes de régularisation et de sélection de variables dans la suite de l'étude.

1.1 Importation des données

```
data <- read.csv("data/telecat.csv")
dim(data)
```

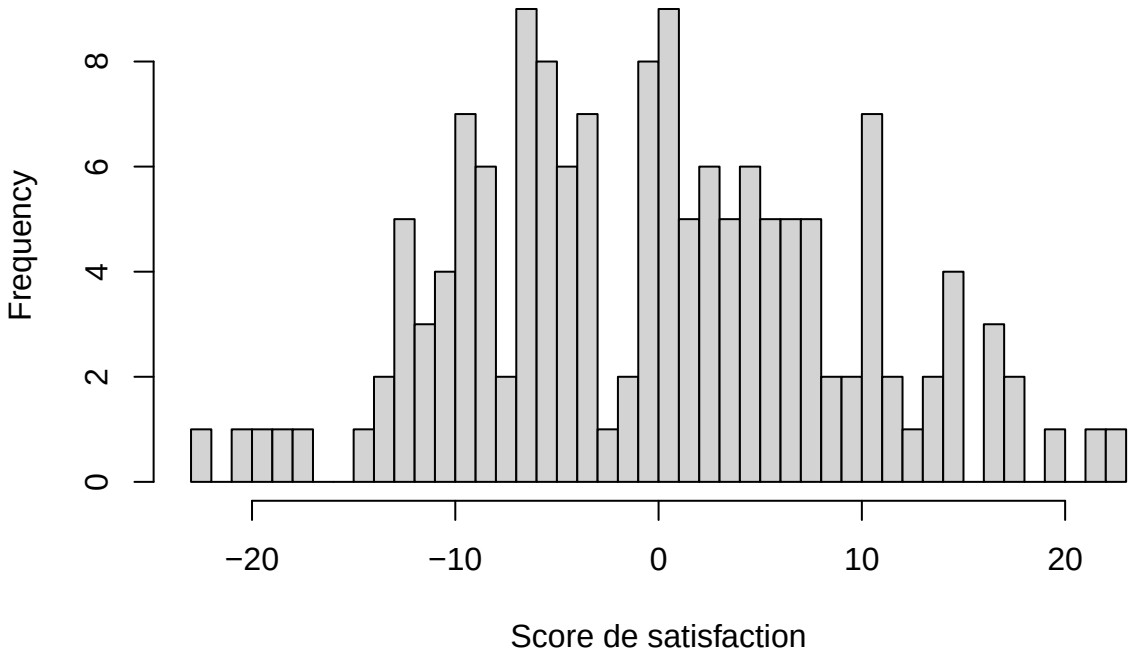
Le nombre d'observations dans le jeu de données est $n = 150$, pour $p = 160$ variables explicatives (en mettant de côté les variables X, Y et $sexe$). Le nombre de variables explicatives étant supérieur au nombre d'observations, une estimation par moindres carrés ordinaires n'est pas possible. Ce contexte justifie l'utilisation de méthodes bayésiennes pénalisées.

1.2 Distribution de la variable de réponse

```
summary(data$Y)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -22.2264  -6.4615  -0.2807  -0.1717   6.1077  22.0661
```

```r
hist(data$Y, breaks = 35, xlab = "Score de satisfaction", main = "")
```



La variable réponse présente une distribution globalement symétrique, ce qui rend plausible l'hypothèse d'erreurs gaussiennes retenue dans les modèles de régression linéaire. Les variables explicatives étant nombreuses et déjà normalisées, une exploration détaillée variable par variable ou par corrélation croisée n'apporte que peu d'information exploitable à ce stade. L'identification des variables influentes sera donc réalisée directement à l'aide des méthodes de régularisation et de sélection bayésiennes.

1.3 Vérification de la normalisation

En considérant un lot de variables explicatives, on remarque bien que leurs moyennes sont proches de 0 et que les écarts-types valent 1.

```r
apply(data[, 5:10], 2, mean)[1:5]
```

```
##        Film.3        Film.4        Film.5        Film.6        Film.7
##  2.965481e-16  1.892866e-15 -4.972515e-15 -4.536926e-15  8.228487e-16
```

```r
apply(data[, 5:10], 2, sd)[1:5]
```

```
## Film.3 Film.4 Film.5 Film.6 Film.7
##      1      1      1      1      1
```

Compte tenu de la structure du jeu de données et du caractère fortement dimensionné du problème, nous commençons l'analyse par une régression aléatoire de type RR-BLUP, qui constitue une référence bayésienne simple reposant sur un shrinkage global des coefficients.

1.4 Division du jeu de données

```
y <- data$Y
X <- as.matrix(data[, -which((names(data) == "Y") | (names(data) == "X") | names(data) == "sexe")

length(y)
```

```
## [1] 150
```

```
dim(X)
```

```
## [1] 150 160
```

```
n <- nrow(X)

ind_train <- sample(1:n, 100)
ind_test <- setdiff(1:n, ind_train)

X_train <- X[ind_train, ]
y_train <- y[ind_train]

X_test <- X[ind_test, ]
y_test <- y[ind_test]
```

2 Utilisation du package `rrBLUP`

2.1 Estimation

2.1.1 Rappels et objectif

La méthode RR-BLUP (Random Regression BLUP) est une approche bayésienne simple de la régression pénalisée. Elle est particulièrement adaptée aux situations où le nombre de variables explicatives est élevé par rapport au nombre d'observations ($p \leq n$). Dans ce cadre, les coefficients de régression sont considérés comme des effets aléatoires, soumis à un a priori gaussien centré, ce qui induit un effet de shrinkage global.

