# Régression linéaire et sélection de variables : approches fréquentistes et bayésiennes

Liam Lajeunesse *

**Résumé**

Dans ce rapport, nous comparons les approches fréquentistes et bayésiennes d'estimation des coefficients de régression dans un modèle linéaire.

L'estimateur Ridge s'inscrit dans la famille des estimateurs à rétrécissement, dont l'intérêt est notamment motivé par le phénomène de James-Stein. Le Lasso, quant à lui, rend possible la sélection de variables conjointement à l'estimation des paramètres de la régression linéaire. Un intermédiaire entre ces deux approches fréquentistes est l'*Elastic Net* qui propose un compromis entre les deux estimateurs précédents. Il est notamment utile dans les contextes comportant plus de variables explicatives que d'observations. Les algorithmes LARS et LARS-EN permettent d'approximer le chemin des solutions du Lasso et de l'*Elastic Net*.

D'un point de vue bayésien, le Lasso peut être interprété comme le mode *a posteriori* de la distribution des paramètres de régression lorsque des lois *a priori* de Laplace indépendantes sont attribuées aux coefficients de régression. Un échantillon de la loi *a posteriori* est accessible à l'aide d'un échantillonneur de Gibbs en utilisant des normales conjuguées pour les coefficients du modèle linéaire ainsi que des lois exponentielles indépendantes pour les hyperparamètres de variance. Les lois *a priori* de type *Spike and Slab* sont aussi utiles afin de faire de la sélection de variables.

Les différents algorithmes et leur efficacité sont testés sur un jeu de données simulé ainsi que sur un jeu de données réel.

**Mots-clés** : Parcimonie, pénalisation, Ridge, Lasso, *Elastic Net*, algorithmes LARS et LARS-EN, échantillonneur de Gibbs, Lasso bayésien, loi *a priori Spike and Slab*

*Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, H3C 3J7, Canada. Email: liam.lajeunesse@umontreal.ca

# Table des matières

# 1 Introduction

Considérons un vecteur aléatoire suivant un modèle gaussien multivarié dont nous cherchons à estimer le vecteur de moyenne $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\intercal$. Formellement, nous avons

$$\mathbf{Y} = (Y_1, \ldots, Y_d)^\intercal \sim \mathcal{N}_d(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_{d\times d}) \tag{1}$$

avec $\sigma^2 = 1$. Il est facile de montrer que l'estimateur du maximum de vraisemblance de $\boldsymbol{\mu}$ est $\boldsymbol{\mu}_{EMV} = \mathbf{Y}$. Toutefois, Charles Stein démontra qu'en considérant certaines fonctions de perte $\mathcal{L} : \Theta \times \mathcal{A} \to [0, \infty)$, où $\Theta \subset \mathbb{R}^d$ est l'espace des paramètres et $\mathcal{A} \subset \mathbb{R}^d$ est l'espace décisionnel, cet estimateur présente des faiblesses. Spécifiquement, pour la fonction de perte quadratique définie par

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{a}) = \|\boldsymbol{\theta} - \mathbf{a}\|_2^2 = (\boldsymbol{\theta} - \mathbf{a})^\intercal(\boldsymbol{\theta} - \mathbf{a}) \ ,$$

où $\|\cdot\|_2$ est la norme euclidienne standard, cet estimateur est inadmissible lorsque $d \geq 3$ (Stein 1956). Rappelons qu'un estimateur $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ est dit inadmissible sous la perte quadratique s'il existe un autre estimateur $\tilde{\boldsymbol{\theta}}$ tel que pour tout $\boldsymbol{\theta} \in \mathbb{R}^d$

$$\mathbb{E}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \leq \mathbb{E}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2$$

et pour au moins un $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\mathbb{E}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 < \mathbb{E}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 \ .$$

En d'autres termes, un estimateur admissible n'est pas dominé uniformément sur $\Theta$.

Quelques années plus tard, James, Stein et al. (1961) proposèrent une forme explicite d'estimateur à rétrécissement dominant l'estimateur usuel sous la perte quadratique. Intuitivement, l'estimation d'un vecteur de paramètres est un point dans l'espace. Un estimateur à rétrécissement ramène ce point vers l'origine, ce qui produit un effet de *shrinkage* sur les coefficients du vecteur. Nous parlons ainsi d'estimateur à rétrécissement introduisant un biais, puisqu'il s'éloigne de la solution usuelle, mais qui possède de meilleures propriétés que l'estimateur du maximum de vraisemblance habituel. Le message central de ce résultat est qu'un estimateur biaisé peut parfois être préférable à un estimateur non biaisé lorsque le critère d'évaluation est l'erreur quadratique moyenne. Cette idée est au cœur des méthodes de pénalisation en régression, qui consistent à ajouter un terme de pénalisation au critère d'optimisation que l'on souhaite minimiser. Par exemple, nous pouvons ajouter une contrainte sur la norme de l'estimateur afin de rétrécir l'amplitude de ses coefficients. Dans un tel cas, même si l'estimateur minimise la fonction de perte, il peut être pénalisé par l'amplitude de sa norme, qui peut être considérée comme trop grande.

Parallèlement, dans le cadre de la statistique, un objectif courant consiste à tenter d'expliquer les relations entre plusieurs variables. En général, nous cherchons à déceler des structures dans les données afin d'expliquer des phénomènes et guider des prises de décisions

à la lumière de ces connaissances. Pour y arriver, les statisticiens analysent la relation entre des covariables et une variable réponse. Une méthode courante d'apprentissage supervisé est la régression linéaire. Formellement, nous considérons le modèle de régression linéaire

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \ , \tag{2}$$

où $\mathbf{y}$ est le vecteur réponse de dimension $n \times 1$, $\alpha$ est une constante interprétée comme l'ordonnée à l'origine, $\mathbf{X}$ est la matrice de design des variables explicatives de dimension $n \times p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ est le vecteur des coefficients de régression et $\boldsymbol{\epsilon}$ est un vecteur d'erreurs normales, indépendantes et identiquement distribuées, de taille $n \times 1$, d'espérance nulle et de variance $\sigma^2$. Dans ce contexte, les observations contenues dans la matrice de design $\mathbf{X}$ et les réponses de la variable d'intérêt contenues dans $\mathbf{y}$ sont reliées par la structure (2). Le premier problème qui se pose est donc l'estimation du paramètre d'intérêt $\boldsymbol{\beta}$ puisque dans un contexte supervisé, les observations et les réponses sont connues. En fait, nous cherchons des manières d'estimer le vecteur $\boldsymbol{\beta}$ ainsi que d'identifier quels coefficients sont réellement importants afin d'expliquer $\mathbf{y}$. Dans un jeu de données quelconque, plusieurs covariables peuvent s'avérer inutiles dans l'explication de la variable d'intérêt.

Il est facile de constater le lien entre le modèle de régression linéaire et le modèle énoncé en (1). Pour l'étude des estimateurs de $\boldsymbol{\beta}$, l'ordonnée à l'origine peut être traitée séparément. Nous supposons donc, dans les développements qui suivent, que les colonnes de la matrice de design $\mathbf{X}$ sont centrées et standardisées et que la réponse a été centrée. Afin d'alléger la notation, nous continuons de noter $\mathbf{y}$ la réponse centrée et travaillons avec le modèle

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \ . \tag{3}$$

Ce modèle sans ordonnée à l'origine permet de mettre en évidence le parallèle avec (1). Désormais, $\mathbf{y}$ suit une loi normale multidimensionnelle d'espérance $\mathbf{X}\boldsymbol{\beta}$. Ce parallèle suggère qu'en régression aussi, il peut être avantageux de considérer des estimateurs biaisés, mais plus stables, plutôt que de se limiter à l'estimateur usuel des moindres carrés. Face à ce problème, les statisticiens introduisirent des formes plus générales d'estimateur à rétrécissement à l'aide de fonctions de perte pénalisées. Ces fonctions modifient les problèmes d'optimisation en utilisant des versions altérées des fonctions de perte usuelles ; elles peuvent comporter un terme supplémentaire, par exemple. Une forme commune de fonction de perte pénalisée est celle des estimateurs Bridge, introduite par Frank et Friedman (1993) et menant à des estimations définies comme

$$\hat{\boldsymbol{\beta}}_\gamma = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \, \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{i=1}^{p} |\beta_i|^\gamma \ , \tag{4}$$

où $\lambda$ est un paramètre de pénalisation. Ainsi, même si un certain candidat $\boldsymbol{\beta}$ minimise le premier terme, il peut être pénalisé par sa taille selon une pénalité de la forme $\sum_i |\beta_i|^\gamma$. L'estimateur Ridge, qui est la solution du cas particulier où $\gamma = 2$ dans (4), fut notamment formalisé par Hoerl et Kennard (1970).

Dans la section 2, nous explorons des méthodes d'inférence fréquentistes. L'estimateur Ridge est présenté comme une méthode de rétrécissement du vecteur de coefficients $\boldsymbol{\beta}$. Bien que cette méthode réponde à un premier objectif consistant à bien estimer $\boldsymbol{\beta}$, cette technique peut créer des modèles peu interprétables en incluant des estimations pour chacun des paramètres de régression. Dans ce contexte, l'estimateur Lasso est également introduit, permettant d'accomplir simultanément de bonnes approximations des coefficients de régression ainsi que de la sélection de variables. Ensuite, l'*Elastic Net* est mentionné comme un bon compromis entre le Ridge et le Lasso. Finalement, les techniques d'estimation sont testées sur un jeu de données simulé, ainsi qu'un jeu de données réel. Pour y arriver, l'algorithme LARS ainsi qu'une de ses variantes LARS-EN sont utilisés pour approximer les solutions du Lasso et de l'*Elastic Net*. L'estimateur Ridge, quant à lui, possède une forme explicite et ne nécessite pas l'utilisation d'un algorithme complexe pour approximer sa solution.

La section 3 porte sur des approches bayésiennes d'estimation du paramètre $\boldsymbol{\beta}$ permettant de traiter celui-ci comme une variable aléatoire. Dans un premier temps, une interprétation bayésienne du Lasso est présentée en posant des lois *a priori* de Laplace indépendantes sur les coefficients de régression. Contrairement au Lasso classique, cette méthode ne rend pas possible la sélection de variables lorsque nous observons les moyennes ou les médianes *a posteriori* des paramètres. Les lois *a priori* de type *Spike and Slab* se présentent ensuite comme des alternatives bayésiennes efficaces, permettant d'introduire de la parcimonie dans l'approximation du vecteur $\boldsymbol{\beta}$. Dans la partie bayésienne, un échantillonneur de Gibbs est ensuite utilisé afin d'obtenir un échantillon de la loi *a posteriori* de $\boldsymbol{\beta}$ à partir d'un jeu de données simulé et d'un jeu de données réel pour les différentes méthodes.

## 2 Méthodes fréquentistes

Dans un premier temps, nous nous positionnons dans un contexte fréquentiste. Conséquemment, nous considérons que les données prises en compte proviennent d'une certaine population et que les modèles éventuels auxquels nous nous intéresserons ont des paramètres fixes et inconnus. Dans un contexte de régression linéaire, on suppose ainsi que le vecteur $\boldsymbol{\beta}$ est fixe, mais inconnu et nous cherchons à l'estimer le mieux possible à travers un sous-ensemble de données provenant de la population d'intérêt. Une piste de solution est de minimiser l'erreur quadratique en introduisant l'estimateur des moindres carrés (MCO ou OLS pour *Ordinary Least Squares*)

$$\hat{\boldsymbol{\beta}}_{OLS} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \ .$$

Ce critère d'optimisation se résout facilement en cherchant le minimum de la norme euclidienne sur l'ensemble $\mathbb{R}^p$ si $\mathbf{X}$ est de plein rang, donc si $\mathbf{X}^\intercal\mathbf{X}$ est inversible. Sous les hypothèses de Gauss-Markov sur le terme d'erreur, la solution à ce critère d'optimisation est BLUE (*Best Linear Unbiased Estimator*) (Montgomery, Peck et Vining 2021). En

optimisant, la solution obtenue est

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y} \ .$$

Étant donné que cet estimateur est souvent sous-optimal et sensible à la colinéarité, à la lumière des résultats de James-Stein, nous considérons maintenant des estimateurs biaisés issus de fonctions de perte pénalisées comme Ridge ou Lasso.

## 2.1 Estimateur Ridge

Considérons le modèle (3) avec le terme d'erreur $\boldsymbol{\epsilon}$ respectant les conditions de Gauss-Markov. La procédure d'estimation par les moindres carrés est bonne dans le cas où $\mathbf{X}^\intercal \mathbf{X}$ est quasiment une matrice unitaire et que la matrice de design $\mathbf{X}$ est de rang plein. Plus généralement, lorsque $\mathbf{X}^\intercal \mathbf{X}$ est mal conditionnée, notamment en présence d'une forte colinéarité entre les prédicteurs, l'estimation par les moindres carrés peut devenir instable (Hoerl et Kennard 1970). L'estimateur Ridge cherche alors à stabiliser la matrice $\mathbf{X}^\intercal \mathbf{X}$ pour la rendre inversible dans le cas où $\mathbf{X}$ n'est pas de rang plein, ou encore, améliorer la gestion de la colinéarité. Notons que, contrairement à l'estimateur des moindres carrés usuels, nous n'avons pas besoin de postuler que la matrice de design est de rang plein. Nous cherchons alors une manière de pénaliser la perte quadratique initialement considérée afin d'améliorer les propriétés de l'estimateur de base en diminuant sa variance par exemple.

Soit $\hat{\boldsymbol{\mu}}$, un estimateur de la moyenne dans (1) que l'on définit comme l'argument qui minimise le risque empirique pour l'erreur quadratique, donc

$$\hat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\arg\min} \|\mathbf{y} - \boldsymbol{\mu}\|_2^2 \ .$$

Afin de créer un effet de rétrécissement sur l'amplitude des coefficients estimés, nous pouvons modifier le critère de minimisation précédent en introduisant une contrainte sur la norme du paramètre d'intérêt :

$$\tilde{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu} \in \mathbb{R}^d : \|\boldsymbol{\mu}\|_2 \leq \delta}{\arg\min} \|\mathbf{y} - \boldsymbol{\mu}\|_2^2 \ , \tag{5}$$

pour un certain $\delta > 0$. Tel que mentionné précédemment, $\boldsymbol{\mu}$ est un point dans $\mathbb{R}^d$. Or, avec la nouvelle contrainte, nous contraignons son estimation à appartenir à $\overline{\mathcal{B}(0, \delta)}$, une boule fermée, centrée à l'origine et de rayon $\delta$. Nous remarquons que si $\delta \to \infty$, nous obtenons essentiellement une boule recouvrant l'ensemble des solutions possibles et nous retrouvons l'estimateur non contraint $\hat{\boldsymbol{\mu}}$, qui respectera nécessairement la contrainte. Dans le cas contraire, si $\delta \to 0$, $\overline{\mathcal{B}(0, \delta)}$ se réduit à l'origine dans la limite et nous obtenons une estimation nulle du paramètre. De manière concrète, $\tilde{\boldsymbol{\mu}}$ est la projection de $\hat{\boldsymbol{\mu}}$ sur $\overline{\mathcal{B}(0, \delta)}$. Ainsi, si $\hat{\boldsymbol{\mu}} \notin \overline{\mathcal{B}(0, \delta)}$, ou en d'autres termes $\|\hat{\boldsymbol{\mu}}\|_2 > \delta$, l'estimateur $\tilde{\boldsymbol{\mu}}$ sera un point sur la surface de $\overline{\mathcal{B}(0, \delta)}$. La figure 1 présente un exemple en trois dimensions où $\boldsymbol{\mu}$ a trois

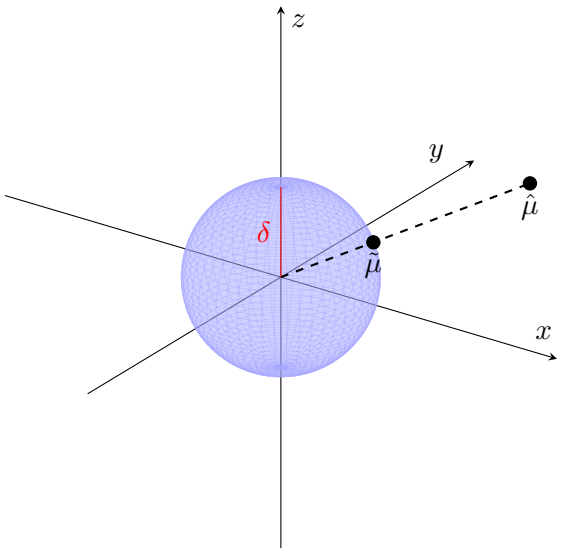


FIGURE 1 – Représentation de la contrainte sur la norme euclidienne du vecteur $\hat{\boldsymbol{\mu}}$ dans $\mathbb{R}^3$.
Dans cet exemple, $\hat{\boldsymbol{\mu}} = (2, \frac{3}{2}, 1)$ et $\tilde{\boldsymbol{\mu}}$ est sa projection sur $\overline{\mathcal{B}(0,\delta)}$, où $\delta = 1$. Notons que cette figure donne seulement une intuition géométrique de pénalisation. En réalité, l'estimateur Ridge impose une pénalité sur $\boldsymbol{\beta}$ et pas directement sur $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Si $\mathbf{X}$ n'est pas orthonormale, la boule dans l'espace des coefficients $\mathbb{R}^p$ est transformée par $\mathbf{X}$, de sorte que sa géométrie dans l'espace des prédictions $\mathbb{R}^n$ est plus complexe.

composantes et est limité par une boule de rayon $\delta = 1$. Nous pouvons facilement voir l'effet de *shrinkage* produit par la contrainte.

Dans un contexte de régression, nous cherchons essentiellement à estimer $\boldsymbol{\beta}$ dans $\mathbb{R}^p$ de manière similaire à la figure 1. Aussi, nous pouvons réécrire le problème d'optimisation (5). En combinant ces deux points, nous définissons l'estimateur Ridge comme

$$\hat{\boldsymbol{\beta}}_R^{\lambda} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \ , \tag{6}$$

où $\lambda > 0$ est appelé paramètre de pénalisation. Il est important de noter qu'il existe un lien intuitif entre $\lambda$ et $\delta$, le rayon de la boule. En effet, si $\lambda$ augmente, nous augmentons l'importance de la pénalité induite par la norme de $\boldsymbol{\beta}$. Dans un tel cas, le problème d'optimisation fournira une solution plus proche de l'origine. À l'inverse, si $\lambda$ diminue, alors moins d'importance est accordée à la norme du vecteur $\boldsymbol{\beta}$, ce qui équivaut à augmenter $\delta$ dans le problème d'optimisation (5). En d'autres termes, faire tendre $\lambda$ vers 0 dans (6) est équivalent à $\delta \to \infty$ dans (5).

