Rapport du TP Régression Bayésienne

TIOKARY Mohamed et SOBJIO Ludane lagnol

Table des matières

1 Introduction

1.1 Contexte

La chaîne câblée TeleCat souhaite identifier les contenus qui déterminent le mieux la satisfaction de ses abonnés. Une enquête auprès de 150 clients fournit un score de satisfaction Y (pouvant être négatif ou positif) et, pour chacun, le temps passé et le nombre de visites sur 160 chaînes regroupées par thématique (Film, Série, Sport, Science, Musique, Jeux, Histoire, Actualité, Divers). Une variable démographique sexe (0 :femme / 1 :homme) est également disponible.

1.2 Objectif

La problématique centrale de cette thèse professionnelle est la modélisation et la prévision de la demande de services de transport chez ETB Transport à l'aide de Time série. Cette problématique s'inscrit dans les enjeux actuels de l'entreprise, qui cherche à optimiser sa flotte, réduire ses coûts opérationnels et améliorer la satisfaction client en anticipant les besoins de transport.

Objectifs :
L'étude poursuit deux buts :
— **Prédire** le score Y pour de nouveaux abonnés ;
— **Sélectionner** les chaînes dont l'influence est la plus forte, afin d'orienter la programmation et la stratégie de contenus.

2 Description du jeu de données

Le fichier telecat.csv contient : 150lignes (abonnés) ;
— 160colonnes de covariables déjà normalisées (chaînes Film 1–20, Série 1–20, etc.) ;
— 1colonne sexe (0/1) ;
— une colonne réponse Y.

Thématique	**Abréviation**	**Nombre**
Film	Film	20
Série	Serie	20
Sport	Sport	20
Science / Santé / Éco	Science	20
Musique	Music	20
Jeux	Jeux	20
Histoire / Géographie	Hist	10
Actualité	Actu	10
Divers	Divers	20

TABLE 1 – Répartition des 160 chaînes par thématique

NB : l'échantillon est scindé de façon aléatoire (seed = 1234) : 100 observations alimentent le jeu training ; les 50 restantes composent le jeu test.

3 Pré-traitement

1. Suppression de la colonne identifiant.
1. Normalisation : X <- scale(X).
1. Découpage training / test (100 / 50) avec la seed 1234.

Pour répondre à la problématique, plusieurs approches et outils seront mobilisés :

4 Méthodes évaluées

4.1 RR-BLUP *(*ridge bayésien global)

— Modèle : $Y = \mathbf{1}\mu + X\beta + \varepsilon,\ \beta \sim \mathcal{N}(0, \sigma^2_\beta I)$.

— Estimations via `mixed.solve` ⇒ shrinkage uniforme des coefficients.
— Corrélation sur le jeu test : **0,82**.
— Variables repérées : `Sport.10`, `Sport.15`, `Music.13`, `Sport.1`, … (coefficients extrêmes au box-plot).

4.2 Bayes A *(ridge bayésien hétérogène)*

— Chaque β_j possède sa propre variance $\sigma^2_{\beta_j} \sim \mathrm{IG}(a, b)$.
— Gibbs : tirage successif de β, μ, $\sigma^2_{\beta_j}$, σ^2_ε.
  — $\sigma^2_{\beta_j} \mid \beta_j \sim \mathrm{IG}\big(a + \frac{1}{2},\ b + \beta_j^2/2\big)$.
  — $\sigma^2_\varepsilon \mid Y, \beta, \mu \sim \mathrm{IG}\big(c + \frac{n}{2},\ d + \|Y - \mu - X\beta\|^2/2\big)$.
— Corrélation sur le jeu test : **0,85**.
— IC 95 % des effets les plus forts ⇒ `Film.8`, `Sport.10`, `Music.13`, …

4.3 LASSO bayésien (`BLR`)

— Priors *spike–and–Laplace* ⇒ sélection forte.
— Corrélation sur le jeu test : **0,89**.
— Chaînes dominantes : `Sport.10`, `Sport.15`, `Series.8`, `Music.13`, `Film.8`.