2.1.2 Modèle statistique

Le modèle s'écrit :

$$Y = \mu\mathbb{I} + X\beta + \varepsilon$$

avec

$$\beta \sim \mathcal{N}\left(0, \sigma_\beta^2 \mathbf{I}_p\right), \varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 \mathbf{I}_n\right)$$

Ce modèle est équivalent à une régression ridge bayésienne, où tous les coefficients sont pénalisés de la même manière.

2.1.3 Estimation des paramètres

Les paramètres du modèle sont estimés à partir du jeu d'apprentissage à l'aide d'un estimateur BLUP, qui correspond à la moyenne a posteriori des coefficients sous les hypothèses gaussiennes précédentes.

```
library(rrBLUP)

rr_blup <- mixed.solve(
  y = y_train,
  Z = X_train
)
```

```
beta_rr <- rr_blup$u
mu_rr <- as.numeric(rr_blup$beta)
```

```
beta_rr
```

```
mu_rr
```

```
## [1] -0.05906177
```

La sortie `beta_rr` correspond aux effets estimés des 160 chaînes. La moyenne globale du score de satisfaction est -0.07776509. Tous les coefficients sont non nuls, mais leur amplitude est fortement réduite par l'effet de shrinkage. La méthode RR-BLUP fournit une estimation des effets des chaînes en imposant un rétrécissement global des coefficients. Cette approche permet d'obtenir des estimations stables dans un contexte de grande dimension, mais ne réalise pas une sélection stricte des variables. Les coefficients estimés sont tous non nuls et de faible amplitude, ce qui reflète l'hypothèse a priori d'une variance commune pour l'ensemble des effets.

2.2 Prédiction

2.2.1 Rappels et objectif

Une fois les paramètres du modèle RR-BLUP estimés sur le jeu d'apprentissage, l'objectif est d'évaluer la capacité prédictive du modèle sur des données non utilisées lors de l'estimation. La qualité de la prédiction est mesurée par la corrélation entre les valeurs observées et les valeurs prédites sur le jeu de test. Cette mesure est couramment utilisée en régression pénalisée et en génomique pour évaluer la performance prédictive globale d'un modèle.

2.2.2 Formule de prédiction

Pour une observation du jeu de test, la valeur prédite est donnée par :

$$\hat{Y}_{test} = \hat{\mu} + X_{test}\hat{\beta}$$

avec $\hat{\mu}$, le score moyen estimé et $\hat{beta}$, le vecteur des coefficients estimés sur le jeu d'apprentissage.

```
y_pred_rr <- as.numeric(mu_rr + X_test %*% beta_rr)
```

```
cor_rr <- cor(y_pred_rr, y_test)
cor_rr
```

```
## [1] 0.8480824
```

La corrélation entre les scores observés et les scores prédits sur le jeu de test est de $r = 0.80757$. Cette valeur indique que le modèle RR-BLUP parvient à capturer une partie de la variabilité du score de satisfaction, tout en restant limité par le caractère fortement dimensionné du problème. Cette performance prédictive servira de référence pour la comparaison avec les méthodes bayésiennes plus flexibles étudiées par la suite.

2.3 Sélection des variables explicatives

2.3.1 Rappels et objectif

La méthode RR-BLUP ne réalise pas de sélection de variables au sens strict : tous les coefficients sont estimés et pénalisés de manière identique par l'a priori gaussien. Cependant, il est possible d'identifier les variables les plus influentes en examinant la distribution des coefficients estimés et en retenant ceux dont l'amplitude est la plus élevée. Cette sélection est donc heuristique et repose sur une analyse visuelle ou sur un seuil.