Remarquons que la fonction à optimiser en (6) est différentiable. Ainsi, nous pouvons facilement trouver sa solution en différentiant et en trouvant son minimum. Posons $\mathcal{S}(\boldsymbol{\beta}) =$

$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$, alors :

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \mathcal{S}(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} \Big[ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \Big] \\
&= \nabla_{\boldsymbol{\beta}} \Big[ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\intercal \boldsymbol{\beta} \Big] \\
&= \nabla_{\boldsymbol{\beta}} \Big[ \mathbf{y}^\intercal \mathbf{y} - \mathbf{y}^\intercal \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{y} + \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta} + \lambda \boldsymbol{\beta}^\intercal \boldsymbol{\beta} \Big] \ .
\end{aligned}$$

Or, $\mathbf{y}^\intercal \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{y}$ puisque ces produits matriciels donnent des scalaires. Ainsi,

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \mathcal{S}(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} \Big[ \mathbf{y}^\intercal \mathbf{y} - 2\mathbf{y}^\intercal \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta} + \lambda \boldsymbol{\beta}^\intercal \boldsymbol{\beta} \Big] \\
&= -2\mathbf{X}^\intercal \mathbf{y} + 2\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} \ .
\end{aligned}$$

Puis, en égalant le gradient à 0 pour trouver les points critiques, nous obtenons :

$$\begin{aligned}
& 0 = -2\mathbf{X}^\intercal \mathbf{y} + 2\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} \\
\Leftrightarrow\ & (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})\boldsymbol{\beta} = \mathbf{X}^\intercal \mathbf{y} \\
\Leftrightarrow\ & \boldsymbol{\beta} = (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\intercal \mathbf{y} \ .
\end{aligned} \tag{7}$$

Notons que $(\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I}) \succ 0$, puisque $\mathbf{X}^\intercal \mathbf{X} \succeq 0$ et $\lambda \mathbf{I} \succ 0$ [1]. Ainsi, ce terme est bien inversible. En revanche, nous devons aussi vérifier que cette solution est bien un minimum. En prenant la dérivée seconde :

$$\begin{aligned}
\nabla^2_{\boldsymbol{\beta}} \mathcal{S}(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} \Big[ -2\mathbf{X}^\intercal \mathbf{y} + 2\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} \Big] \\
&= 2\mathbf{X}^\intercal \mathbf{X} + 2\lambda \mathbf{I}_{p\times p} \succ 0 \ .
\end{aligned}$$

Ainsi, puisque la matrice hessienne est définie positive, nous avons bien que la solution en (7) est un minimum. Donc, nous pouvons réécrire l'estimateur Ridge sous une forme explicite comme

$$\hat{\boldsymbol{\beta}}_R^\lambda = (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I}_{p\times p})^{-1} \mathbf{X}^\intercal \mathbf{y} \ .$$

Bien que l'estimateur Ridge possède une forme explicite, il peut être plus complexe de choisir la valeur de $\lambda$. Dans notre contexte d'apprentissage supervisé, le vecteur $\mathbf{y}$ ainsi que la matrice de design $\mathbf{X}$ sont connus. Alors, il reste seulement à trouver une manière de sélectionner la meilleure valeur de $\lambda$ à partir de ces informations, c'est-à-dire celle qui mène au meilleur modèle.

1. Les symboles $\succ 0$ et $\succeq 0$ signifient que la matrice de gauche est définie positive et semi-définie positive respectivement.

### 2.1.1 Validation croisée à $k$ blocs et validation croisée généralisée

Une méthode courante pour déterminer une valeur de $\lambda$ est d'utiliser une procédure de validation croisée à $k$ blocs (*k-fold cross-validation*). Cette technique consiste à former une grille de valeurs de $\lambda$, puis à calculer l'erreur de prédiction moyenne pour chacun de ces candidats. La valeur de $\lambda$ retenue est alors celle qui est associée à la plus petite erreur de prédiction moyenne. Spécifiquement, les données sont découpées en $k$ blocs desquels nous choisissons un groupe à laisser de côté. Pour une valeur de $\lambda$ fixée, un modèle est ensuite ajusté en utilisant les $k-1$ blocs restants. Ensuite, l'erreur de prédiction est calculée à l'aide du bloc laissé de côté. L'opération est répétée jusqu'à ce que chacun des $k$ blocs ait été mis de côté exactement une fois. La moyenne des erreurs de prédiction est calculée pour la valeur de $\lambda$ en question. Cette moyenne est ainsi considérée comme un score attribué au candidat $\lambda$ de la grille. Après avoir répété les étapes pour chaque valeur de $\lambda$ dans la grille, nous sélectionnons la valeur de $\lambda$ menant à la plus petite erreur moyenne de prédiction parmi les candidats (Frank et Friedman 1993). Ensuite, un modèle final est ajusté avec cette valeur $\lambda$ sur l'ensemble des données. Bien qu'assez simple conceptuellement, cette méthodologie peut être très coûteuse numériquement. Cette méthode est toutefois utilisée avec dix blocs dans la section 2.4.2.

Une alternative moins coûteuse, particulièrement adaptée aux estimateurs linéaires en $\mathbf{y}$ comme Ridge, est la validation croisée généralisée (*generalized cross-validation*, GCV) proposée par Golub, Heath et Wahba (1979). Cette méthode peut être vue comme une approximation efficace d'une erreur de prédiction de type *leave-one-out*, sans devoir réajuster le modèle en retirant successivement chaque observation. Intuitivement, cette technique approxime une mesure d'erreur de prédiction hors-échantillon pour un candidat $\lambda$ provenant d'une grille de candidats. Sous cette méthode, Golub, Heath et Wahba (1979) ont démontré que l'estimateur $\hat{\lambda}$ défini comme le minimiseur de

$$\mathcal{V}(\lambda) = \frac{\frac{1}{n}\left\|(\boldsymbol{I} - \boldsymbol{A}(\lambda))\mathbf{y}\right\|_2^2}{\left[\frac{1}{n}\operatorname{Tr}(\boldsymbol{I} - \boldsymbol{A}(\lambda))\right]^2},$$

où $\boldsymbol{A}(\lambda) = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal$, est une bonne variante de l'estimateur PRESS de Allen (1974). L'idée est de rechercher un compromis entre un bon ajustement et une complexité raisonnable. Si $\lambda$ est trop petit, alors le modèle est trop flexible et l'erreur d'entraînement sera faible, mais la trace de $\mathbf{A}(\lambda)$ sera grande, ce qui diminue le dénominateur et fait augmenter $\mathcal{V}(\lambda)$. Si $\lambda$ est trop grand, le modèle devient trop contraint et l'erreur d'ajustement au numérateur augmente et le score pour le candidat aussi. Cette méthodologie est utilisée dans la section 2.4.1.

En résumé, l'estimateur Ridge est une bonne alternative à l'estimateur MCO usuel

puisqu'il réduit notamment la variance de celui-ci ou encore l'erreur de prédiction [2]. C'est aussi un estimateur utile lorsque l'on souhaite conserver l'ensemble des prédicteurs dans le modèle plutôt que d'effectuer une sélection explicite. En revanche, cet avantage peut aussi s'avérer être un défaut lorsque les données explicatives ne présentent pas de lien intuitif et direct avec la variable réponse. L'estimation par la méthode Ridge a tendance à produire des modèles peu interprétables et elle ne règle que l'un des deux problèmes précédemment mentionnés dans le cadre de la régression linéaire. L'estimateur Ridge permet de bien approximer $\boldsymbol{\beta}$ avec un certain biais, mais ne permet pas de déterminer facilement lesquels de ses coefficients sont réellement utiles dans le modèle. Face à cette difficulté, l'estimateur Lasso sera présenté dans la prochaine section.

## 2.2 Estimateur Lasso

L'estimateur Lasso, pour *Least Absolute Shrinkage and Selection Operator*, est une méthode de rétrécissement combinant l'estimation des coefficients de régression et la sélection de variables. Comme l'estimateur Ridge, il est obtenu en minimisant une perte quadratique à laquelle nous ajoutons une pénalité sur la taille des coefficients. Toutefois, contrairement à Ridge, qui utilise une pénalité sur la norme euclidienne du vecteur, le Lasso utilise la norme $\ell_1$ afin de limiter l'estimation à une région spécifique dans l'espace.

Considérons d'abord le contexte (1) afin d'avoir une intuition géométrique de la contrainte. Soit $\hat{\boldsymbol{\mu}} \in \mathbb{R}^d$, un estimateur non contraint de $\boldsymbol{\mu}$. Nous pouvons définir son analogue contraint par la norme $\ell_1$ comme

$$\tilde{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu}\in\mathbb{R}^d:\|\boldsymbol{\mu}\|_1\leq\delta}{\arg\min} \|\mathbf{y}-\boldsymbol{\mu}\|_2^2 \ ,$$

pour un certain $\delta > 0$. Autrement dit, $\tilde{\boldsymbol{\mu}}$ est la projection euclidienne de $\hat{\boldsymbol{\mu}}$ sur la boule de norme $\ell_1$ de rayon $\delta$. La particularité de cette contrainte est que la boule $\ell_1$ possède des sommets et des arêtes alignés avec les axes dans $\mathbb{R}^d$. Ainsi, la projection peut facilement se situer sur une région où une ou plusieurs composantes sont exactement nulles. Cette intuition géométrique explique en partie pourquoi le Lasso produit des solutions parcimonieuses. La figure 2 offre une représentation intuitive de la contrainte dans $\mathbb{R}^3$.

Dans le cas de la régression linéaire, l'objectif n'est plus de contraindre directement la moyenne $\boldsymbol{\mu}$, mais plutôt le vecteur de coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. L'estimateur Lasso, formalisé par R. Tibshirani (1996), est alors défini par

$$\hat{\boldsymbol{\beta}}_L^{\lambda} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p}|\beta_j| \ , \tag{8}$$

2. Notons que dans cette section, les conditions de Gauss-Markov n'ont pas été nécessaires afin de trouver une solution explicite pour l'estimateur Ridge. Ces conditions concernent plutôt les énoncés sur l'estimateur MCO. En revanche, elles sont utiles afin de trouver une forme explicite pour la variance de l'estimateur Ridge et montrer qu'elle est plus petite que celle de l'estimateur MCO en général.

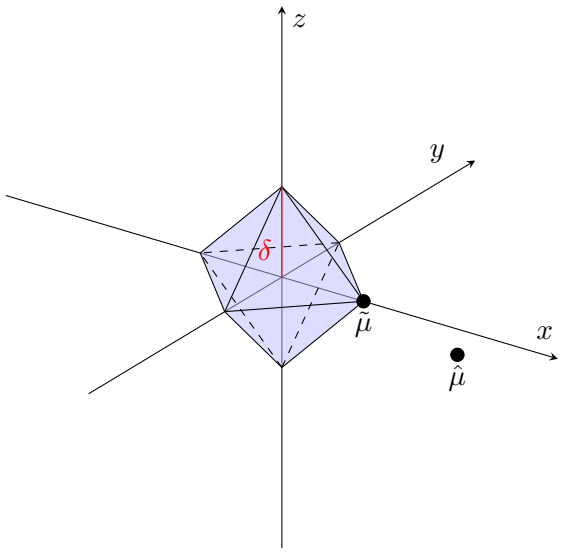


FIGURE 2 – Représentation de la contrainte sur la norme $\ell_1$ du vecteur $\hat{\boldsymbol{\mu}}$ dans $\mathbb{R}^3$.
Dans cet exemple, $\hat{\boldsymbol{\mu}} = (2.5, -0.5, 0)$ et $\tilde{\boldsymbol{\mu}}$ est sa projection sur $\overline{\mathcal{B}(0,\delta)}$ sous la norme $\ell_1$, où $\delta = 1$. Notons que cette figure donne seulement une intuition géométrique de pénalisation $\ell_1$ comme dans la discussion de la figure 1. Remarquons aussi que malgré la composante non nulle en $y$ de $\hat{\boldsymbol{\mu}}$, le Lasso fournit une solution nulle en $y$. La géométrie de l'espace sous la norme $\ell_1$ admet des solutions introduisant de la parcimonie dans le vecteur. Une explication simple est que la procédure juge que la taille du vecteur en $y$ est assez petite relativement à celle en $x$ et la méthode ramène $y$ exactement à 0.

où $\lambda > 0$ est un paramètre de pénalisation [3]. Similairement à l'estimateur Ridge, lorsque $\lambda$ augmente, la pénalité devient plus importante et les coefficients sont davantage contractés vers 0. De plus, pour des valeurs suffisamment grandes de $\lambda$, certains coefficients peuvent devenir exactement nuls, ce qui permet d'interpréter le Lasso comme une méthode de sélection de variables. La figure 2 représente géométriquement la contrainte de la boule $\ell_1$ dans $\mathbb{R}^3$.

### 2.2.1 Cas orthonormal

Afin de mieux saisir pourquoi le Lasso permet de donner des solutions parcimonieuses, commençons par considérer le cas où la matrice de design $\mathbf{X}$ est orthonormale, donc elle est constituée de vecteurs unitaires, deux à deux orthogonaux. Dans ce cas, nous avons la relation $\mathbf{X}^\intercal\mathbf{X} = \mathbf{I}_{p\times p}$. Parallèlement, notons $\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ l'estimateur MCO de $\boldsymbol{\beta}$.

3. Notons que le terme $\frac{1}{2}$ devant la fonction de perte quadratique est seulement une convention pour définir la fonction de perte pénalisée. Lorsque nous prenons la dérivée, le terme à l'exposant vient s'annuler avec celui-ci. La différence de la constante précédant la fonction de perte quadratique est en quelque sorte prise en compte dans la sélection du paramètre de pénalisation $\lambda$.

Sous cette condition, nous avons que

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 &= \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\boldsymbol{\beta} \\ &= \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\hat{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}^\intercal\boldsymbol{\beta} \ ;\end{aligned}$$

puis, en complétant le carré,

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 &= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^\intercal(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0) + \mathbf{y}^\intercal\mathbf{y} - \hat{\boldsymbol{\beta}}_0^\intercal\hat{\boldsymbol{\beta}}_0 \\ &= \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0\|_2^2 + \mathbf{y}^\intercal\mathbf{y} - \hat{\boldsymbol{\beta}}_0^\intercal\hat{\boldsymbol{\beta}}_0 \ .\end{aligned}$$

On remarque alors que minimiser $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ revient à minimiser $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0\|_2^2$ à une constante près, qui ne dépend pas de $\boldsymbol{\beta}$. Dans le cas orthonormal, les colonnes de $\mathbf{X}$ ne sont pas colinéaires. Ainsi, ajuster un certain coefficient du vecteur $\boldsymbol{\beta}$ n'influence pas directement l'ajustement d'une autre de ses entrées. Nous pouvons réécrire le critère d'optimisation sur la perte quadratique comme

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{j=1}^{p} (\beta_j - \hat{\beta}_j^0)^2 \ .$$

Ainsi, pour chacun des coefficients de $\boldsymbol{\beta}$, nous pouvons élaborer un critère d'optimisation en introduisant le paramètre de pénalisation $\lambda > 0$, puisque le critère d'optimisation est séparable selon les coordonnées. En d'autres termes, le critère d'optimisation peut s'écrire sous une forme de somme que nous pouvons minimiser terme par terme. La prochaine étape est de minimiser la fonction

$$f(\beta_j) = \frac{1}{2}(\beta_j - \hat{\beta}_j^0)^2 + \lambda|\beta_j| \ . \tag{9}$$

Remarquons que la fonction est convexe car elle est la somme d'une fonction quadratique et d'une fonction de valeur absolue qui sont toutes deux convexes. Dans ce cas, tout minimum local est un minimum global de la fonction.

Dans ce contexte, trois cas se distinguent pour finalement trouver une forme explicite de l'estimateur. Afin de trouver le minimum de la fonction (9), nous commençons par résoudre le problème lorsque nous considérons $\beta_j > 0$. Dans un tel cas, $|\beta_j| = \beta_j$ et nous devons minimiser la fonction

$$f(\beta_j) = \frac{1}{2}(\beta_j - \hat{\beta}_j^0)^2 + \lambda\beta_j \ .$$

En dérivant et en égalant à 0, nous trouvons $\hat{\beta}_j^L = \hat{\beta}_j^0 - \lambda$, qui est valide seulement si $\hat{\beta}_j^0 > \lambda$ car nous considérons le cas $\beta_j > 0$. Un critère sur la dérivée seconde permet de voir que c'est bien un minimum. Autrement dit, si le coefficient MCO est assez positif, l'estimation par la technique du Lasso le réduit de $\lambda$ et le rapproche de 0. De façon similaire, si $\beta_j < 0$, alors $|\beta_j| = -\beta_j$ et nous trouvons la solution $\hat{\beta}_j^L = \hat{\beta}_j^0 + \lambda$, qui est valide seulement si

$\hat{\beta}_j^0 < -\lambda$ puisque nous avons postulé que $\beta_j < 0$. En d'autres mots, si le coefficient MCO est assez négatif, le Lasso l'augmente de $\lambda$ pour le rapprocher de 0.

Finalement, il reste à vérifier le cas où $|\hat{\beta}_j^0| \leq \lambda$. Les deux solutions trouvées pour les cas $\beta_j > 0$ et $\beta_j < 0$ ne fonctionnent pas. En effet, pour $\beta_j > 0$, le point critique serait $\hat{\beta}_j^L = \hat{\beta}_j^0 - \lambda$. Or, dans cette situation, $\hat{\beta}_j^0 - \lambda \leq 0$, qui n'est pas dans la région $\beta_j > 0$. De manière similaire, si $\beta_j < 0$, le point critique serait $\hat{\beta}_j^L = \hat{\beta}_j^0 + \lambda$ qui n'est pas dans l'ensemble solution. Ainsi, il n'y a aucun minimum du côté positif ni du côté négatif. Il reste seulement le candidat $\beta_j = 0$. Nous devons alors analyser les dérivées à gauche et à droite, qui sont respectivement

$$f'(0^-) = -\hat{\beta}_j^0 - \lambda \ ,$$
$$f'(0^+) = -\hat{\beta}_j^0 + \lambda \ .$$

Pour que 0 soit un minimum, nous devons avoir que la fonction est décroissante en arrivant vers 0 par la gauche et croissante ensuite. Cette condition se traduit par

$$f'(0^-) = -\hat{\beta}_j^0 - \lambda \leq 0 \iff \hat{\beta}_j^0 \geq -\lambda$$

$$f'(0^+) = -\hat{\beta}_j^0 + \lambda \geq 0 \iff \hat{\beta}_j^0 \leq \lambda$$

et en combinant ces deux restrictions, on obtient bien $|\hat{\beta}_j^0| \leq \lambda$. Lorsque le coefficient MCO est dans l'intervalle $[-\lambda, \lambda]$, les minimiseurs des régions positives et négatives ne sont pas valides et en analysant la fonction autour de l'origine, nous constatons qu'elle atteint bien un minimum en 0 pour ce domaine.