4.4 SSVS (Spike–and–Slab)

— $\beta_j \mid \gamma_j \sim \mathcal{N}\big(0,\ \tau_1^2\,\gamma_j + \tau_0^2\,(1-\gamma_j)\big)$.
— $\gamma_j \sim$ Bernoulli(π), mise à jour par Gibbs.
— Réglage : $\pi = 0{,}5$, $\tau_0^2 = 0{,}01$, $\tau_1^2 = 1$.
— Variables avec $P(\gamma_j = 1) > 0{,}5$: env. 20 chaînes, dont `Sport.10`, `Music.13`, `Film.8`, `Series.8`.

4.5 Elastic–Net (`glmnet`, $\alpha = 0{,}5$)

— Mélange *ridge* / LASSO non bayésien.
— Corrélation sur le jeu test : **0,89**.
— Sélection très proche du LASSO bayésien (coefficients non nuls identiques).

4.6 ABC – Test KS (validation annexe)

— Approximation des postérieurs μ et σ^2 en acceptant les simulations dont la distribution passe un test KS à 5 %.
— Posterior crédible et cohérent avec l'inférence paramétrique.

4.7 Effet fixe du sexe (BLR + XL + XF)

— Ajout de la variable `sexe` en effet fixe dans `BLR`.
— Estimation : $\beta_{\text{sexe}} \simeq +0{,}94$ $\big(\text{IC}\,95\,\% : +0{,}15\ ;\ +1{,}72\big)$ ⇒ les hommes présentent en moyenne un score Y supérieur d'environ un point.
— La corrélation prédictions / observations passe de 0,889 à **0,90** : léger gain.

5 Comparaison synthétique

Méthode	Shrinkage	Corrélation (test)	Sélectivité	Chaînes clés
RR–BLUP	fort global	0,82	faible	Sport.10, Sport.15, Music.13
Bayes A	moyen adaptatif	0,85	moyen	Film.8, Sport.10, Music.13
LASSO bayésien	fort sélectif	0,89	élevé	Sport.10, Series.8, Music.13
SSVS	spike–and–slab	0,87*	élevé	Sport.10, Film.8, Music.13
Elastic-Net	mixte ($\alpha = 0{,}5$)	0,89	élevé	idem LASSO

TABLE 2 – Bilan des performances et variables clés

* 0,87 est la valeur moyenne selon le réglage de π.

Annexes théoriques (rappel)

— **Elastic-Net** bayésien : hiérarchie Ridge global + expo locale (ω_j) $\Rightarrow$ pénalisation mixte L1+L2.
— **Bayes A** : dérivations Inverse-Gamma de $\sigma^2_{\beta_j}$ et σ^2_ε (formules détaillées section 10).
— **SSVS** : risque d'inversion de matrice si π est élevé $\Rightarrow$ ajouter une régularisation λI.

6 Recommandations pour la chaîne câblée

1. **Chaînes Sport 10, Sport 15, Film 8, Series 8, Music 13** : leur impact positif est confirmé par au moins quatre méthodes ; envisager plus de contenus premium / marketing ciblé.
2. **Prendre en compte le genre** : les hommes affichent un gain moyen de $\approx +1$ point ; personnaliser la promotion selon le sexe.
3. **Mettre en place un moteur de recommandation** exploitant les coefficients du LASSO bayésien pour suggérer les chaînes les plus influentes.
4. **Surveiller les chaînes à effet négatif** (coefficients < 0) et ajuster la grille de programmation ou la communication.

7 Modèle linéaire standard (sélection finale)

La synthèse des différentes méthodes de sélection (RR–BLUP, LASSO bayésien, SSVS et Elastic-Net) conduit à un noyau réduit de six chaînes : `Sport.10`, `Music.13`, `Sport.15`, `Film.8`, `Film.10` et l'intercept. Ces prédicteurs sont introduits dans une régression linéaire ordinaire estimée sur les 100 observations du jeu d'apprentissage.