2.3.2 Principe de sélection

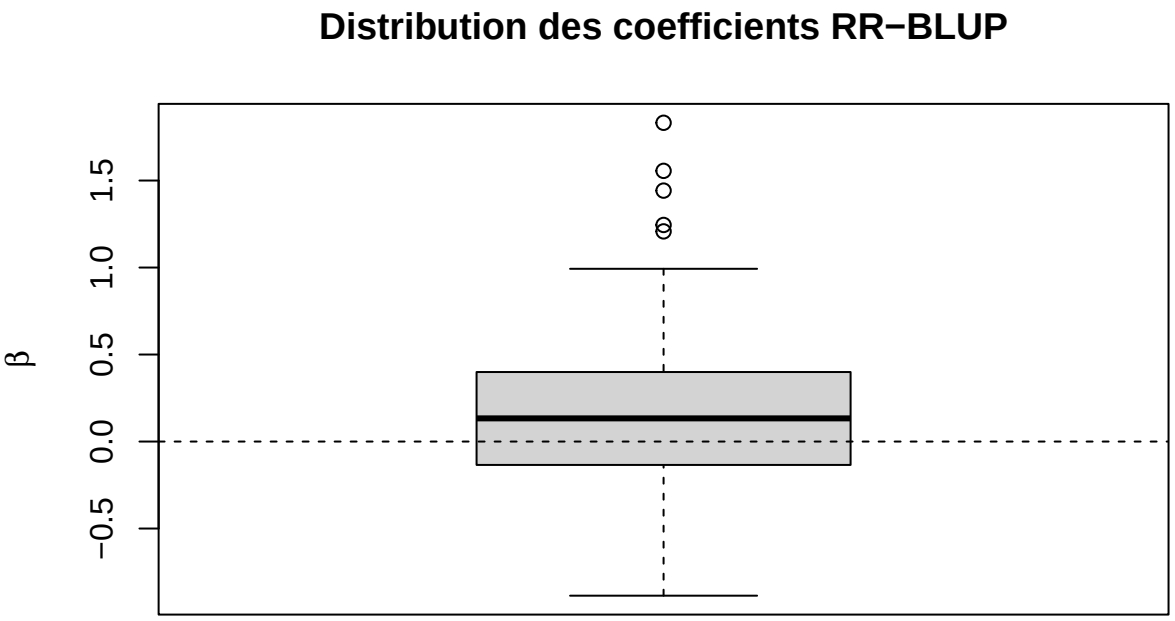
Deux approches sont possibles :

- Analyse visuelle à l'aide d'un boxplot des coefficients estimés ;
- Sélection par seuillage, en retenant les coefficients dont la valeur absolue dépasse un certain quantile élevé de la distribution.

Ces méthodes permettent d'identifier un petit nombre de variables susceptibles d'avoir un impact important sur le score de satisfaction

```
boxplot(beta_rr, main = "Distribution des coefficients RR-BLUP",
        ylab = expression(hat(beta)))

abline(h = 0, lty = 2)
```




```
seuil_rr <- quantile(abs(beta_rr), 0.95)
sel_rr <- which(abs(beta_rr) > seuil_rr)

length(sel_rr)
```

```
## [1] 8
```

```
sel_rr
```

```
##   Film.8  Film.10  Serie.8  Sport.1  Sport.2 Sport.10 Sport.15 Music.13
##        8       10       28       41       42       50       55      133
```

La méthode RR-BLUP ne permet pas une sélection franche des variables, mais l'examen de la distribution des coefficients estimés met en évidence un petit nombre de chaînes dont les effets estimés s'écartent davantage de zéro. Ces variables peuvent être considérées comme les plus influentes selon le modèle RR-BLUP, bien que cette sélection repose sur un critère heuristique et soit fortement dépendante de l'effet de shrinkage imposé par le modèle.

Les limites de la sélection obtenue par RR-BLUP motivent l'utilisation de modèles bayésiens plus flexibles, capables d'introduire un shrinkage différencié entre les variables. Nous étudions dans la section suivante le modèle Bayes A, qui repose sur une hiérarchie de variances spécifiques à chaque coefficient.

3 La régression bayésienne A

3.1 Différence entre le modèle Bayes A et le modèle Random Regression (RR-BLUP)

3.1.1 Rappels et objectif

Le modèle Bayes A est une extension hiérarchique du modèle de régression bayésienne utilisé dans le RR-BLUP. L'objectif principal est de relâcher l'hypothèse de variance commune imposée à l'ensemble des coefficients de régression dans le RR-BLUP, afin de permettre une pénalisation plus flexible et mieux adaptée aux données.

3.1.2 Modèles hiérarchiques comparés

RR-BLUP

$$\beta_j \sim \mathcal{N}\left(0, \sigma^2_\beta\right), \forall j$$

- Une seule variance pour tous les coefficients
- Shrinkage uniforme
- Tous les coefficients sont pénalisés de la même façon

Bayes A

$$\beta_j | \sigma^2_{\beta_j} \sim \mathcal{N}\left(0, \sigma^2_{\beta_j}\right); \sigma^2_{\beta_j} \sim Inverse - Gamma\,(a, b)$$

- Une variance spécifique par coefficient
- Shrinkage adaptatif
- Les coefficients importants peuvent être moins pénalisés

Marginalement, chaque coefficient suit une loi de Student, plus épaisse en queue qu'une loi gaussienne.

3.1.3 Avantage principal de Bayes A

- Meilleure détection des variables à fort effet
- Moins de sur-rétrécissement des coefficients réellement influents
- Modèle plus flexible que RR-BLUP

Contrairement au modèle RR-BLUP, qui impose une variance commune à l'ensemble des coefficients, le modèle Bayes A introduit une hiérarchie bayésienne permettant à chaque coefficient de disposer de sa propre variance. Cette structure conduit à un shrinkage adaptatif, favorisant la conservation des effets importants tout en pénalisant davantage les effets faibles. Le modèle Bayes A est ainsi plus flexible et potentiellement plus performant en présence de variables réellement influentes.