Ainsi, sous l'hypothèse d'orthonormalité de $\mathbf{X}$, nous obtenons la solution

$$\hat{\beta}_j^L = \begin{cases} \hat{\beta}_j^0 - \lambda & \text{si } \hat{\beta}_j^0 > \lambda \\ 0 & \text{si } |\hat{\beta}_j^0| \leq \lambda \\ \hat{\beta}_j^0 + \lambda & \text{si } \hat{\beta}_j^0 < -\lambda \end{cases} .$$

Sous une forme plus compacte, nous avons donc

$$\hat{\beta}_j^L = \operatorname{sgn}(\hat{\beta}_j^0) \max\{0, |\hat{\beta}_j^0| - \lambda\} \ .$$

Lorsque nous cherchons un minimum à la fonction (9), le terme quadratique tend à pousser $\beta_j$ vers $\hat{\beta}_j^0$. En revanche, le terme de la contrainte Lasso $\lambda|\beta_j|$ cherche plutôt à faire tendre la solution vers 0. Si $\hat{\beta}_j^0$ est petit, alors la contrainte est plus *forte* et est en mesure d'amener le coefficient complètement à 0 (R. Tibshirani 1996).

### 2.2.2 Cadre général

Dans un cadre plus général, l'apparition de coefficients exactement nuls peut être comprise à partir des conditions de Karush-Kuhn-Tucker (Boyd et Vandenberghe 2004).

Elles sont des critères d'optimalité permettant de résoudre des problèmes sous contraintes. Nous nous penchons plus précisément sur la condition de stationnarité. Les conditions KKT peuvent être vues comme une version généralisée de la résolution d'un point critique d'une fonction différentiable. Dans notre cas, la norme $\ell_1$ possède un coin lorsqu'un coefficient vaut 0. On remplace donc la dérivée ordinaire par un ensemble de pentes admissibles appelé sous-différentiel. Ces conditions impliquent que la solution Lasso $\hat{\boldsymbol{\beta}}_L^{\lambda}$ doit satisfaire

$$\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda}) = \lambda s \ ,$$

où $s \in \partial\|\hat{\boldsymbol{\beta}}_L^{\lambda}\|_1 = \{s : \|s\|_\infty \leq 1, s^\intercal\hat{\boldsymbol{\beta}}_L^{\lambda} = \left\|\hat{\boldsymbol{\beta}}_L^{\lambda}\right\|_1\}$, un sous-gradient de la norme $\ell_1$ évalué en $\hat{\boldsymbol{\beta}}_L^{\lambda}$ (R. J. Tibshirani et Wasserman 2017). Plus précisément, cela signifie que

$$s_j \in \begin{cases} \{1\} & \text{si } \hat{\beta}_j^L > 0 \\ \{-1\} & \text{si } \hat{\beta}_j^L < 0 \\ [-1, 1] & \text{si } \hat{\beta}_j^L = 0 \end{cases} \qquad \text{pour } j = 1, \ldots, p \ .$$

Remarquons que si $\mathbf{X}$ est standardisée, $\mathbf{x}_j^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda})$ est proportionnelle à la corrélation entre le vecteur explicatif $\mathbf{x}_j$ et les résidus du modèle sous l'estimation Lasso. Dans le cas actif où $\hat{\beta}_j^L \neq 0$, si $\hat{\beta}_j^L > 0$, alors $s_j = 1$ et $\mathbf{x}_j^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda}) = \lambda$. Si $\hat{\beta}_j^L < 0$, alors $s_j = -1$ et $\mathbf{x}_j^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda}) = -\lambda$. En combinant les deux cas, nous obtenons la forme

$$\mathbf{x}_j^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda}) = \lambda \operatorname{sgn}(\hat{\beta}_j^L) \tag{10}$$

si $\hat{\beta}_j^L \neq 0$. Alors, une variable active a un produit scalaire avec le résidu qui atteint exactement le seuil $\lambda$ avec un signe cohérent avec celui du coefficient. Dans le cas où $\hat{\beta}_j^L = 0$, la norme $\ell_1$ n'a pas de dérivée unique et son sous-gradient est l'intervalle $s_j \in [-1, 1]$. La condition KKT devient alors

$$\mathbf{x}_j^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda}) = \lambda s_j \tag{11}$$

avec $s_j \in [-1, 1]$. Cette condition équivaut à

$$-\lambda \leq \mathbf{x}_j^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda}) \leq \lambda \iff |\mathbf{x}_j^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda})| \leq \lambda \ .$$

C'est cette contrainte qui explique pourquoi un coefficient peut rester exactement nul. Intuitivement, la quantité aux équations (10) et (11) est le produit scalaire entre la variable $\mathbf{x}_j$ et les résidus du modèle. Lorsque les données sont centrées et standardisées, elle est proportionnelle à leur corrélation. Cette quantité mesure à quel point la variable $j$ est encore utile pour expliquer la partie des résidus qui n'est pas expliquée par le modèle. Elle mesure à quel point la variable $j$ est encore utile pour expliquer la partie des résidus qui n'est pas expliquée par le modèle. Ainsi, une variable devient active lorsque la valeur absolue de

son produit scalaire usuel avec le résidu atteint le seuil $\lambda$. En revanche, si la valeur absolue de cette quantité est inférieure ou égale à $\lambda$, la pénalisation est assez forte pour garder son coefficient à 0. Ainsi, une variable reste inactive lorsque son produit scalaire avec le résidu est insuffisant pour dépasser le seuil imposé par $\lambda$. Les conditions KKT donnent donc une interprétation analytique de la sélection de variables effectuée par le Lasso.

En ayant en tête la discussion précédente, une question naturelle se pose : existe-t-il une valeur de $\lambda$ pour laquelle tous les coefficients sont nuls ? Il s'avère que

$$\lambda_{\max} = \max\left\{|\mathbf{x}_j^\intercal \mathbf{y}| : j \in \{1, \dots, p\}\right\}$$

est la plus petite valeur de $\lambda$ pour laquelle tous les coefficients sont nuls (El Ghaoui, Viallon et Rabbani 2010). En pensant à la géométrie du problème, nous pouvons aussi nous demander s'il existe un moyen d'enlever certaines variables explicatives préalablement à la résolution pour le Lasso afin d'alléger le modèle et faciliter les calculs. Une règle de base, proposée par El Ghaoui, Viallon et Rabbani (2010) et provenant d'un ensemble de règles surnommé SAFE, est d'éliminer préalablement le prédicteur $j$ dans $\mathbf{X}$ si

$$|\mathbf{x}_j^\intercal \mathbf{y}| < \lambda - \|\mathbf{x}_j\|_2 \|\mathbf{y}\|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}} \ . \tag{12}$$

R. Tibshirani et al. (2012) proposèrent une règle forte, généralement plus agressive en pratique, consistant à éliminer temporairement le prédicteur $j$ si

$$|\mathbf{x}_j^\intercal \mathbf{y}| < 2\lambda - \lambda_{\max} \ . \tag{13}$$

Dans le cas où les variables explicatives sont standardisées, nous pouvons montrer que le côté droit de (12) est inférieur au côté droit de (13). En revanche, lorsque les prédicteurs ne sont pas standardisés, il est plus difficile de distinguer l'ordre de grandeur entre les deux contraintes, mais la règle forte tend à mettre de côté plus de variables inutiles en pratique (R. Tibshirani et al. 2012). Aussi, contrairement aux règles SAFE, la règle forte peut rejeter à tort une variable active. Il faut donc vérifier les conditions KKT après l'ajustement et réintégrer les variables rejetées en cas de violation de ces critères.

Ces règles facilitent les calculs d'estimation de la solution Lasso. En effet, contrairement à l'estimateur Ridge, le Lasso ne possède pas de forme explicite. La contrainte sur la norme $\ell_1$ de $\boldsymbol{\beta}$ n'est pas différentiable ce qui ne permet pas de résoudre facilement le critère d'optimisation. Afin d'approximer la solution, des algorithmes numériques itératifs sont utilisés comme la descente par composante (Wu et Lange 2008), ou encore, une méthode homotopique proposée par Osborne, Presnell et Turlach (2000) ayant notamment inspiré les algorithmes de type *Least Angle Regression* (Efron et al. 2004). Les règles précédentes sont particulièrement utiles dans les algorithmes numériques qui résolvent le problème Lasso sur une grille de valeurs de $\lambda$. Dans la section 2.4, nous utilisons l'algorithme LARS-Lasso afin de construire le chemin de solutions.

### 2.2.3 Les algorithmes LARS

Le but des algorithmes de sélection de variables est de choisir un modèle linéaire en se basant directement sur les données sur lesquelles le modèle est utilisé. L'algorithme *Least Angle Regression*, ou LARS pour inclure ses variantes, est une méthode homotopique d'estimation des coefficients de $\boldsymbol{\beta}$. Dans sa forme la plus simple, cette méthode permet de trouver efficacement la solution MCO d'un modèle. Contrairement à la régression par sélection ascendante construisant un modèle séquentiellement en ajoutant une seule variable à la fois (Weisberg 2005), LAR progresse dans l'espace en fonction des corrélations entre les variables explicatives et la variable réponse.

Dans un premier temps, l'algorithme consiste à calculer la corrélation entre chacune des variables explicatives et la variable réponse. La covariable la plus corrélée avec $\mathbf{y}$, disons $\mathbf{x}_{j_1}$, est alors ajoutée dans le modèle. En revanche, au lieu de directement ajouter l'estimation MCO du coefficient le plus corrélé dans le modèle, l'algorithme LAR progresse dans l'espace vers son estimation MCO, mais pas directement jusqu'à celle-ci. En fait, l'algorithme calcule le plus grand pas possible dans la direction de la composante MCO et s'arrête lorsqu'un autre prédicteur, disons $\mathbf{x}_{j_2}$, est autant corrélé avec les résidus du modèle (Efron et al. 2004).

Dans un second temps, au lieu de continuer dans la direction de $\mathbf{x}_{j_1}$, l'algorithme déplace l'estimation dans une direction équiangle entre $\mathbf{x}_{j_1}$ et $\mathbf{x}_{j_2}$, en maintenant égales, en valeur absolue, leurs corrélations avec le résidu. Le plus grand pas est alors emprunté dans cette direction jusqu'à ce qu'une troisième variable soit autant corrélée avec les deux premières, et ainsi de suite jusqu'à la saturation du modèle ou jusqu'à la fin du chemin ; lorsque $p \leq n-1$ et sous les conditions usuelles, toutes les variables peuvent éventuellement être incluses (Efron et al. 2004).

Une modification de la procédure de base permet aussi de retracer l'ensemble des solutions Lasso. L'algorithme, tiré de Hastie, R. Tibshirani et Wainwright (2015), est présenté à l'algorithme 1. Contrairement au LAR classique, la modification Lasso permet à une variable de quitter l'ensemble actif lorsqu'un coefficient atteint 0. Le nombre d'étapes n'est donc pas nécessairement égal à $\min\{n-1, p\}$, puisqu'un retrait peut ajouter un nœud au chemin et qu'une variable retirée peut éventuellement redevenir active. La procédure est poursuivie jusqu'à la fin du chemin, correspondant à $\lambda = 0$.

En revanche, l'algorithme LAR classique suppose implicitement que les variables actives sont assez indépendantes pour que la direction équiangulaire soit bien définie. Si ce n'est pas le cas, l'algorithme prend plutôt une trajectoire possible dans l'ensemble des solutions Lasso du problème. En effet, contrairement à la fonction

$$\boldsymbol{\beta} \to \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \ ,$$

la fonction

$$\boldsymbol{\beta} \to \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

n'est pas nécessairement strictement convexe, notamment lorsque $\mathbf{X}$ n'est pas de plein rang. Ainsi, $\hat{\boldsymbol{\beta}}_L^\lambda$ n'est pas nécessairement unique. En revanche, il existe des conditions sur $\mathbf{X}$ qui garantissent l'unicité de $\hat{\boldsymbol{\beta}}_L^\lambda$. Par exemple, si les colonnes de $\mathbf{X}$ sont en position générale, alors la solution Lasso est unique [4]. Cette condition est satisfaite presque sûrement lorsque les covariables proviennent de lois continues (R. J. Tibshirani 2013). De plus, si $\text{rang}(\mathbf{X}_E) = \text{Card}(E)$ où $E$ est défini comme l'ensemble d'équicorrélation $E = \left\{j \in \{1, \ldots, p\} : |\mathbf{x}_j^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L^\lambda)| = \lambda\right\}$, alors la solution Lasso est aussi unique (R. J. Tibshirani 2023).

**Algorithme 1** *Least Angle Regression* pour une solution Lasso

1. Standardiser les prédicteurs de manière à avoir une moyenne nulle et une norme $\ell_2$ unitaire. Poser $\mathbf{r}_0 = \mathbf{y} - \overline{\mathbf{y}}\mathbf{1}_n$ et $\boldsymbol{\beta}^0 = (\beta_1, \ldots, \beta_p) = \mathbf{0}_p$ comme valeurs initiales pour les résidus et le vecteur de coefficients de régression, respectivement.
2. Trouver le prédicteur $\mathbf{x}_j$ le plus corrélé avec $\mathbf{r}_0$, c'est-à-dire celui qui maximise $|\langle \mathbf{x}_j, \mathbf{r}_0 \rangle|$. Appeler cette valeur $\lambda_0$ comme valeur de départ pour la constante de pénalisation. Définir l'ensemble actif de variables explicatives dans le modèle comme $\mathcal{H} = \{j\}$, ainsi que $\mathbf{X}_\mathcal{H}$, la matrice constituée de cette seule variable explicative.
3. Pour $k = 1, 2, \ldots$, tant que $\lambda_{k-1} > 0$ :
   - **(a)** Définir la direction des moindres carrés comme $\boldsymbol{\zeta} = \frac{1}{\lambda_{k-1}} \left(\mathbf{X}_\mathcal{H}^\intercal \mathbf{X}_\mathcal{H}\right)^{-1} \mathbf{X}_\mathcal{H}^\intercal \mathbf{r}_{k-1}$ et définir le vecteur $\boldsymbol{\mathcal{Z}}$ de dimension $p$ tel que $\boldsymbol{\mathcal{Z}}_\mathcal{H} = \boldsymbol{\zeta}$ et que le reste des éléments sont nuls.
   - **(b)** Déplacer les coefficients $\boldsymbol{\beta}$ depuis $\boldsymbol{\beta}^{k-1}$ dans la direction $\boldsymbol{\mathcal{Z}}$ vers leurs solutions MCO en calculant $\boldsymbol{\beta}(\lambda) = \boldsymbol{\beta}^{k-1} + (\lambda_{k-1} - \lambda)\boldsymbol{\mathcal{Z}}$ pour $0 < \lambda < \lambda_{k-1}$. Garder en mémoire l'évolution des résidus en calculant $\mathbf{r}(\lambda) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}(\lambda) = \mathbf{r}_{k-1} - (\lambda_{k-1} - \lambda)\mathbf{X}\boldsymbol{\mathcal{Z}}$.
   - **(c)** Garder une trace de $|\langle \mathbf{x}_\ell, \mathbf{r}(\lambda) \rangle|$ pour $\ell \notin \mathcal{H}$ et identifier la plus grande valeur de $\lambda$ à laquelle la corrélation d'une variable rattrape celle de l'ensemble actif de variables ; si la variable a l'index $j$, cela veut dire que $|\langle \mathbf{x}_j, \mathbf{r}(\lambda) \rangle| = \lambda$ et cette valeur définit le nouveau $\lambda_k$. Simultanément, si un coefficient non nul traverse 0 avant que la prochaine variable soit ajoutée, on l'enlève de $\mathcal{H}$ et nous recalculons la direction des moindres carrés jointe. Il faut identifier quel événement se produit en premier et mettre à jour $\mathcal{H}$ en conséquence. Si aucun événement d'entrée ou de sortie ne survient avant d'atteindre $\lambda = 0$, poser $\lambda_k = 0$ et terminer le chemin après la mise à jour des coefficients et des résidus.
   - **(d)** Mettre à jour l'ensemble actif selon l'événement rencontré. Si une variable inactive atteint le seuil de corrélation actif, l'ajouter à $\mathcal{H}$. Si un coefficient actif atteint 0, retirer la variable correspondante de $\mathcal{H}$. Poser ensuite $\boldsymbol{\beta}^k = \boldsymbol{\beta}(\lambda_k)$ et $\mathbf{r}_k = \mathbf{r}(\lambda_k)$.
4. Retourner la suite $\{(\lambda_k, \boldsymbol{\beta}^k)\}_{k=0}^{K_{\text{fin}}}$, où $K_{\text{fin}}$ désigne le nombre d'étapes effectuées avant d'atteindre la fin du chemin.

L'algorithme présenté suppose que les matrices $\mathbf{X}_\mathcal{H}^\intercal \mathbf{X}_\mathcal{H}$ rencontrées sont inversibles. Des extensions utilisant notamment des inverses généralisés sont nécessaires dans les cas dégénérés (Efron et al. 2004).

4. On dit que la matrice $\mathbf{X} \in \mathbb{R}^{n \times p}$ a des colonnes en position générale si tout sous-espace affine $L \subseteq \mathbb{R}^n$ de dimension $k < n$ contient au plus $k + 1$ éléments de l'ensemble $\left\{\pm \mathbf{X}_1, \ldots, \pm \mathbf{X}_p\right\}$, en excluant les paires symétriques.

Bien que le Lasso offre une bonne estimation de $\boldsymbol{\beta}$ en plus d'effectuer une sorte de sélection de variables grâce à la géométrie de la contrainte, il possède néanmoins des limites [5]. Par exemple, s'il y a une paire de variables fortement corrélées dans le modèle, le Lasso tend à en rejeter une arbitrairement (Zou et Hastie 2005). De plus, dans les situations où $n > p$, s'il y a une grande corrélation entre les covariables, il a été observé à travers des simulations que la performance de prédiction du Lasso est dominée par la régression Ridge (R. Tibshirani 1996). Cette limite du Lasso a mené à l'élaboration de l'*Elastic Net*, un compromis entre les estimateurs Ridge et Lasso.

## 2.3 L'*Elastic Net*

L'*Elastic Net* est un compromis entre les estimateurs Ridge et Lasso. Nous avons vu précédemment que bien que l'estimateur Ridge diminue la variance de l'estimateur usuel, il crée des modèles souvent peu interprétables. Pour remédier à ce problème, le Lasso est présenté comme une solution similaire à Ridge en utilisant une contrainte sur la norme de $\boldsymbol{\beta}$ dans la fonction de perte à optimiser, mais cette fois-ci, elle permet aussi d'effectuer de la sélection de variables. Comme Ridge, le Lasso peut parfois ne pas être bien adapté à certains contextes. Alors, l'idée est d'essayer de trouver une technique de régularisation palliant les inconvénients du Ridge et du Lasso, tout en conservant une bonne performance prédictive.