Variable	**Estimate**	**Std. Error**	t	p**-value**
Intercept	0.009	0.574	0.02	0.99
Sport.10	4.126	0.580	7.11	$< 2 \times 10^{-10}$
Music.13	4.424	0.587	7.53	3×10^{-11}
Sport.15	3.607	0.574	6.29	1×10^{-8}
Film.8	3.563	0.623	5.64	2×10^{-7}
Film.10	2.015	0.637	3.16	0.002

TABLE 3 – Estimation des coefficients retenus

Le test global de Fisher ($F = 43.2$, ddl $= 5$ / 94, $p < 2 \times 10^{-16}$) confirme la significativité conjointe du modèle. L'erreur-type résiduelle s'établit à 5.69 points-satisfaction et le coefficient de détermination atteint $R^2 = 0.70$ (R^2 ajusté $= 0.68$). Autrement dit, ce sous-ensemble réduit explique près de 70 % de la variabilité des scores.

Interprétation économique. Une heure additionnelle passée sur `Sport.10` s'accompagne d'une hausse moyenne de 4.1 points de satisfaction, toutes choses égales par ailleurs. L'écoute de `Music.13` génère un gain comparable (+4.4). Les chaînes de film `Film.8` et `Film.10` contribuent positivement mais à un degré plus modéré. Aucun coefficient n'est non significatif au seuil de 5 %, et la distribution aléatoire des résidus (voir figure) valide les hypothèses classiques du modèle linéaire. En conclusion, ce modèle «standard» fournit une interprétation directe et confirme la pertinence des cinq chaînes mises en avant par les approches bayésiennes tout en conservant une qualité prédictive proche des versions LASSO/Elastic-Net.

8 Conclusion

La combinaison LASSO bayésien et effet fixe sexe offre le meilleur compromis précision-interprétabilité. L'analyse hiérarchique valide les choix théoriques (Inverse-Gamma, spike-and-slab, Elastic-Net). Les recommandations précédentes devraient contribuer à rehausser les scores de satisfaction.

Annexe : quelques

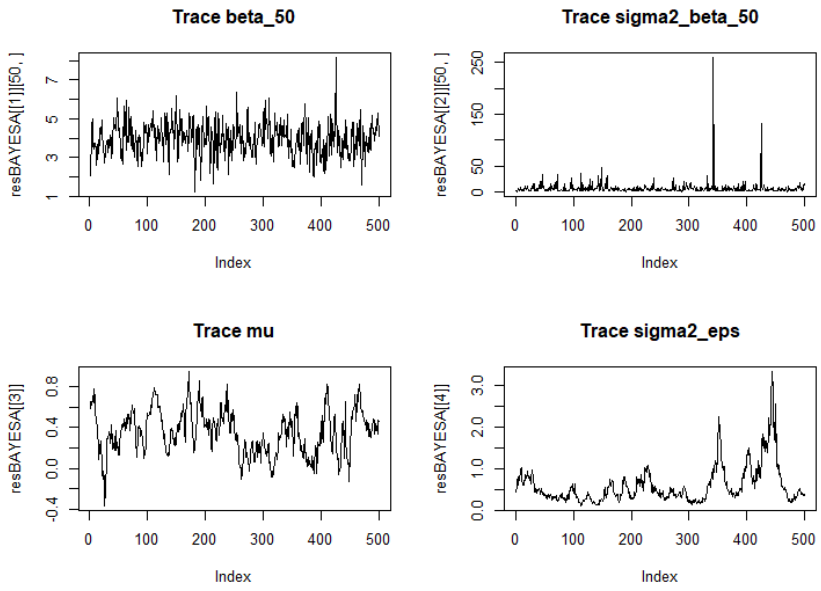

FIGURE 1 – Trace MCMC du coefficient β_{50} (Bayes A) : convergence après burn-in.