Considérons toujours le modèle (3). L'*Elastic Net*, proposé par Zou et Hastie (2005), permet d'effectuer de la sélection de variables, de créer un effet de *shrinkage* sur les coefficients de $\boldsymbol{\beta}$, de gérer des groupes de variables corrélées efficacement et peut être utilisé dans le cas où $p > n$. À l'instar des fonctions de perte pénalisées des sections 2.1 et 2.2, l'*Elastic Net* naïf est défini comme

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}}_{EN_{naïf}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\lambda_1\left\|\boldsymbol{\beta}\right\|_1+\frac{\lambda_2}{2}\left\|\boldsymbol{\beta}\right\|_2^2 \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\frac{\lambda_2}{2}\sum_{j=1}^{p}\beta_j^2\ ,\end{aligned} \tag{14}$$

où $\boldsymbol{\lambda}=(\lambda_1,\lambda_2)$ est un vecteur de paramètres de pénalisation [6]. En posant

$$\wp=\lambda_1+\lambda_2 \qquad \text{et} \qquad \xi=\frac{\lambda_2}{\lambda_1+\lambda_2},$$

5. Similairement à la section sur l'estimateur Ridge, les résultats présentés pour le Lasso ne requièrent pas que les conditions de Gauss-Markov soient respectées. Par contre, nous nous positionnons tout de même dans ce contexte dans les sections 2.1 et 2.2 afin de comparer facilement les estimateurs présentés avec la solution des moindres carrés usuelle.

6. Les facteurs $\frac{1}{2}$ sont introduits par convention afin de simplifier les dérivées des termes quadratiques, en faisant disparaître le facteur 2. Ils ne modifient pas la famille d'estimateurs obtenue, mais seulement l'échelle numérique des paramètres de pénalisation.

nous avons $\lambda_1 = \wp(1-\xi)$ et $\lambda_2 = \wp\xi$. Le critère d'optimisation (14) peut donc être réécrit exactement comme

$$\hat{\boldsymbol{\beta}}^{(\wp,\xi)}_{EN_{naïf}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \wp\left[(1-\xi)\sum_{j=1}^{p}|\beta_j| + \frac{\xi}{2}\sum_{j=1}^{p}\beta_j^2\right].$$

L'*Elastic Net* naïf est donc un mélange des pénalités du Ridge et du Lasso. Lorsque $\xi = 1$, nous retrouvons le critère d'optimisation de la régression Ridge. À l'inverse, lorsque $\xi = 0$, le critère devient celui du Lasso. Pour $0 < \xi < 1$, la pénalité de l'*Elastic Net* est donnée par

$$P_\xi(\boldsymbol{\beta}) = (1-\xi)\|\boldsymbol{\beta}\|_1 + \frac{\xi}{2}\|\boldsymbol{\beta}\|_2^2 \ .$$

La présence de la norme $\ell_1$ rend cette pénalité non différentiable lorsqu'au moins une composante de $\boldsymbol{\beta}$ est nulle, ce qui permet de produire des coefficients exactement nuls. Pour $\xi > 0$, la présence du terme quadratique rend la pénalité strictement convexe et favorise un effet de groupement entre les variables corrélées (Zou et Hastie 2005). Ainsi, pour $0 < \xi < 1$, l'*Elastic Net* partage certaines propriétés du Ridge et du Lasso. La figure 3 permet de bien voir comment l'*Elastic Net* est un mélange des contraintes sur les normes $\ell_1$ et $\ell_2$ dans $\mathbb{R}^2$ précisément.

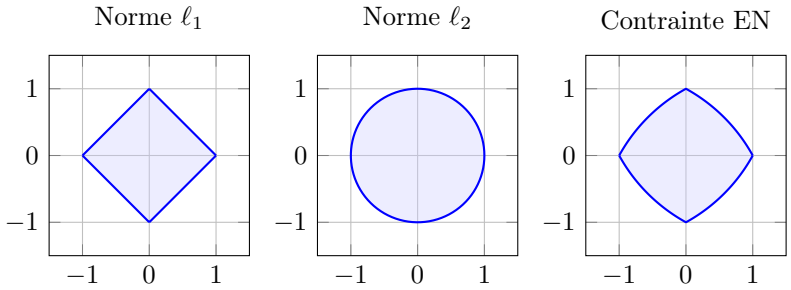


Figure 3 – Représentation des contraintes sous la norme $\ell_1$, la norme $\ell_2$ et pour l'*Elastic Net* dans $\mathbb{R}^2$. Dans cette figure, nous pouvons observer la géométrie des contraintes associées à ces trois pénalisations. Dans la figure de gauche, nous pouvons observer la géométrie de la contrainte du Lasso dans $\mathbb{R}^2$ : la boule fermée $\ell_1$ de rayon 1. De façon similaire, la figure du centre présente la contrainte utilisée dans la régression Ridge : la boule $\ell_2$ de rayon 1. La figure de droite présente la contrainte utilisée dans l'*Elastic Net* pour une valeur de $\xi = 0.45$. Nous pouvons observer ses sommets permettant d'introduire de la parcimonie dans l'estimation de $\boldsymbol{\beta}$.

Il s'avère que résoudre le critère d'optimisation (14) est équivalent à résoudre un critère de type Lasso, ce qui permet d'utiliser les méthodes computationnelles présentées à la section 2.2. Plus précisément, considérons la variable expliquée $\mathbf{y}$, les variables explicatives $\mathbf{X}$ ainsi que le vecteur de paramètres de pénalisation $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. Définissons le jeu de données augmenté $(\mathbf{y}*, \mathbf{X}^*)$ comme

$$\mathbf{X}^*{}_{(n+p)\times p} = \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{I}_{p\times p}\end{pmatrix}, \quad \mathbf{y}^*{}_{(n+p)} = \begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}. \tag{15}$$

En posant $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$, on peut vérifier que

$$\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \lambda_1\left\|\boldsymbol{\beta}\right\|_1 + \frac{\lambda_2}{2}\left\|\boldsymbol{\beta}\right\|_2^2 = \frac{1}{2}\left\|\mathbf{y}^*-\mathbf{X}^*\boldsymbol{\beta}^*\right\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\left\|\boldsymbol{\beta}^*\right\|_1 \ .$$

Dans un tel cas, en posant

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}^*\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2}\left\|\mathbf{y}^*-\mathbf{X}^*\boldsymbol{\beta}^*\right\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\left\|\boldsymbol{\beta}^*\right\|_1\right\}.$$

alors,

$$\hat{\boldsymbol{\beta}}_{EN_{naïf}} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^* \ ,$$

où $\hat{\boldsymbol{\beta}}^*$ est la solution intermédiaire de type Lasso. Ainsi, nous pouvons transformer le critère d'optimisation pour l'*Elastic Net* naïf en un critère Lasso sur des données augmentées. Cette forme permet donc de résoudre un premier problème relevé avec le Lasso classique. Contrairement à celui-ci, étant donné que nous travaillons sur des données augmentées de taille $(n+p)$, l'*Elastic Net* naïf peut potentiellement sélectionner tous les prédicteurs, même dans le cas où $p > n$ lorsque $\lambda_2 > 0$.

De plus, les deux méthodes partagent l'avantage d'effectuer de la sélection de variables. Pour l'*Elastic Net* naïf, le développement est similaire à celui fait pour le Lasso dans les sections 2.2.1 et 2.2.2. En résolvant pour le cas orthogonal dans lequel nous avons $\hat{\boldsymbol{\beta}}_0 = \mathbf{X}^\intercal\mathbf{y}$, nous obtenons que

$$\hat{\beta}_j^{EN_{naïf}} = \frac{\max\{0, |\hat{\beta}_j^0| - \lambda_1\}}{1+\lambda_2}\operatorname{sgn}(\hat{\beta}_j^0) \ .$$

Ce résultat permet alors de comprendre pourquoi l'*Elastic Net* produit également une forme de sélection de variables.

Additionnellement, cette nouvelle méthode gère mieux la colinéarité entre les variables explicatives. Effectivement, la présence du terme quadratique dans la pénalité favorise un effet de groupement entre les variables corrélées. Dans le cas extrême où deux variables explicatives sont identiques, l'*Elastic Net* leur attribue des coefficients égaux, par un résultat de Zou et Hastie (2005). En particulier, une méthode de régression montre un effet de groupement si elle a tendance à assigner des valeurs presque égales à des groupes de coefficients fortement corrélés. Dans la situation extrême où certaines variables sont identiques, la méthode devrait leur donner des valeurs de coefficients égales. Dans un cadre plus réaliste, la méthodologie amène les coefficients associés à des prédicteurs fortement corrélés à être proches. Cet effet de groupement est particulièrement utile dans les contextes où plusieurs variables explicatives fortement corrélées peuvent contribuer conjointement à la réponse, comme dans certaines applications en génétique. L'*Elastic Net* est alors souvent plus adapté que le Lasso, qui tend plutôt à sélectionner une seule variable parmi un groupe de variables corrélées.

Bien que l'*Elastic Net* naïf soit plus adapté aux situations où $p > n$ et en présence de variables fortement corrélées, il s'avère que ses performances peuvent être faibles lorsque le paramètre $\xi$ est loin de 0 et 1. En d'autres termes, les performances de prédiction ne sont pas satisfaisantes lorsque l'*Elastic Net* n'est pas proche d'être une régression Ridge ou Lasso. Ce défaut vient d'un effet de double rétrécissement dans la procédure. En effet, dans sa résolution, pour une valeur de $\lambda_2$ fixée, nous trouvons d'abord les coefficients de la régression Ridge pour ensuite effectuer un second rétrécissement le long de la solution Lasso. Face à ce problème, Zou et Hastie (2005) proposent l'*Elastic Net*, non naïf cette fois-ci, de la forme

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}}_{EN} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\boldsymbol{\beta}^\intercal \left( \frac{\mathbf{X}^\intercal\mathbf{X} + \lambda_2 \mathbf{I}_{p\times p}}{1+\lambda_2} \right) \boldsymbol{\beta} - \mathbf{y}^\intercal \mathbf{X}\boldsymbol{\beta} + \lambda_1 \left\| \boldsymbol{\beta} \right\|_1 \right\} \ ,$$

à une constante additive indépendante de $\boldsymbol{\beta}$ près. Dans ce cas-ci, l'*Elastic Net* corrigé est

$$\hat{\boldsymbol{\beta}}_{EN} = (1+\lambda_2)\hat{\boldsymbol{\beta}}_{EN_{naïf}} \ . \tag{16}$$

Ainsi, les coefficients de l'*Elastic Net* corrigé ne sont que les coefficients de l'*Elastic Net* naïf réajustés. Cet ajustement permet de contourner l'effet de double rétrécissement et améliore la capacité de prédiction de l'*Elastic Net* naïf.

En somme, cette méthode règle certains problèmes inhérents à la régression Lasso tout en préservant ses propriétés de sélection de variables [7]. Le lien intrinsèque entre ces deux méthodes permet aussi de facilement modifier l'algorithme LARS du Lasso pour résoudre le problème de l'*Elastic Net*. Il suffit d'augmenter les données comme montré précédemment, puis d'appliquer l'algorithme LARS à ces données augmentées afin d'obtenir $\hat{\boldsymbol{\beta}}^*$. L'estimateur *Elastic Net* corrigé est ensuite obtenu par le réajustement

$$\hat{\boldsymbol{\beta}}_{EN} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^* \ ,$$

ou, de manière équivalente, par la relation (16). Le choix des paramètres de pénalisation se fait par validation croisée à $k$ blocs. Ce sont ces algorithmes qui sont utilisés dans la section 2.4.

Tout au long de la section sur les estimateurs fréquentistes, nous remarquons que certains d'entre eux ne possèdent pas nécessairement de forme explicite, ce qui complique la quantification de certaines de leurs caractéristiques, comme leur variance. Il est facile de déterminer la variance théorique de l'estimateur Ridge, mais ce n'est pas le cas pour l'estimateur Lasso ou encore pour l'*Elastic Net*. Ainsi, dans la section bayésienne, nous montrerons une interprétation bayésienne du Lasso afin d'étudier plus facilement sa structure.

7. Les hypothèses de Gauss-Markov ne sont pas nécessaires dans les équations présentées dans cette section.

## 2.4 Exemples fréquentistes

Dans cette section, nous présentons les estimations du vecteur de paramètres $\boldsymbol{\beta}$ obtenues à l'aide des méthodes fréquentistes présentées précédemment. Dans un premier temps, l'estimateur Ridge, le Lasso et l'*Elastic Net* sont comparés sur un jeu de données simulé. Dans un second temps, nous utilisons les données sur le cancer de la prostate de Stamey et al. (1989), également étudiées par Zou et Hastie (2005).

### 2.4.1 Comparaison sur un jeu de données simulé

Dans cette première sous-section, nous estimons le vecteur de coefficients $\boldsymbol{\beta}$ à partir d'un jeu de données simulé. Nous fixons $n = 72$ et $p = 16$. Bien que $p < n$, ce choix correspond à un problème de dimension modérée dans lequel l'estimateur des moindres carrés ordinaires demeure calculable, mais où les méthodes de rétrécissement peuvent réduire la variance de l'estimation.

Une matrice de prédicteurs de dimension $n \times p$ est d'abord générée à partir de variables aléatoires indépendantes suivant une loi $\Gamma(2, 2)$, dont les paramètres de forme et de taux sont égaux à 2. Les colonnes de cette matrice sont ensuite centrées et réduites. Un vecteur d'erreurs est également simulé selon une loi $\mathcal{N}(0, 1)$. Le vecteur de réponses $\mathbf{y}$ est alors construit à partir du modèle donné à l'équation (3).

Le vrai vecteur de coefficients utilisé dans cette simulation est

$$\boldsymbol{\beta} = (25, -10, 0, 2, -5, 0, 0, 0.2, 0, 0, 0, 0, 0, 0, 0, 0)^\intercal \ . \tag{17}$$

Le choix de ces composantes répond à plusieurs objectifs. Premièrement, 11 des 16 coefficients sont fixés à zéro afin d'évaluer la capacité du Lasso et de l'*Elastic Net* à effectuer une sélection de variables. Deuxièmement, la présence de coefficients positifs et négatifs permet de vérifier que les méthodes fondées sur l'algorithme LARS retrouvent correctement leur signe. Finalement, les coefficients non nuls présentent des ordres de grandeur différents, allant de 25 à 0.2. Puisque les prédicteurs sont standardisés, leurs coefficients peuvent être comparés sur une même échelle.

Afin de comparer les performances des différents estimateurs, plusieurs critères sont considérés. Dans un premier temps, l'erreur quadratique moyenne de prédiction sur l'échantillon d'entraînement est calculée selon

$$\mathrm{EQMP}_{\mathrm{ent}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{18}$$

où $\hat{y}_i$ désigne la prévision associée à la $i$-ième observation obtenue à partir du vecteur estimé $\hat{\boldsymbol{\beta}}$.

Une EQMP de test est également calculée sur un échantillon indépendant de l'échantillon d'entraînement. Pour ce faire, $N = 10\,000$ nouvelles observations sont générées selon

le même mécanisme que les données d'entraînement. Les nouveaux prédicteurs sont transformés en utilisant les paramètres de centrage et de réduction estimés à partir de l'échantillon d'entraînement. Les réponses sont ensuite générées à partir du même vecteur $\boldsymbol{\beta}$ et de nouvelles erreurs indépendantes suivant une loi $\mathcal{N}(0,1)$.

Finalement, la précision de l'estimation des coefficients est mesurée par l'erreur quadratique

$$D_{\boldsymbol{\beta}} = \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\|_2^2 . \tag{19}$$

Dans le tableau 1, $\lambda_1$ désigne le paramètre associé à la pénalisation $\ell_1$, tandis que $\lambda_2$ désigne celui associé à la pénalisation quadratique. Le tableau présente les résultats obtenus avec les MCO, la régression Ridge, le Lasso et l'*Elastic Net*. Les estimations individuelles des coefficients du vecteur $\boldsymbol{\beta}$ sont présentées à l'annexe A.

TABLE 1 – Performances prédictives et erreur d'estimation des coefficients fréquentistes sur les données simulées.

| | $\lambda_1$ | $\lambda_2$ | $EQMP_{ent}$ | $EQMP_{test}$ | Distance | Nombre non nuls |
|---|---|---|---|---|---|---|
| MCO | | | 0.7965 | 1.2582 | 0.2216 | 16 |
| Ridge | | 0.0242 | 0.7966 | 1.2650 | 0.2271 | 16 |
| Lasso | 3.9192 | | 0.8286 | 1.1516 | 0.1389 | 12 |
| Elastic Net | 7.6168 | 0.0075 | 0.8854 | 1.0971 | 0.0937 | 10 |

Comme prévu, l'estimateur des MCO possède la plus faible EQMP d'entraînement, soit 0.7965, puisqu'il minimise directement la somme des carrés des résidus sur l'échantillon utilisé pour l'ajustement. Les estimateurs pénalisés introduisent quant à eux un certain biais, ce qui explique que le Lasso et l'*Elastic Net* présentent des erreurs d'entraînement plus élevées.

Cette situation est toutefois inversée lorsque les modèles sont évalués sur les données de test. Les EQMP de test du Lasso et de l'*Elastic Net* sont respectivement de 1.1516 et de 1.0971, comparativement à 1.2582 pour les MCO. Ces résultats illustrent le compromis biais-variance. En effet, la pénalisation détériore légèrement l'ajustement aux données d'entraînement, mais elle peut améliorer la capacité de généralisation du modèle.

Puisque la variance du terme d'erreur est égale à 1, le risque quadratique de prédiction possède une borne inférieure de 1. Les EQMP de test peuvent donc être comparées à cette valeur de référence. L'*Elastic Net* est donc la méthode dont l'EQMP de test se rapproche le plus de cette valeur dans la présente réalisation. L'estimateur Ridge obtient, pour sa part, une EQMP de test de 1.2650, soit légèrement supérieure à celle des MCO. Cette différence demeure toutefois faible et ne permet pas, à elle seule, de conclure que l'estimateur Ridge est systématiquement moins efficace.

Le Lasso et l'*Elastic Net* présentent également l'avantage d'effectuer une sélection de variables. D'après les valeurs affichées à l'annexe A, aucun coefficient estimé par les MCO ou par la régression Ridge n'est nul, comme prévu. Le Lasso retrouve 4 des 11 coefficients

réellement nuls, tandis que l'*Elastic Net* en retrouve 6. Les deux méthodes conservent par ailleurs les cinq coefficients réellement non nuls.

L'*Elastic Net* présente également la plus faible erreur d'estimation des coefficients, soit

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 = 0.0937 \ ,$$

comparativement à 0.1389 pour le Lasso, à 0.2216 pour les MCO et à 0.2271 pour la régression Ridge.

Ainsi, dans cette réalisation, l'*Elastic Net* offre le meilleur compromis entre la qualité de la prévision, la précision de l'estimation et la sélection de variables. Ce résultat ne doit toutefois pas être interprété comme une démonstration de sa supériorité générale. Une étude plus approfondie, dans laquelle la génération des données et l'estimation des modèles seraient répétées un grand nombre de fois, serait nécessaire afin de comparer plus solidement les performances moyennes des différentes méthodes.

### 2.4.2 Comparaison sur un jeu de données réel

Dans cette seconde comparaison, les méthodes fréquentistes d'estimation du vecteur de coefficients de régression $\boldsymbol{\beta}$ sont testées sur un jeu de données réel. Le jeu de données provient d'une étude portant sur la concentration de l'antigène prostatique spécifique comme indicateur de la progression du cancer de la prostate (Stamey et al. 1989). Ces données sont notamment utilisées dans Zou et Hastie (2005). Plus précisément, la variable réponse est le logarithme de la quantité de l'antigène prostatique spécifique (`lpsa`) et les huit variables explicatives sont le logarithme du volume de la masse cancéreuse (`lcavol`), le logarithme du poids de la prostate (`lweight`), l'âge du patient (`age`), le logarithme de la quantité d'hyperplasie bénigne de la prostate (`lbph`), un indicateur de la présence d'envahissement des vésicules séminales (`svi`), le logarithme du franchissement capsulaire (`lcp`), le score de Gleason (`gleason`) ainsi que le pourcentage de grades de Gleason 4 ou 5 (`pgg45`). Le jeu de données, composé de 97 observations, est séparé en un groupe d'entraînement de 67 observations et un groupe de test de 30 observations pour calculer l'EQMP des différentes méthodes.

Premièrement, l'estimateur MCO est calculé à partir des données d'entraînement, ainsi que son $EQMP_{ent}$. Des prédictions sont faites avec les données de test afin d'obtenir son $EQMP_{test}$.

Deuxièmement, l'estimateur Ridge est calculé. Contrairement à la section 2.4.1 dans laquelle la validation croisée généralisée est utilisée, la validation croisée à $k$ blocs est utilisée dans cette section avec $k = 10$. Pour y arriver, les 67 observations d'entraînement sont séparées en 10 blocs de taille presque égale. Une grille de candidats pour la valeur de $\lambda$ est parcourue et pour chaque bloc et chaque $\lambda$, le modèle est ajusté sur les neuf autres blocs. Son EQMP est calculée sur le bloc restant et les dix EQMP sont moyennées. La valeur de $\lambda$ avec la plus faible EQMP est retenue et Ridge est réajusté sur les 67 observations. L'EQMP de test est ensuite calculée sur les 30 observations de test.

Troisièmement, l'algorithme LARS-Lasso est utilisé pour construire le chemin de solutions Lasso. Cet algorithme retourne essentiellement les nœuds du chemin de solutions dans $\mathbb{R}^p$ lorsqu'une variable entre ou sort du modèle. Une variable $s$, calculée comme

$$s(\lambda_1) = \frac{\left\|\hat{\boldsymbol{\beta}}(\lambda_1)\right\|_1}{\max_{\iota \geq 0} \left\|\hat{\boldsymbol{\beta}}(\iota)\right\|_1} ,$$

est utilisée afin de repérer une position sur le chemin des coefficients produit par l'algorithme (Zou et Hastie 2005). Essentiellement, une valeur de $s = 0$ représente un modèle dans lequel tous les coefficients sont nuls, ce qui correspond à une pénalisation maximale. À l'inverse, $s = 1$ représente la fin du chemin Lasso et donc une pénalité nulle. Nous procédons encore par validation croisée sur les mêmes blocs établis pour le Ridge. Ainsi, une grille de valeurs de $s$ est testée. Un chemin de solutions LARS est calculé à partir des 9 blocs d'entraînement. Pour une valeur de $s$ entre les nœuds retournés par l'algorithme, une interpolation linéaire est effectuée entre les deux points pour trouver une valeur de $\boldsymbol{\beta}$ correspondante. La valeur de $s$ minimisant l'EQMP de validation croisée est retenue. Le chemin de solutions LARS est recalculé sur les 67 observations et les coefficients associés à la valeur de $s$ retenue sont interpolés, ainsi que la valeur du paramètre de pénalisation $\lambda_1$. L'EQMP d'entraînement et de test sont finalement calculées.

Quatrièmement, la solution de l'*Elastic Net* est calculée à partir des données augmentées (15). Nous procédons ensuite par validation croisée en choisissant un couple de valeurs $(s, \lambda_2)$. De manière analogue à la procédure pour le Lasso, une grille de valeurs de $\lambda_2$ est testée. Pour chacune de ces valeurs, un chemin LARS sur les données augmentées est calculé. Sur ce chemin, des valeurs de $s$ sont testées. Le couple de paramètres minimisant l'EQMP de validation croisée est retenu, le modèle final est ajusté sur les 67 observations et les mesures de performance sont calculées. Le tableau 2 compare les mesures de performance des méthodes, tandis que les estimations des coefficients de $\boldsymbol{\beta}$ sont présentées à l'annexe B.

Table 2 – Performances d'ajustement, performances prédictives et parcimonie des modèles sous les méthodes fréquentistes sur les données du cancer de la prostate.

| | $s$ | $\lambda_1$ | $\lambda_2$ | $EQMP_{ent}$ | $EQMP_{test}$ | Nombre non nuls |
|---|---|---|---|---|---|---|
| MCO | | | | 0.4392 | 0.5213 | 8 |
| Ridge | | | 3.1518 | 0.4433 | 0.5000 | 8 |
| Lasso | 0.8800 | 0.8222 | | 0.4423 | 0.4949 | 7 |
| Elastic Net | 0.5000 | 7.2346 | 0.1084 | 0.4929 | 0.4357 | 5 |

Encore une fois, tel qu'attendu par la théorie, l'EQMP d'entraînement est bien minimisée par la méthode des MCO avec une valeur de 0.4392. Cette méthode d'estimation est pourtant suivie de près par Ridge et le Lasso pour cette métrique avec des valeurs respec-

tives de 0.4433 et 0.4423. L'*Elastic Net*, quant à lui, présente une EQMP d'entraînement plus élevée, soit 0.4929.

En revanche, les méthodes de pénalisation semblent être plus performantes en termes de prédictions. Sur l'ensemble test, les trois méthodes pénalisées obtiennent une EQMP inférieure à celle des MCO, qui s'élève à 0.5213. Les valeurs sélectionnées des paramètres de pénalisation sont strictement positives, de sorte que les modèles retenus diffèrent effectivement de la solution non pénalisée. Ces méthodes semblent donc améliorer les performances prédictives sur cet ensemble test. Plus précisément, l'estimateur Ridge diminue l'erreur quadratique moyenne de prévision dans le jeu de test de 0.0213 unités. Le Lasso améliore légèrement le résultat avec une diminution de 0.0264. En revanche, l'*Elastic Net* réduit davantage cette erreur avec une différence de 0.0856 comparativement aux MCO.

Aussi, nous pouvons observer l'avantage du Lasso et de l'*Elastic Net* qui introduisent de la parcimonie dans le modèle. En plus d'obtenir une EQMP test inférieure à celles de Ridge et des MCO, ces méthodes proposent bien des modèles plus simples. L'Elastic Net présente ici une amélioration par rapport au Lasso en ayant une EQMP de test moindre ainsi que seulement 5 variables sélectionnées au lieu de 7. Une explication possible de cette performance est la présence de corrélations entre certains prédicteurs. Cette méthode tend à bien gérer ce phénomène. Grâce à la combinaison des pénalités $\ell_1$ et $\ell_2$, l'*Elastic Net* peut à la fois mieux traiter les groupes de variables corrélées et produire des coefficients exactement nuls.

Ainsi, pour ces données spécifiques, l'*Elastic Net* domine encore les autres techniques d'estimation en offrant un modèle plus simple que les autres avec seulement 5 variables retenues et en obtenant la plus faible EQMP de test parmi les quatre méthodes étudiées.

# 3 Méthodes bayésiennes

Nous explorons maintenant la régression linéaire d'un point de vue bayésien. Contrairement à l'approche fréquentiste, dans laquelle le paramètre $\boldsymbol{\beta}$ est considéré comme fixe, mais inconnu, l'approche bayésienne représente l'incertitude entourant sa valeur en lui attribuant une distribution de probabilité.

Le principe fondamental de la statistique bayésienne est la mise à jour de la distribution du paramètre d'intérêt à l'aide de données. Plus précisément, une loi *a priori* est attribuée au paramètre aléatoire $\boldsymbol{\beta}$ considéré. Cette loi est ensuite mise à jour à l'aide des données observées. Cette mise à jour permet d'accéder à la loi *a posteriori* de la variable aléatoire $\boldsymbol{\beta}$ et d'étudier sa distribution à la lumière des données recueillies. Ce processus se fait à l'aide du théorème de Bayes. Plus particulièrement, considérons un paramètre quelconque $\boldsymbol{\theta}$ ainsi que des observations $\mathbf{x}$. Le théorème de Bayes stipule alors que la densité *a posteriori* de $\boldsymbol{\theta}$ est

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\displaystyle\int_{\Theta} f(\mathbf{x}|\boldsymbol{u})\pi(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}}\ , \tag{20}$$

où $\pi(\boldsymbol{\theta})$ est la densité *a priori* de $\boldsymbol{\theta}$ et $\Theta$ est son espace paramétrique. La densité des observations $\mathbf{x}$ conditionnellement à $\boldsymbol{\theta}$ est $f(\mathbf{x}|\boldsymbol{\theta})$ et, lorsque $\mathbf{x}$ est fixé, cette expression considérée comme une fonction de $\boldsymbol{\theta}$ correspond à la vraisemblance. En pratique, il est souvent possible de travailler à une constante de proportionnalité près. En effet, le dénominateur dans (20) ne dépend pas de $\boldsymbol{\theta}$ et agit seulement comme une constante de normalisation. Nous pouvons donc écrire

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \ ,$$

ce qui permet d'identifier le noyau de la densité *a posteriori*. Lorsque ce noyau correspond à celui d'une loi connue, la distribution *a posteriori* peut être reconnue directement. Si, au contraire, la distribution est difficilement reconnaissable, des méthodes numériques peuvent être nécessaires pour effectuer les calculs à l'aide de celle-ci [8].

Dans le contexte de la régression linéaire, en attribuant une loi *a priori* à $\boldsymbol{\beta}$ et $\sigma^2$ et en conditionnant sur la matrice de design $\mathbf{X}$, la loi *a posteriori* devient

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}, \sigma^2) \ . \tag{21}$$

Cette approche permet d'exprimer nos croyances à propos du paramètre d'intérêt $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, tout en tenant compte des données observées $(\mathbf{y}, \mathbf{X})$. La loi *a priori* permet d'encoder l'information disponible, ou l'incertitude, à l'égard du paramètre $\boldsymbol{\theta}$ avant l'observation des données. Son choix constitue une étape importante de la modélisation, puisqu'il influence directement la distribution *a posteriori* du paramètre. Le modèle échantillonnal $f(\mathbf{x}|\boldsymbol{\theta})$ est généralement motivé par les hypothèses formulées sur le mécanisme ayant produit les données, tandis que la loi *a priori* nécessite des choix supplémentaires concernant le paramètre. Sous certaines conditions, l'influence de la loi a priori tend à diminuer à mesure que la taille de l'échantillon augmente. Cette propriété n'est toutefois pas automatique et dépend du modèle ainsi que du choix de la loi *a priori* (Lee 1989).

Il existe plusieurs avantages à utiliser une approche bayésienne. Par exemple, la loi *a posteriori* permet de quantifier l'incertitude entourant les paramètres et de construire des intervalles de crédibilité. Cette loi peut également être résumée par sa moyenne, sa médiane ou son mode, selon la fonction de perte considérée. Aussi, plusieurs applications statistiques impliquent plusieurs paramètres qui peuvent être connectés à travers la structure du problème. Le cadre bayésien permet de tenir compte de ces dépendances au moyen de distributions jointes et conditionnelles (Gelman et al. 1995). Une manière efficace et facilement implémentable de considérer ces liens est d'utiliser des modèles hiérarchiques. Un modèle bayésien hiérarchique est obtenu lorsque la loi *a priori* d'un paramètre est spécifiée conditionnellement à un ou plusieurs hyperparamètres, auxquels sont à leur tour attribuées des distributions. Cette distribution *a priori* est composée de distributions conditionnelles

$$\pi_1(\boldsymbol{\theta}|\omega_1), \pi_2(\omega_1|\omega_2), \quad \ldots \quad , \pi_L(\omega_{L-1}|\omega_L)$$

8. Ces techniques numériques peuvent parfois être nécessaires dans un contexte avec des distributions connues également.

et d'une distribution *a priori* pour l'hyperparamètre du dernier niveau, $\pi_{L+1}(\omega_L)$, telles que

$$\pi(\boldsymbol{\theta}) = \int_{\Omega_1 \times \cdots \times \Omega_L} \pi_1(\boldsymbol{\theta}|\omega_1)\pi_2(\omega_1|\omega_2)\ldots\pi_L(\omega_{L-1}|\omega_L)\cdot\pi_{L+1}(\omega_L)\mathrm{d}\omega_1\ldots\mathrm{d}\omega_L \ ,$$

où les paramètres $\omega_i$ sont appelés les hyperparamètres de niveau $i$, pour $i = 1, \ldots, L$.

En revanche, pour la majorité des modèles avec plusieurs paramètres et une hiérarchie complexe, il peut être difficile d'échantillonner directement selon la distribution jointe *a posteriori*. Il s'avère, en général, qu'il est beaucoup plus facile d'échantillonner celle-ci à partir des distributions conditionnelles complètes de chaque paramètre. La distribution conditionnelle complète d'un paramètre est sa distribution *a posteriori* conditionnellement aux données et à tous les autres paramètres du modèle. Pour une composante $\theta_j$ du paramètre d'intérêt, elle s'écrit

$$\pi(\theta_j|\boldsymbol{\theta}_{-j}, \mathbf{x}) \ ,$$

où $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$. Dans ce cas, l'échantillonneur de Gibbs peut être utilisé pour étudier la distribution *a posteriori*. Il s'agit d'une méthode de Monte-Carlo par chaînes de Markov qui construit une chaîne dont la distribution stationnaire est la loi a posteriori cible (Hoff 2009).

Intuitivement, l'algorithme de Gibbs génère des observations selon une distribution cible multidimensionnelle, une composante à la fois, en utilisant les distributions conditionnelles complètes de ces composantes (Bédard 2025). Plus précisément, considérons maintenant un hyperparamètre $\omega$ pour $\theta$. Dans sa forme la plus simple, une fois que nous avons choisi des valeurs initiales pour les paramètres, une itération de l'échantillonneur de Gibbs s'effectue comme décrit dans l'algorithme 2 en supposant que la chaîne se trouve à l'état $(\theta^{(i)}, \omega^{(i)})$ à l'itération $i$.

**Algorithme 2** Itération d'un échantillonneur de Gibbs à un paramètre $\theta$ avec un hyperparamètre $\omega$

**1.** Générer $\theta^{(i+1)}$ selon la distribution conditionnelle complète $\pi(\cdot|\omega^{(i)}, \mathbf{x})$.

**2.** Générer $\omega^{(i+1)}$ selon la distribution conditionnelle complète $\pi(\cdot|\theta^{(i+1)}, \mathbf{x})$.

$\rightarrow$ Le couple produit à l'itération $i+1$, $(\theta^{(i+1)}, \omega^{(i+1)})$, est ensuite utilisé lors de l'itération suivante.

À chaque étape, nous conditionnons donc toujours sur les valeurs les plus récentes des paramètres. Sous des conditions usuelles d'ergodicité, la distribution de la chaîne converge vers la distribution *a posteriori* cible. Les moyennes empiriques calculées à partir des observations générées peuvent alors être utilisées pour approximer les espérances *a posteriori*. L'algorithme s'étend naturellement à plusieurs paramètres, pourvu qu'il soit possible de générer des observations à partir de leur distribution conditionnelle complète.

L'algorithme de Gibbs est particulièrement utile pour l'estimation du paramètre $\boldsymbol{\beta}$ de notre régression linéaire, dorénavant considéré comme une variable aléatoire. Dans cette

section, nous présentons d'abord l'interprétation du Lasso comme un mode *a posteriori* obtenu en attribuant des lois *a priori* de Laplace indépendantes entre elles à chacun des coefficients de régression. Ensuite, différents modèles issus de lois *a priori* de type *Spike and Slab* sont explorés. Afin d'étudier ces différentes approches d'estimation de $\boldsymbol{\beta}$, l'échantillonneur de Gibbs est utilisé pour générer des observations provenant approximativement des lois *a posteriori* correspondantes. Ces observations permettent ensuite d'estimer les coefficients et de quantifier l'incertitude qui les entoure.

## 3.1 Le Lasso bayésien

Dans cette sous-section, positionnons-nous dans le contexte de l'équation (2) avec $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_{n\times n})$. Dans notre contexte de régression, nous pouvons attribuer des lois *a priori* aux différents paramètres contenus dans le problème, donc modéliser ces paramètres avec des lois spécifiques. Plus précisément, nous cherchons des lois *a priori* sur le paramètre $\boldsymbol{\beta}$ qui accomplissent des objectifs similaires aux techniques fréquentistes explorées précédemment. Il s'avère que le critère d'optimisation menant à la solution Lasso (8) possède une interprétation bayésienne à travers l'utilisation d'un certain type de loi *a priori* sur les coefficients de régression. En effet, ce critère peut être interprété comme le maximum *a posteriori* (MAP) lorsque des lois de Laplace indépendantes sont utilisées sur les coefficients de régression. Sous cette condition, la loi *a priori* de $\boldsymbol{\beta}$ est

$$\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda_b}{2}\, \mathrm{e}^{-\lambda_b|\beta_j|} = \left(\frac{\lambda_b}{2}\right)^p \exp\left\{ -\lambda_b \sum_{j=1}^{p} |\beta_j| \right\},$$

où $\lambda_b > 0$. De manière concise, nous avons $\beta_j \sim \mathcal{L}aplace(0, \frac{1}{\lambda_b})$ pour $j = 1, \ldots, p$. Comme mentionné précédemment, dans notre contexte de régression sous normalité des erreurs, nous avons que $\mathbf{y}|\alpha, \boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim \mathcal{N}_n(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_{n\times n})$. Avec une loi *a priori* indépendante $\pi(\sigma^2)$ pour $\sigma^2 > 0$, nous pouvons trouver la distribution *a posteriori* dont le maximum est équivalent à trouver la solution au critère de minimisation Lasso. En effet, à l'aide du théorème de Bayes (21) et en attribuant une loi *a priori* indépendante plate sur $\alpha$, nous avons que

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2, \alpha) &\propto f(\mathbf{y}|\alpha, \sigma^2, \mathbf{X}, \boldsymbol{\beta})\pi(\boldsymbol{\beta}) \\
&\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\} \prod_{j=1}^{p} \frac{\lambda_b}{2} e^{-\lambda_b|\beta_j|} \\
&\propto_{\boldsymbol{\beta}} \exp\left\{ -\frac{1}{2\sigma^2} \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\} \exp\left\{ -\lambda_b \sum_{j=1}^{p} |\beta_j| \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2} \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \lambda_b \sum_{j=1}^{p} |\beta_j| \right\}.
\end{aligned}$$

Donc, le MAP s'obtient en maximisant une fonction proportionnelle en $\boldsymbol{\beta}$ à l'équation précédente. Pour maximiser ce terme, nous devons maximiser le terme dans l'exponentielle, ce qui est équivalent à minimiser le même terme de signe contraire. On définit alors l'estimateur par la méthode du maximum *a posteriori* comme :

$$\hat{\boldsymbol{\beta}}_{MAP} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2} \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sigma^2\lambda_b \sum_{j=1}^{p} |\beta_j| \ .$$

Il est alors très facile de voir le lien avec le critère du Lasso en centrant les données et en posant le paramètre de pénalisation Lasso comme $\lambda_L = \sigma^2\lambda_b$. Donc, pour toute valeur fixe $\sigma^2 > 0$, le vecteur $\boldsymbol{\beta}$ qui maximise la loi *a posteriori* est une estimation Lasso [9].