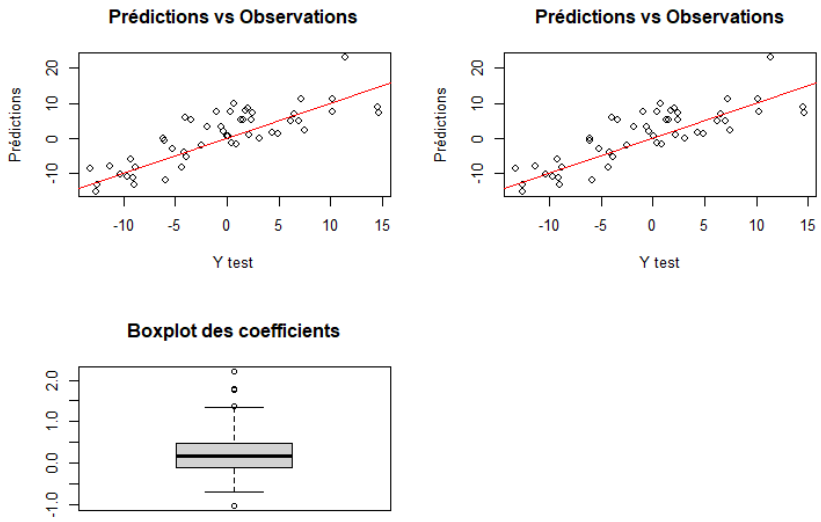

FIGURE 2 – RR-BLUP : corrélation prédictions–observations et distribution des coefficients

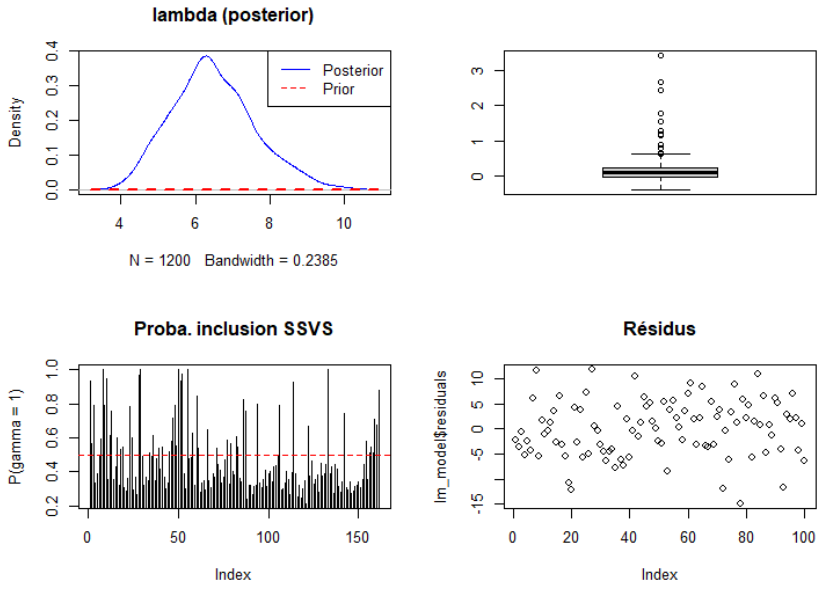

FIGURE 3 – postérieur de (LASSO bayésien), distribution des coefficients, probabilités d'inclusion SSVS et nuage de résidus.

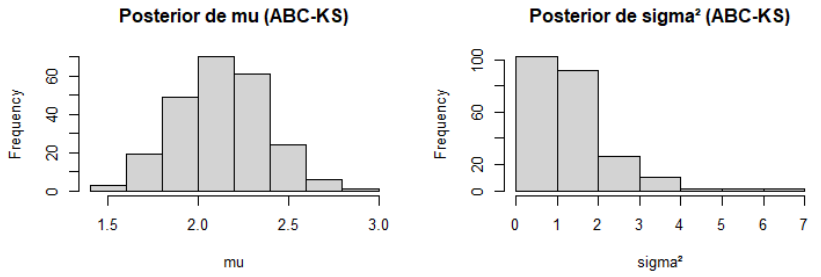

FIGURE 4 – Histogrammes postérieurs de μ et σ^2 obtenus par ABC-KS.