En revanche, Park et Casella (2008) soulignent que, bien que la maximisation de la loi *a posteriori* soit pratique, elle ne constitue pas toujours une procédure bayésienne naturelle. Un problème soulevé est que le mode *a posteriori* n'est pas nécessairement préservé sous marginalisation, c'est-à-dire que

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\sigma^2>0}{\arg\max} \quad \pi(\boldsymbol{\beta}, \sigma^2|\alpha, \mathbf{y}, \mathbf{X})$$

ne donnera pas forcément la même solution en $\boldsymbol{\beta}$ que

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max} \quad \pi(\boldsymbol{\beta}|\alpha, \mathbf{y}, \mathbf{X}) \ .$$

Une analyse pleinement bayésienne suggérerait plutôt d'utiliser les médianes ou les moyennes *a posteriori* afin d'obtenir des estimations spécifiques du paramètre de régression. Conséquemment, la propriété de sélection de variables caractérisant si bien le Lasso n'est pas préservée sous ces estimations.

L'approche bayésienne du Lasso présente tout de même un avantage de taille en permettant de quantifier l'incertitude entourant le paramètre d'intérêt. En effet, il est très facile d'estimer des intervalles de crédibilité pour les coefficients de régression à l'aide d'un échantillonneur de Gibbs. Pour ce faire, l'échantillonneur repose sur l'usage d'une représentation de la loi de Laplace comme un mélange de lois normales. Plus précisément, nous avons la décomposition suivante, avec $a > 0$, de la loi de Laplace :

$$\begin{aligned} I = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} \mathrm{d}s &= \frac{a^2}{2} \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left\{ -\frac{(z^2 + s^2a^2)}{2s} \right\} \mathrm{d}s \\ &= \frac{a^2}{2} \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left\{ -\frac{(|z|^2 + s^2a^2)}{2s} \right\} \mathrm{d}s \\ &= \frac{a^2}{2} \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left\{ -\frac{(|z| - sa)^2}{2s} - \frac{2|z|sa}{2s} \right\} \mathrm{d}s \ . \end{aligned}$$

9. Dans la suite, afin d'alléger la notation, nous noterons simplement $\lambda$ comme étant le paramètre de la loi *a priori*.

Puis, en factorisant le terme exponentiel indépendant de $s$, nous pouvons développer et résoudre l'intégrale $I$ :

$$\begin{aligned} I &= \frac{a^2}{2} e^{-a|z|} \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left\{ -\frac{(|z| - sa)^2}{2s} \right\} \mathrm{d}s \\ &= \frac{a^2}{2} e^{-a|z|} \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left\{ -\frac{a^2(s - \frac{|z|}{a})^2}{2s} \right\} \mathrm{d}s \\ &= \frac{a^2}{2} e^{-a|z|} \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left\{ -\frac{|z|^2(s - \frac{|z|}{a})^2}{2(\frac{|z|}{a})^2 s} \right\} \mathrm{d}s \\ &= \frac{a^2}{2|z|} e^{-a|z|} \int_0^\infty s \left[ \frac{|z|^2}{2\pi s^3} \right]^{1/2} \exp\left\{ -\frac{|z|^2(s - \frac{|z|}{a})^2}{2(\frac{|z|}{a})^2 s} \right\} \mathrm{d}s \ . \end{aligned}$$

Remarquons que la dernière intégrale correspond à l'espérance d'une loi gaussienne inverse pour $|z| \neq 0$. En effet, nous avons que $S \sim IG(\frac{|z|}{a}, |z|^2)$, alors,

$$I = \frac{a^2}{2|z|} e^{-a|z|} \frac{|z|}{a} = \frac{a}{2} e^{-a|z|} \tag{22}$$

et nous retrouvons la loi de Laplace. Cette représentation de la loi de Laplace est notamment efficace dans les simulations MCMC (Andrews et Mallows 1974). Plus précisément, il s'avère que la loi de Laplace n'est pas conjuguée avec la vraisemblance normale des observations [10]. Nous voulons donc exprimer cette loi *a priori* en faisant intervenir des lois normales, puisque ces lois sont conjuguées pour une vraisemblance normale. La représentation précédente permet donc de plus facilement implémenter l'échantillonneur de Gibbs à travers un mélange de lois normales pour le vecteur $\boldsymbol{\beta}$. Autrement dit, au lieu de générer directement la loi de Laplace $Z \sim \mathcal{L}aplace(0, \frac{1}{a})$, nous pouvons procéder en deux étapes : d'abord en générant la variance $S \sim \mathcal{E}xp(\frac{a^2}{2})$, puis en tirant de la loi conditionnelle $Z|S = s \sim \mathcal{N}(0, s)$. Dans cette décomposition, lorsque nous marginalisons, nous retrouvons exactement la densité de Laplace.

Précédemment, la loi de Laplace indépendante de $\sigma^2$ a été utilisée afin de mettre en évidence le lien entre le MAP et l'estimateur Lasso. Pour l'analyse bayésienne complète, Park et Casella (2008) privilégient plutôt une loi *a priori* conditionnelle à $\sigma^2$, de la forme $\beta_j|\sigma^2 \sim \mathcal{L}aplace(0, \sqrt{\sigma^2}/\lambda)$, qui mène à la formulation hiérarchique utilisée dans l'échantillonneur de Gibbs. Ainsi, nous posons essentiellement $z = \beta_j$ et $a = \frac{\lambda}{\sqrt{\sigma^2}}$. Park et Casella (2008) introduisent aussi une variable latente permettant de contrôler la variance individuelle de chacun des coefficients de régression. Les variables latentes $\tau_j^2$ jouent le rôle de paramètres locaux d'échelle et déterminent conditionnellement la variance $\sigma^2\tau_j^2$

10. Une densité *a priori* est dite conjuguée à une fonction de vraisemblance si la loi *a posteriori* obtenue après application du théorème de Bayes appartient à la même famille de distributions que la loi *a priori*. Plus formellement, pour un paramètre $\theta$, une famille de densités $\mathcal{F}$ sur $\Theta$ est dite conjuguée si $\forall \pi \in \mathcal{F}, \pi(\theta|\mathbf{x}) \in \mathcal{F}$.

de chaque coefficient $\beta_j$. Plus précisément, nous avons que $\beta_j|\tau_j^2, \sigma^2 \sim \mathcal{N}(0, \sigma^2\tau_j^2)$ avec $\tau_j^2 \sim \mathcal{E}xp(\frac{\lambda^2}{2})$. Puis, en marginalisant $\tau_j^2$ et à l'aide de la décomposition (22) de l'intégrale $I$, nous avons

$$\begin{aligned}
\pi(\beta_j|\sigma^2) &= \int_0^\infty \pi(\beta_j|\sigma^2, \tau_j^2)\pi(\tau_j^2)\mathrm{d}\tau_j^2 \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left\{ -\frac{\beta_j^2}{2\sigma^2\tau_j^2} \right\} \frac{\lambda^2}{2} \exp\left\{ -\frac{\lambda^2\tau_j^2}{2} \right\} \mathrm{d}\tau_j^2 \qquad (23) \\
&= \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{ -\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}} \right\} .
\end{aligned}$$

Conditionnellement aux $\tau_j^2$, la loi *a priori* de $\boldsymbol{\beta}$ devient normale, donc conjuguée avec le modèle de régression gaussien, ce qui facilite l'implémentation de l'échantillonneur de Gibbs. Ces variables latentes contrôlent la variance locale de chaque $\beta_j$. Cette décomposition suggère alors la hiérarchie complète suivante à implémenter dans l'algorithme :

$$\mathbf{y}|\alpha, \sigma^2, \boldsymbol{\beta}, \mathbf{X} \sim \mathcal{N}_n(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_{n\times n}) \qquad (24)$$

$$\boldsymbol{\beta}|\tau_1^2, \ldots, \tau_p^2, \sigma^2 \sim \mathcal{N}_p(\mathbf{0}_p, \sigma^2\mathbf{D}_\tau), \quad \mathbf{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2) \qquad (25)$$

$$\tau_j^2 \overset{iid}{\sim} \mathcal{E}xp\left(\frac{\lambda^2}{2}\right), \quad j = 1, \ldots, p \qquad (26)$$

$$\sigma^2 \sim \pi(\sigma^2) \qquad (27)$$

avec $\tau_1^2, \ldots, \tau_p^2 > 0$ et $\sigma^2$ indépendants. On peut attribuer au paramètre $\alpha$ une loi *a priori* plate et indépendante de celle des autres paramètres.

### 3.1.1 Implémentation de l'échantillonneur de Gibbs

En pratique, bien qu'il existe plusieurs choix de distribution *a priori* pour le paramètre $\sigma^2$, nous utilisons une distribution inverse gamma pour celui-ci conjointement à la hiérarchie (27) pour l'implémentation de l'algorithme [11]. Ainsi, supposons que

$$\pi(\sigma^2) = \frac{\gamma^t}{\Gamma(t)}(\sigma^2)^{-t-1}e^{-\gamma/\sigma^2} , \qquad (28)$$

11. Notons que certains symboles sont réutilisés dans le rapport, comme le paramètre $\gamma$ de la loi inverse gamma pour $\sigma^2$. Bien entendu, il ne s'agit pas du même $\gamma$ qu'à l'équation (4).

où $\sigma^2 > 0$ et $t, \gamma > 0$. Alors, en posant $\pi(\alpha) \propto 1$ et en définissant $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_p^2)$, nous obtenons la densité jointe conditionnelle à $\lambda^2$ suivante :

$$\begin{aligned}\pi(\mathbf{y}, \alpha, \boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2|\lambda^2) &\propto f(\mathbf{y}|\alpha, \boldsymbol{\beta}, \sigma^2)\pi(\sigma^2)\pi(\alpha)\prod_{j=1}^{p}\pi(\beta_j|\tau_j^2, \sigma^2)\pi(\tau_j^2) \\ &\propto \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\|_2^2\Big\}\frac{\gamma^t}{\Gamma(t)}(\sigma^2)^{-t-1}e^{-\gamma/\sigma^2} \\ &\quad \times \prod_{j=1}^{p}\frac{1}{(2\pi\sigma^2\tau_j^2)^{1/2}}\exp\Big\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\Big\}\frac{\lambda^2}{2}\exp\Big\{-\frac{\lambda^2\tau_j^2}{2}\Big\}\,.\end{aligned}$$

Il est ensuite assez simple de trouver les densités conditionnelles complètes des différents paramètres. En posant $\tilde{\mathbf{y}} = \mathbf{y} - \overline{\mathbf{y}}\mathbf{1}_n$ et en se rappelant que les colonnes de X sont standardisées, nous avons que $\alpha$ suit une loi normale de moyenne $\overline{\mathbf{y}}$ et de variance $\sigma^2/n$. Une fois le paramètre $\alpha$ intégré, nous avons que la densité jointe conditionnelle est proportionnelle à

$$\begin{aligned}\pi(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2|\lambda^2) &\propto \frac{1}{(\sigma^2)^{(n-1)/2}}\exp\Big\{-\frac{1}{2\sigma^2}\|\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}\|_2^2\Big\}(\sigma^2)^{-t-1}e^{-\gamma/\sigma^2} \\ &\quad \times \prod_{j=1}^{p}\frac{1}{(\sigma^2\tau_j^2)^{1/2}}\exp\Big\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\Big\}\frac{\lambda^2}{2}\exp\Big\{-\frac{\lambda^2\tau_j^2}{2}\Big\}\,.\end{aligned}$$

En posant $\mathbf{A} = \mathbf{X}^\intercal\mathbf{X} + \mathbf{D}_\tau^{-1}$ et en complétant quelques manipulations algébriques simples, nous trouvons que la densité conditionnelle complète du paramètre $\boldsymbol{\beta}$ est une loi normale multidimensionnelle d'espérance $\mathbf{A}^{-1}\mathbf{X}^\intercal\tilde{\mathbf{y}}$ et de variance $\sigma^2\mathbf{A}^{-1}$. La distribution conditionnelle complète de $\sigma^2$ est alors inverse gamma de paramètre de forme $(n-1)/2 + p/2 + t$ et de paramètre d'échelle $(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\intercal(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})/2 + \boldsymbol{\beta}^\intercal\mathbf{D}_\tau^{-1}\boldsymbol{\beta}/2 + \gamma$ car sa loi *a priori* est conjuguée.

Dans le cas des variables $\tau_j^2$, nous pouvons utiliser une reparamétrisation de celles-ci afin de les échantillonner selon une loi inverse gaussienne. En effet, les termes dépendant des $\tau_j^2$ dans la distribution conditionnelle jointe sont proportionnels à

$$(\tau_j^2)^{-1/2}\exp\Big\{-\frac{1}{2}\Big(\frac{\beta_j^2/\sigma^2}{\tau_j^2} + \lambda^2\tau_j^2\Big)\Big\}\,,$$

qui est proportionnel à la densité d'une loi inverse gaussienne réciproque. En posant $\eta_j = 1/\tau_j^2$, la densité de $\eta_j$ est proportionnelle à la densité d'une loi inverse gaussienne. Plus précisément, en posant $F$ comme une fonction de répartition d'une variable aléatoire et pour une certaine valeur $a > 0$, nous avons que

$$F_{\eta_j}(a) = \mathbb{P}\big\{\eta_j < a\big\} = \mathbb{P}\big\{\frac{1}{\tau_j^2} < a\big\} = \mathbb{P}\big\{\frac{1}{a} < \tau_j^2\big\} = 1 - F_{\tau_j^2}(1/a)\,.$$

Puis, en dérivant, nous pouvons trouver le noyau de la fonction de densité de $\eta_j$ et nous obtenons :

$$f_{\eta_j}(a) = \frac{1}{a^2} f_{\tau_j^2}(1/a)$$

Finalement, en remplaçant $a$ par $\eta_j$, nous avons

$$\begin{aligned} f_{\eta_j}(\eta_j) &\propto \left(\frac{1}{\eta_j}\right)^2 (\eta_j)^{1/2} \exp\left\{ -\frac{1}{2}\left(\frac{\beta_j^2}{\sigma^2}\eta_j + \frac{\lambda^2}{\eta_j}\right)\right\} \\ &\propto (\eta_j)^{-3/2} \exp\left\{ -\frac{\beta_j^2\big(\eta_j - \sqrt{\lambda^2\sigma^2/\beta_j^2}\big)^2}{2\sigma^2\eta_j}\right\}, \end{aligned}$$

qui est une paramétrisation de la loi inverse gaussienne. Ainsi, dans notre échantillonneur de Gibbs, nous pouvons simuler les paramètres $\eta_j$, puis appliquer la transformation inverse pour retrouver les $\tau_j^2$. Plus précisément, la distribution de $1/\tau_j^2$ est inverse gaussienne de moyenne $\sqrt{\lambda^2\sigma^2/\beta_j^2}$ et de paramètre d'échelle $\lambda^2$.

Il s'avère que, sous les lois *a priori* (23) et (28), la densité jointe *a posteriori* est proportionnelle à

$$(\sigma^2)^{-(n+p-1)/2-t-1} \exp\left\{ -\frac{1}{\sigma^2}\big( \|\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2 + \gamma\big) - \frac{\lambda}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j|\right\},$$

ce qui suggère que nous pouvons plutôt adopter directement la loi *a priori* impropre invariante d'échelle $\pi(\sigma^2) \propto 1/\sigma^2$, dont le noyau correspond formellement au cas $t = \gamma = 0$. Cette approche permet d'introduire relativement peu d'information *a priori* sur l'échelle $\sigma^2$.

### 3.1.2 Choix du paramètre de pénalisation

Finalement, il reste seulement à identifier une manière d'estimer le paramètre de pénalisation $\lambda$. Park et Casella (2008) proposent notamment l'utilisation d'un algorithme de type espérance-maximisation. Il s'agit d'un algorithme itératif permettant d'estimer les paramètres d'un modèle par maximisation de la vraisemblance (Dempster, Laird et Rubin 1977). Ici, il est utilisé afin de maximiser la vraisemblance marginale par rapport au paramètre $\lambda$. En revanche, il a été observé que le taux de convergence de l'algorithme varie énormément selon le choix initial de la valeur de $\lambda$. Une grande valeur de $\lambda$ amène notamment l'algorithme à être particulièrement lent et lorsqu'il converge relativement rapidement, il est limité par le niveau d'approximation des espérances. Les problèmes inhérents à cet algorithme ne sont pas seulement abordés par Park et Casella (2008), mais aussi par Booth et Hobert (1999) et C. E. McCulloch (1997) dans leurs tentatives de régression dans des modèles linéaires généralisés.

Une autre piste afin d'estimer le paramètre $\lambda$ est de directement le considérer comme une variable aléatoire faisant partie intrinsèque du modèle de régression. En effet, en abordant le paramètre de pénalisation avec une loi *a priori* comme les autres paramètres, nous prenons en compte l'incertitude entourant celui-ci tout en pouvant facilement l'implémenter dans notre échantillonneur de Gibbs conjointement à la hiérarchie explorée précédemment. Park et Casella (2008) rappellent cependant que le choix de cette loi est crucial car certaines peuvent introduire plusieurs modes *a posteriori* ou encore rendre la distribution postérieure non-intégrable. Pour pallier ces problèmes, nous échantillonnons des valeurs de $\lambda^2$ depuis une loi *a priori* gamma. Comme pour les paramètres $\tau_j^2$ précédents, nous pouvons ensuite retransformer la chaîne de valeurs fournies par l'échantillonneur afin de retrouver des valeurs de $\lambda$. Ainsi, nous utilisons la loi *a priori*

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2} \ , \tag{29}$$

où $\lambda^2 > 0$, $r > 0$ et $\delta > 0$. L'utilisation de cette distribution permet d'implémenter facilement le paramètre $\lambda$ dans l'échantillonneur de Gibbs grâce aux propriétés des distributions conjuguées. Aussi, lorsque nous utilisons cette loi dans la hiérarchie (27), le produit des facteurs dépendant de $\lambda$ dans la densité jointe est

$$(\lambda^2)^{p+r-1} \exp\left\{ -\lambda^2\left(\delta + \frac{1}{2}\sum_{j=1}^{p}\tau_j^2\right)\right\}$$

et nous avons donc que la distribution conditionnelle complète de $\lambda^2$ est une loi gamma de paramètre de forme $p+r$ et de paramètre de taux $\delta + \sum_{j=1}^{p}\tau_j^2/2$. Avec cette spécification, $\lambda^2$ peut ainsi être ajouté directement au cycle de Gibbs ; les autres lois conditionnelles conservent la même forme et sont évaluées à la valeur courante de $\lambda^2$. En revanche, Park et Casella (2008) attirent l'attention sur le fait que le paramètre $\delta$ doit être relativement plus grand que 0 afin d'éviter des problèmes computationnels.

Ainsi, nous pouvons interpréter le Lasso comme le maximum *a posteriori* lorsque des lois *a priori* de Laplace indépendantes sont utilisées sur les coefficients de régression. Nous pouvons ensuite obtenir un échantillon de la loi *a posteriori* à l'aide d'un échantillonneur de Gibbs car nous pouvons relativement facilement trouver les distributions conditionnelles complètes des paramètres du modèle. Une fois l'algorithme terminé, nous pouvons construire des estimateurs ponctuels à partir des moyennes ou médianes *a posteriori*.

Il existe aussi une variante plus flexible du Lasso bayésien. Le Lasso bayésien adaptatif, proposé par Leng, Tran et Nott (2014), s'appuie sur des paramètres de pénalisation distincts pour chacun des coefficients de régression. Cette approche permet ainsi d'adapter l'intensité du rétrécissement au niveau d'importance associé à chaque variable et fournit des mécanismes permettant d'obtenir des modèles parcimonieux à partir des modes conditionnels *a posteriori*. Cette idée est notamment inspirée du Lasso adaptatif fréquentiste de Zou (2006), dans lequel des poids de pénalisation distincts sont construits à partir

d'un estimateur préliminaire des coefficients tel que l'estimateur MCO. D'autres approches bayésiennes permettent également d'effectuer directement de la sélection de variables. Par exemple, George et R. E. McCulloch (1993) proposent une méthode de sélection de sous-ensembles reposant sur un modèle hiérarchique et un échantillonneur de Gibbs. Les lois *a priori* de type *Spike and Slab* s'inscrivent dans cette famille d'approches en associant aux coefficients des distributions permettant de distinguer les variables actives des variables inactives.

## 3.2 Loi *a priori Spike and Slab*

Dans un contexte parcimonieux, plusieurs coefficients de $\boldsymbol{\beta}$ sont nuls ou très près de 0. L'idée des lois *a priori Spike and Slab* est donc de modéliser chaque coefficient à l'aide d'un mélange de deux composantes : une composante fortement concentrée autour de zéro et une composante plus diffuse. Pour y arriver, il existe plusieurs postulats sur les deux composantes de la loi *a priori* visant toujours à accomplir le même but : effectuer de la sélection de variables dans un cadre bayésien. Intuitivement, le *spike* représente une forte concentration autour de la valeur 0. Cette partie peut être modélisée de diverses manières. Par exemple, certaines méthodes qui seront explorées utilisent une masse en 0, alors que d'autres ont plutôt recours à des densités de probabilité avec une faible variance autour de 0. Le *slab* est plutôt une loi plus diffuse suivie par les coefficients que nous considérons importants dans le modèle.

### 3.2.1 *Spike and Slab* uniforme

Mitchell et Beauchamp (1988) proposent une formulation du *Spike and Slab* dans laquelle chaque coefficient susceptible d'être exclu reçoit une loi *a priori* composée d'une masse ponctuelle en zéro et d'une densité uniforme diffuse. Cette construction permet d'attribuer une probabilité *a posteriori* à chacun des $2^p$ sous-modèles obtenus en incluant ou en excluant les différents prédicteurs. Une règle naturelle de sélection consiste alors à retenir le sous-modèle ayant la plus grande probabilité *a posteriori*. La méthode permet également de calculer les probabilités marginales *a posteriori* d'inclusion ou d'exclusion de chacune des variables.

Dans cette approche, quelques hypothèses sont postulées. En particulier, nous supposons que $\mathbf{X}$ est de rang plein et que $p < n$. Nous pouvons remarquer que nous devons essentiellement considérer les $2^p$ sous-modèles obtenus en incluant ou en excluant chacun des $p$ prédicteurs. Pour la suite, notons $\mathbf{X}_m$ la matrice de design de dimension $n \times p_m$ dont les colonnes correspondent aux covariables incluses dans le $m$-ième sous-modèle $A_m$ pour $m = 1, 2, \ldots, 2^p$. Notons aussi l'estimateur MCO d'un sous-modèle comme

$$\hat{\boldsymbol{\beta}}_m^0 = (\mathbf{X}_m^\intercal \mathbf{X}_m)^{-1} \mathbf{X}_m^\intercal \mathbf{y} \ ,$$

ainsi que sa somme des carrés des résidus comme

$$S_m^2 = (\mathbf{y} - \mathbf{X}_m \hat{\boldsymbol{\beta}}_m^0)^\intercal (\mathbf{y} - \mathbf{X}_m \hat{\boldsymbol{\beta}}_m^0) \ .$$

Considérons le modèle (3) où la variable réponse est centrée, la matrice $\mathbf{X}$ est standardisée et les erreurs sont gaussiennes indépendantes. À titre de rappel, nous avons donc que $\mathbf{y}$ suit une loi normale multidimensionnelle d'espérance $\mathbf{X}\boldsymbol{\beta}$ et de variance $\sigma^2\mathbf{I}$. Mitchell et Beauchamp (1988) utilisent aussi la loi *a priori* non informative $\pi(\sigma) \propto \frac{1}{\sigma}$. Ils supposent aussi que la distribution *a priori* de $\boldsymbol{\beta}$ est indépendante de $\sigma^2$ et que les composantes de $\boldsymbol{\beta}$ sont mutuellement indépendantes et suivent chacune une distribution *Spike and Slab*. Plus précisément, nous attribuons une loi uniforme à chacun des $\beta_j$ sur l'intervalle $[-f_j, f_j]$, avec une partie de la masse de probabilité centrée à 0 si la variable $x_j$ est considérée comme susceptible d'être exclue du modèle. Formellement, nous posons que

$$\begin{aligned} \mathbb{P}\big\{\beta_j = 0\big\} &= h_{0j} \ , \\ \mathbb{P}\big\{\beta_j < b, \beta_j \neq 0\big\} &= (b + f_j) h_{1j} \ , \quad b \in (-f_j, f_j) \ , \\ \mathbb{P}\big\{|\beta_j| > f_j\big\} &= 0 \ , \end{aligned}$$

où $h_{0j} \geq 0, h_{1j} > 0$ et $h_{0j} + 2h_{1j}f_j = 1$. Intuitivement, $h_{0j}$ représente la probabilité *a priori* que le $j$-ième coefficient soit nul. Si $h_{0j}$ est grand, nous croyons fortement que la variable $j$ ne devrait pas être dans le modèle. Dans le cas contraire, si $h_{0j} = 0$, aucune masse ponctuelle en zéro n'est attribuée à $\beta_j$ *a priori*, de sorte à ce que la variable correspondante soit nécessairement incluse dans le modèle. Cette partie représente donc une masse de probabilité ponctuelle. Le terme $h_{1j}$ est plutôt la hauteur de la partie *slab* uniforme sur $[-f_j, f_j]$. Ainsi, la probabilité totale de la loi *a priori* doit être la somme de la partie *spike* et de l'aire du rectangle uniforme. De manière concise, nous pouvons écrire la loi *a priori* de chacun des coefficients comme

$$\pi(\beta_j) = h_{0j}\delta_0(\beta_j) + h_{1j}\mathbb{1}_{[-f_j, f_j]}(\beta_j) \ ,$$

où $h_{0j} + 2h_{1j}f_j = 1$. La figure 4 présente un exemple de loi *a priori Spike and Slab* uniforme pour un coefficient $\beta_j$ quelconque.

Mitchell et Beauchamp (1988) utilisent deux paramètres pour la distribution afin de paramétrer cette loi *a priori*. En effet, la distribution peut être entièrement caractérisée par les paramètres

$$\varphi_j = \frac{h_{0j}}{h_{1j}} = \frac{2h_{0j}f_j}{1 - h_{0j}}$$

et $f_j$. Ainsi, $\varphi_j$ est le rapport de la masse de probabilité du *spike* sur la hauteur du *slab*. En posant $\mathcal{S}_m = \big\{j : \beta_j \neq 0 \text{ dans } A_m\big\}$ comme étant l'ensemble d'indices des termes inclus dans le sous-modèle $A_m$ et $\mathcal{S}_m^c$ l'ensemble d'indices des termes omis, la distribution *a priori* pour un sous-modèle est

$$\mathbb{P}\big\{A_m\big\} = \prod_{j \in \mathcal{S}_m^c} h_{0j} \prod_{j \in \mathcal{S}_m} (2f_j h_{1j}) = \prod_{j \in \mathcal{S}_m^c} \varphi_j \prod_{j \in \mathcal{S}_m} (2f_j) \prod_{j=1}^{p} (\varphi_j + 2f_j)^{-1} \ .$$

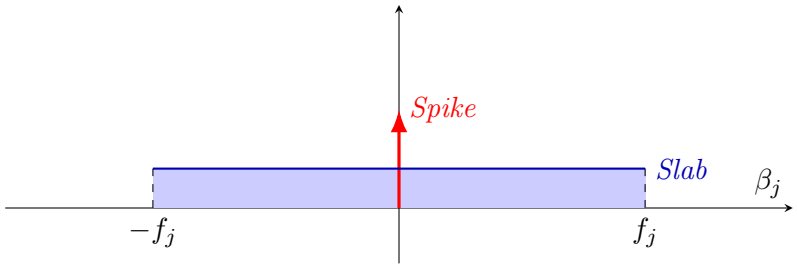


FIGURE 4 – Représentation de la loi *a priori Spike and Slab* d'un coefficient de régression $\beta_j$.
Dans cette figure, la partie en rouge est une masse en 0 de 0.5 représentant le *spike* de la loi, tandis que la partie bleue est le *slab* uniforme d'une hauteur $h_{1j} = 0.2$. Ainsi, nous avons que $f_j = 1.25$.

Sous les spécifications précédentes, nous avons que la fonction de densité de $\mathbf{y}$, sachant les autres paramètres, est

$$f(\mathbf{y}|A_m, \boldsymbol{\beta}_m, \sigma) = (2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\big(S_m^2 + (\boldsymbol{\beta}_m - \hat{\boldsymbol{\beta}}_m^0)^\intercal \mathbf{X}_m^\intercal \mathbf{X}_m(\boldsymbol{\beta}_m - \hat{\boldsymbol{\beta}}_m^0)\big)\right\}.$$

En multipliant cette vraisemblance par la loi *a priori* conditionnelle de $\boldsymbol{\beta}_m$, puis en intégrant par rapport à $\boldsymbol{\beta}_m$, nous obtenons la densité marginale de $\mathbf{y}$ conditionnellement à $A_m$ et à $\sigma$. Plus précisément, MITCHELL et BEAUCHAMP (1988) supposent les $f_j$ suffisamment grands pour que les bornes d'intégration puissent être remplacées, avec une erreur négligeable, par $\mathbb{R}^{p_m}$. Nous intégrons ensuite par rapport à $\sigma$, puis appliquons le théorème de Bayes afin d'obtenir la probabilité *a posteriori* de chacun des sous-modèles. Plus précisément, nous avons que

$$\mathbb{P}\{A_m|\mathbf{y}\} \propto \left(\prod_{j\in\mathcal{S}_m^c}\varphi_j\right)\pi^{p_m/2}\Gamma\left(\frac{n-p_m}{2}\right)\left|\mathbf{X}_m^\intercal\mathbf{X}_m\right|^{-1/2}(S_m^2)^{-(n-p_m)/2},$$

où $p_m$ est le nombre de termes dans le sous-modèle $A_m$ et où nous pouvons remplacer le déterminant de la matrice $\mathbf{X}_m^\intercal\mathbf{X}_m$ par 1 dans le cas du modèle nul. Supposons maintenant que les coefficients correspondant aux prédicteurs susceptibles d'être exclus possèdent des lois *a priori* identiques. En particulier, les paramètres $\varphi_j$ prennent alors une valeur commune $\varphi$. Dans ce cas, nous avons que

$$\mathbb{P}\{A_m|\mathbf{y}\} \propto \varphi^{p-p_m}\pi^{p_m/2}\Gamma\left(\frac{n-p_m}{2}\right)\left|\mathbf{X}_m^\intercal\mathbf{X}_m\right|^{-1/2}(S_m^2)^{-(n-p_m)/2}.$$

Cette formule représente bien comment la taille du modèle est contrôlée. En effet, un modèle avec un petit $S_m^2$ est favorisé car il ajuste bien les données, tandis que $\varphi$ contrôle en quelque sorte l'importance *a priori* donnée aux exclusions. À $f_j$ fixé, une augmentation de $\varphi$ accroît l'importance *a priori* accordée à l'exclusion des variables et favorise ainsi les modèles comportant moins de prédicteurs.

Nous pouvons facilement calculer ces probabilités en normalisant les poids associés aux différents sous-modèles :

$$\mathbb{P}\big\{A_m|\mathbf{y}\big\} = \frac{w_m}{\sum_{u=1}^{2^p} w_u} \ ,$$

où le logarithme du poids est donné par

$$\begin{aligned}\ln(w_m) = p_m\big(-\ln(\varphi) + \frac{1}{2}\ln(\pi)\big) + \ln\left(\Gamma\big(\frac{n-p_m}{2}\big)\right) \\ - \frac{1}{2}\ln\big|\mathbf{X}_m^\intercal\mathbf{X}_m\big| - \frac{1}{2}(n-p_m)\ln(S_m^2) \ ,\end{aligned}$$

à une constante additive commune à tous les modèles près. Ces probabilités sont notamment utiles afin de calculer les probabilités *a posteriori* qu'un certain coefficient de régression soit nul en sommant les probabilités des sous-modèles dans lesquels le coefficient est nul :

$$\mathbb{P}\big\{\beta_j = 0|\mathbf{y}\big\} = \sum_{m:j\notin\mathcal{S}_m} \mathbb{P}\big\{A_m|\mathbf{y}\big\} \ .$$

Nous pouvons également calculer l'espérance du nombre de termes dans le modèle de la manière suivante :

$$\mathbb{E}\big[p_m|\mathbf{y}\big] = \sum_m p_m\mathbb{P}\big\{A_m|\mathbf{y}\big\} \ .$$

Afin de sélectionner un modèle, l'idée est donc de sélectionner le sous-modèle avec la plus grande probabilité *a posteriori*, ce qui correspond à une règle de sélection MAP sur l'espace des modèles. Une autre possibilité consiste à considérer les probabilités marginales d'inclusion *a posteriori*. Barbieri et Berger (2004) proposent à cet effet un modèle de probabilité médiane, qui contient les variables dont la probabilité *a posteriori* d'inclusion est supérieure ou égale à $1/2$. Ainsi, dans notre notation, une variable $j$ peut être incluse lorsque $\mathbb{P}\{\beta_j = 0|\mathbf{y}\} \leq 1/2$. Sous certaines conditions, les auteurs montrent que ce modèle possède des propriétés d'optimalité pour la prédiction sous une perte quadratique.

Ensuite, Mitchell et Beauchamp (1988) n'attribuent pas de loi *a priori* au paramètre commun $\varphi$. Ils le considèrent plutôt comme un paramètre ajustable. Ils proposent notamment d'étudier son influence à l'aide d'une procédure de validation croisée bayésienne de type *leave-one-out*. Pour chaque observation, une distribution prédictive est construite à partir des autres observations, puis sa qualité prédictive est évaluée. Les auteurs proposent également d'examiner une mesure d'adéquation du modèle. Ces diagnostics permettent ainsi de guider le choix de $\varphi$ plutôt que de lui attribuer directement une loi *a priori*.

L'estimation des coefficients se fait ensuite par une moyenne sur l'incertitude de sélection. Conditionnellement à un modèle donné $A_m$, seuls les coefficients correspondant aux variables incluses doivent être estimés. Mitchell et Beauchamp (1988) montrent alors que leur loi *a posteriori* est une loi de Student multivariée centrée sur l'estimateur MCO du sous-modèle $\hat{\boldsymbol{\beta}}_m^0$ à $n-p_m$ degrés de liberté. Autrement dit, conditionnellement au sous-modèle, $\boldsymbol{\beta}_m|A_m,\mathbf{y}$ possède une distribution centrée en $\hat{\boldsymbol{\beta}}_m^0$, tandis que tous les coefficients

exclus sont exactement 0. Ainsi, nous pouvons déterminer la distribution *a posteriori* de $\boldsymbol{\beta}$ en moyennant sur l'incertitude de sélection du modèle :

$$\pi(\boldsymbol{\beta}|\mathbf{y}) = \sum_m \mathbb{P}\big\{A_m|\mathbf{y}\big\}\pi(\boldsymbol{\beta}|A_m,\mathbf{y}) \ .$$

Pour un coefficient en particulier, nous pouvons donc avoir une estimation ponctuelle par son espérance

$$\mathbb{E}\big[\beta_j|\mathbf{y}\big] = \sum_{m:j\in\mathcal{S}_m} \mathbb{P}\big\{A_m|\mathbf{y}\big\}\mathbb{E}\big[\beta_j|A_m,\mathbf{y}\big] \ ,$$

puisque les modèles ne contenant pas $j$ contribuent exactement zéro. Avec leurs postérieures conditionnelles centrées sur les estimateurs MCO, cela revient essentiellement à une moyenne des estimations MCO des différents sous-modèles, pondérée par leurs probabilités *a posteriori*. Ainsi, Mitchell et Beauchamp (1988) donnent d'abord une distribution de probabilité sur l'ensemble des modèles, puis une distribution *a posteriori* des coefficients à l'intérieur de chaque modèle.

En somme, cette structure de loi *a priori* peut être reformulée à l'aide d'un indicateur latent

$$Z_j = \begin{cases} 0, & \text{si la variable } j \text{ est exclue,} \\ 1, & \text{si la variable } j \text{ est incluse .} \end{cases}$$

On obtient alors naturellement des probabilités d'inclusion *a posteriori*, des probabilités *a posteriori* sur les sous-modèles et nous pouvons élaborer une procédure de sélection de variables fondée sur ces probabilités. Le cadre du *Spike and Slab* uniforme réalise donc un compromis entre la parcimonie du modèle et sa flexibilité.

### 3.2.2 *Spike and Slab* gaussien

Une autre approche des lois *a priori Spike and Slab* consiste à utiliser un mélange de lois normales pour les coefficients $\beta_j$ afin d'effectuer une sélection de variables. Plus précisément, afin d'éviter l'énumération exhaustive des $2^p$ sous-modèles possibles, George et R. E. McCulloch (1993) ont développé une procédure appelée *stochastic search variable selection* (SSVS). Cette méthode repose sur un mélange de lois normales et sur l'introduction de variables latentes permettant d'indiquer si chaque coefficient est associé à la composante *spike* ou à la composante *slab*. Un échantillonneur de Gibbs est alors utilisé afin d'explorer la loi *a posteriori* des paramètres et des indicateurs de sélection. Les sous-modèles prometteurs peuvent ainsi être identifiés à partir de leur fréquence d'apparition dans l'échantillon obtenu, laquelle fournit une approximation de leur probabilité *a posteriori*.

Considérons toujours notre modèle de régression (3), où $\boldsymbol{\beta}$ et $\sigma^2$ sont inconnus. Désormais, nous ne supposons plus que les coefficients de régression possèdent une masse de probabilité en 0. Nous considérons plutôt un contexte dans lequel chaque coefficient $\beta_j$ suit un mélange de deux lois normales centrées en 0, dont l'une possède une faible variance

et l'autre une variance plus élevée. En introduisant une variable latente $\varrho_j \in \{0, 1\}$, nous pouvons représenter ce mélange de lois normales comme

$$\beta_j|\varrho_j \sim (1 - \varrho_j)\mathcal{N}(0, \tau_j^2) + \varrho_j\mathcal{N}(0, c_j^2\tau_j^2), \tag{30}$$

où

$$\mathbb{P}\{\varrho_j = 1\} = 1 - \mathbb{P}\{\varrho_j = 0\} = q_j.$$

Nous pouvons interpréter $q_j$ comme la probabilité *a priori* que le coefficient soit associé à la composante *slab* et soit ainsi inclus dans le modèle. Lorsque $\varrho_j = 0$, nous avons

$$\beta_j \sim \mathcal{N}(0, \tau_j^2),$$

tandis que lorsque $\varrho_j = 1$,

$$\beta_j \sim \mathcal{N}(0, c_j^2\tau_j^2).$$

Nous choisissons $\tau_j > 0$ suffisamment petit pour que, lorsque $\varrho_j = 0$, le coefficient $\beta_j$ soit fortement concentré autour de 0 et puisse être considéré comme négligeable en pratique. Nous choisissons ensuite $c_j > 1$ suffisamment grand pour que, lorsque $\varrho_j = 1$, la loi *a priori* soit plus diffuse et permette à $\beta_j$ de prendre des valeurs substantiellement différentes de 0. La figure 5 permet de visualiser ce mélange de lois normales. Afin d'obtenir les lois

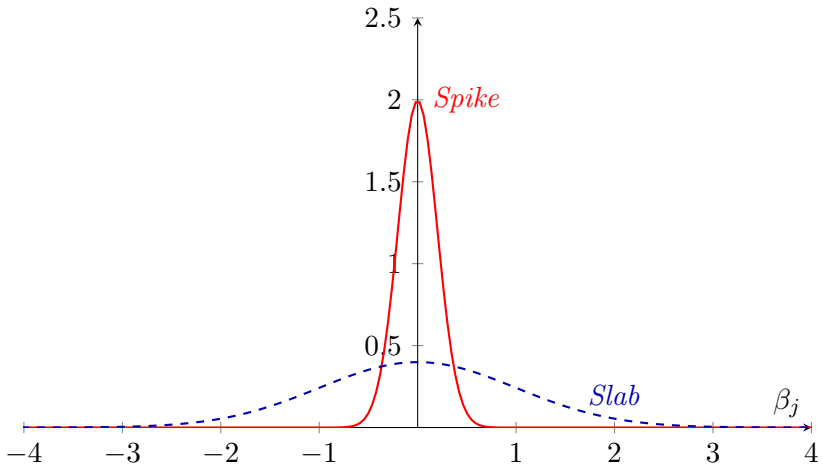


FIGURE 5 – Représentation du mélange de lois normales pour une loi *a priori* de type *Spike and Slab*
Dans cette figure, la partie en rouge est la loi normale représentant le *spike* de la loi *a priori* avec une variance $\tau_j^2 = 0.04$. La partie en bleu représente le *slab* de la fonction avec une variance de $c_j^2\tau_j^2 = 1$. Ainsi, nous avons que $c_j^2 = 25$ dans cet exemple.

marginales présentées dans (30), GEORGE et R. E. MCCULLOCH (1993) considèrent une loi *a priori* normale multidimensionnelle pour le vecteur $\boldsymbol{\beta}$. Plus précisément,

$$\boldsymbol{\beta}|\boldsymbol{\varrho} \sim \mathcal{N}_p\big(\mathbf{0}, \mathbf{D}_{\boldsymbol{\varrho}}\mathbf{R}\mathbf{D}_{\boldsymbol{\varrho}}\big),$$

où $\boldsymbol{\varrho} = (\varrho_1, \ldots, \varrho_p)$, $\mathbf{R}$ est une matrice de corrélation *a priori* et

$$\mathbf{D}_{\boldsymbol{\varrho}} = \text{diag}\left(a_1\tau_1, \ldots, a_p\tau_p\right),$$

avec

$$a_j = \begin{cases} 1, & \text{si } \varrho_j = 0, \\ c_j, & \text{si } \varrho_j = 1\ . \end{cases}$$

Nous devons également spécifier une loi *a priori* pour le vecteur d'indicateurs $\boldsymbol{\varrho}$. Puisque chaque $\varrho_j$ peut prendre les valeurs 0 ou 1, le vecteur $\boldsymbol{\varrho}$ possède $2^p$ configurations possibles. George et R. E. McCulloch (1993) définissent ainsi une loi discrète $f(\boldsymbol{\varrho})$ sur l'ensemble $\{0,1\}^p$, où $f(\boldsymbol{\varrho})$ représente la probabilité *a priori* attribuée au sous-modèle identifié par $\boldsymbol{\varrho}$.

Un choix simple consiste à supposer les variables $\varrho_j$ indépendantes avec

$$\varrho_j \sim \text{Bernoulli}(q_j).$$

Dans ce cas, la loi jointe de $\boldsymbol{\varrho}$ est donnée par

$$f(\boldsymbol{\varrho}) = \prod_{j=1}^{p} q_j^{\varrho_j}(1-q_j)^{1-\varrho_j}.$$

Lorsque les probabilités $q_j$ sont toutes égales à $1/2$, les $2^p$ configurations reçoivent la même probabilité *a priori*. D'autres choix de $f(\boldsymbol{\varrho})$ permettent toutefois d'accorder davantage de poids à certains sous-modèles. Conjointement à $f(\boldsymbol{\varrho})$, la loi *a priori* marginale de $\boldsymbol{\beta}$ devient ainsi un mélange fini de lois normales multidimensionnelles.

Finalement, George et R. E. McCulloch (1993) complètent leur modèle hiérarchique en attribuant une loi *a priori* gamma inverse à la variance $\sigma^2$. Plus précisément, ils posent

$$\sigma^2|\boldsymbol{\varrho} \sim \Gamma^{-1}\left(\frac{\nu_{\boldsymbol{\varrho}}}{2}, \frac{\nu_{\boldsymbol{\varrho}}\lambda_{\boldsymbol{\varrho}}}{2}\right).$$

Les paramètres $\nu_{\boldsymbol{\varrho}}$ et $\lambda_{\boldsymbol{\varrho}}$ peuvent ainsi dépendre de la configuration $\boldsymbol{\varrho}$ considérée. La structure hiérarchique complète permet ainsi d'écrire la loi *a posteriori* jointe sous la forme

$$\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\varrho}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}|\boldsymbol{\varrho})\pi(\sigma^2|\boldsymbol{\varrho})f(\boldsymbol{\varrho}).$$

Cette formulation permet ensuite d'utiliser un échantillonneur de Gibbs afin d'explorer les différentes configurations de $\boldsymbol{\varrho}$ et d'identifier les sous-modèles possédant les probabilités *a posteriori* les plus élevées.

Le choix des différents paramètres de la hiérarchie influence évidemment la capacité de la méthode à distinguer les variables importantes de celles qui devraient être exclues. George et R. E. McCulloch (1993) considèrent principalement $\tau_j$ et $c_j$ comme des paramètres d'ajustement. La valeur de $\tau_j$ doit être suffisamment petite pour que les coefficients associés

au *spike* soient fortement concentrés autour de 0, tandis que $c_j > 1$ doit être suffisamment grand pour que la composante *slab* permette des coefficients substantiellement différents de 0. Les auteurs proposent notamment d'étudier différentes valeurs de ces paramètres afin d'évaluer la sensibilité de la procédure.

La loi *a priori* $f(\boldsymbol{\varrho})$ permet quant à elle de contrôler les probabilités attribuées aux différents sous-modèles. Par exemple, lorsque les variables $\varrho_j$ sont indépendantes avec $\varrho_j \sim \text{Bernoulli}(q_j)$, le choix $q_j = 1/2$ pour tout $j$ attribue la même probabilité *a priori* aux $2^p$ configurations possibles. D'autres choix permettent plutôt de favoriser certains sous-modèles, notamment des modèles plus parcimonieux. Finalement, la matrice $\mathbf{R}$ contrôle la structure de corrélation *a priori* entre les coefficients. Un choix simple consiste à poser $\mathbf{R} = \mathbf{I}$, ce qui correspond à des coefficients indépendants conditionnellement à $\boldsymbol{\varrho}$, alors qu'une matrice tenant compte de la structure de $\mathbf{X}$ peut également être utilisée.

Une fois ces paramètres spécifiés, il n'est pas nécessaire de calculer explicitement les probabilités *a posteriori* des $2^p$ sous-modèles. À l'instar de la procédure utilisée précédemment pour le Lasso bayésien, les lois conditionnelles complètes peuvent être obtenues en conservant successivement les termes qui dépendent du paramètre à échantillonner. Sous la hiérarchie précédente, la loi conditionnelle de $\boldsymbol{\beta}$ est une loi normale multidimensionnelle, celle de $\sigma^2$ est une loi gamma inverse et chacun des indicateurs $\varrho_j$ peut être échantillonné à partir d'une loi de Bernoulli conditionnelle. Un échantillonneur de Gibbs permet ainsi de générer successivement $\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}, \boldsymbol{\varrho}^{(t)}$ et d'explorer les différents sous-modèles sans devoir énumérer directement l'ensemble des $2^p$ possibilités. Les configurations de $\boldsymbol{\varrho}$ apparaissant fréquemment dans l'échantillon correspondent alors aux sous-modèles possédant une probabilité *a posteriori* élevée.

L'échantillonneur de Gibbs permet simultanément d'effectuer la sélection des variables et l'estimation des coefficients de régression. En effet, conditionnellement aux valeurs courantes de $\boldsymbol{\varrho}$ et de $\sigma^2$, la loi *a posteriori* de $\boldsymbol{\beta}$ est une loi normale multidimensionnelle. À chaque itération, une nouvelle valeur $\boldsymbol{\beta}^{(t)}$ est donc générée à partir de cette loi conditionnelle.

Après avoir retiré les premières itérations de la chaîne, une estimation ponctuelle des coefficients peut être obtenue à partir de leur moyenne *a posteriori*,

$$\hat{\beta}_j = \mathbb{E}\big[\beta_j | \mathbf{y}\big] \approx \frac{1}{T} \sum_{t=1}^{T} \beta_j^{(t)}.$$

Cette moyenne est calculée sur l'ensemble des configurations de $\boldsymbol{\varrho}$ visitées par l'échantillonneur de Gibbs et ne correspond donc pas nécessairement à un sous-modèle particulier. Il faut ainsi distinguer l'estimation des coefficients de la sélection d'un modèle.

La sélection peut être effectuée directement à partir de l'échantillon de $\boldsymbol{\varrho}$. Pour une configuration $\mathbf{r} \in \{0,1\}^p$, sa probabilité *a posteriori* peut être estimée par sa fréquence d'apparition dans la chaîne,

$$\mathbb{P}\big\{\boldsymbol{\varrho} = \mathbf{r} | \mathbf{y}\big\} \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\big\{\boldsymbol{\varrho}^{(t)} = \mathbf{r}\big\}.$$

Les configurations apparaissant le plus fréquemment correspondent ainsi aux sous-modèles possédant les probabilités *a posteriori* les plus élevées. Si nous désirons retenir un seul sous-modèle, nous pouvons considérer la configuration possédant la plus grande probabilité *a posteriori* estimée. En pratique, $\hat{\boldsymbol{\varrho}}$ correspond donc à la configuration apparaissant le plus fréquemment dans l'échantillon de Gibbs. Les prédicteurs pour lesquels $\hat{\varrho}_j = 1$ sont alors considérés comme inclus dans le modèle retenu, tandis que ceux pour lesquels $\hat{\varrho}_j = 0$ sont considérés comme exclus.

Contrairement au *spike* ponctuel présenté précédemment, $\varrho_j = 0$ n'impose toutefois pas que $\beta_j$ soit exactement nul. À une itération donnée, cette configuration entraîne plutôt un fort rétrécissement du coefficient vers 0 par l'intermédiaire de la faible variance $\tau_j^2$, alors que $\varrho_j = 1$ associe le coefficient à la composante *slab*. Les valeurs générées de $\beta_j$ servent ainsi à estimer l'amplitude des effets, tandis que les configurations générées de $\boldsymbol{\varrho}$ permettent d'effectuer la sélection des variables.

Cette formulation du *Spike and Slab* permet donc de remplacer la masse ponctuelle en 0 utilisée précédemment par une composante continue fortement concentrée autour de 0. Elle conserve néanmoins une séparation explicite entre les coefficients considérés comme négligeables et ceux pouvant prendre des valeurs plus importantes. Des travaux ultérieurs ont cherché à rendre cette séparation plus flexible en introduisant des structures hiérarchiques supplémentaires sur les variances associées aux coefficients.

Ishwaran et Rao (2005) proposent notamment une généralisation fondée sur des modèles *rescaled Spike and Slab*. Plutôt que de fixer entièrement les variances responsables du rétrécissement des coefficients, leur approche accorde une importance particulière aux hypervariances et à leur distribution. Les auteurs étudient notamment l'utilisation de lois continues bimodales pour ces hypervariances, de manière à obtenir deux comportements distincts : une forte contraction vers 0 pour certaines variables et un rétrécissement beaucoup plus faible pour les variables importantes. Cette propriété, appelée *selective shrinkage*, permet d'adapter l'intensité du rétrécissement à chaque coefficient.

Un autre élément important de leur approche est la mise à l'échelle du modèle. Ishwaran et Rao (2005) montrent que cette transformation modifie le comportement de la moyenne *a posteriori* et permet d'établir un lien avec certaines méthodes fréquentistes. Leur formulation met ainsi en évidence l'importance du choix de la structure hiérarchique de la loi *a priori*, puisque celle-ci détermine directement l'intensité du rétrécissement appliqué aux différents coefficients.

### 3.2.3 Modélisation de la parcimonie à travers la taille du support

Une autre méthode de sélection de variables consiste à choisir la taille du support puis les variables actives et inactives. Considérons le modèle (3) avec des erreurs normales. Castillo, Schmidt-Hieber et Van der Vaart (2015) s'intéressent d'emblée au contexte parcimonieux dans lequel $n \leq p$ et où plusieurs coefficients du paramètre de régression sont

nuls. Leur méthodologie repose sur la structure *a priori* suivante :

$$(\mathcal{S}, \boldsymbol{\beta}) \longmapsto \pi_p(|\mathcal{S}|) \frac{1}{\binom{p}{|\mathcal{S}|}} g_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}}) \delta_0(\boldsymbol{\beta}_{\mathcal{S}^c}) \ ,$$

où $\delta_0(\boldsymbol{\beta}_{\mathcal{S}^c})$ désigne une masse de Dirac en $\mathbf{0}$ sur les coordonnées inactives, c'est-à-dire que $\beta_j = 0$ pour tout $j \in \mathcal{S}^c$. Plus précisément, leur modélisation de la loi *a priori* consiste d'abord à choisir la dimension $s$ d'une loi *a priori* $\pi_p$ sur l'ensemble $\{0, \dots, p\}$, puis choisir aléatoirement un sous-ensemble $\mathcal{S} \subset \{1, 2, \dots, p\}$ de cardinalité $|\mathcal{S}| = s$ pour finalement choisir un vecteur de coefficients non nuls $\boldsymbol{\beta}_{\mathcal{S}} \equiv \{\beta_j : j \in \mathcal{S}\}$ d'une densité *a priori* $g_{\mathcal{S}}$ sur $\mathbb{R}^{\mathcal{S}}$. Ainsi, aux coordonnées actives, les auteurs proposent d'attribuer une densité continue $g_{\mathcal{S}}$ et de mettre une masse en 0 pour les variables inactives. Cette écriture distingue clairement la parcimonie, contrôlée par $\pi_p$, et la modélisation des coefficients actifs, contrôlée par $g_{\mathcal{S}}$.

Un choix fréquent consiste à prendre $g_{\mathcal{S}}$ sous forme produit,

$$g_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}}) = \prod_{j \in \mathcal{S}} g(\beta_j),$$

où $g$ est une densité continue sur $\mathbb{R}$. Une densité de Laplace constitue un exemple classique d'un tel choix. Dans ce contexte, la composante *spike* correspond à une masse ponctuelle en 0, alors que la composante *slab* correspond à une densité continue sur les coefficients actifs, par exemple une densité de Laplace. La figure 6 montre un exemple de cette loi *a priori*.

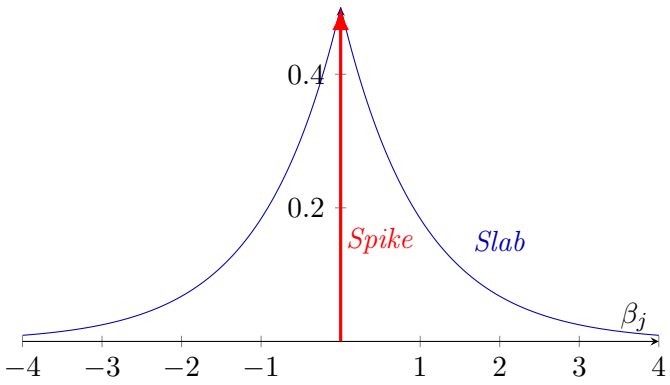


FIGURE 6 – Illustration schématique d'une loi *Spike and Slab* avec densité de Laplace
La composante *spike* est une masse ponctuelle en 0, tandis que la composante *slab* est représentée ici par une densité de Laplace de paramètre $(0, 1)$.

Pour compléter la hiérarchie, nous avons besoin de spécifier la loi *a priori* $\pi_p(s)$ pour $s = 0, 1, \dots, p$. Cette loi est essentielle au bon fonctionnement de cette structure car elle permet la sélection de variables. Dans un contexte parcimonieux, CASTILLO, SCHMIDT-HIEBER et VAN DER VAART (2015) cherchent une loi favorisant les plus petits modèles et

qui évite de répartir trop de masse sur les modèles trop complexes. Les auteurs montrent qu'il est avantageux de choisir une loi $\pi_p$ qui favorise suffisamment les petits supports, en imposant une décroissance exponentielle de la masse *a priori* lorsque la taille du support augmente.

Le document de Castillo, Schmidt-Hieber et Van der Vaart (2015) propose plusieurs résultats d'optimalité reposant notamment sur des conditions de compatibilité et sur le comportement des plus petites valeurs singulières de la matrice de design lorsque celle-ci est restreinte à des vecteurs parcimonieux. Les auteurs définissent la compatibilité d'un modèle comme

$$\phi(\mathcal{S}) = \inf \left\{ \frac{\|\mathbf{X}\boldsymbol{\beta}\|_2 \, |\mathcal{S}|^{1/2}}{\|\mathbf{X}\|_{\max} \|\boldsymbol{\beta}_{\mathcal{S}}\|_1} : \|\boldsymbol{\beta}_{\mathcal{S}^c}\|_1 \leq 7 \|\boldsymbol{\beta}_{\mathcal{S}}\|_1 , \boldsymbol{\beta}_{\mathcal{S}} \neq \mathbf{0} \right\},$$

où $\|\mathbf{X}\|_{\max} = \max_{1 \leq j \leq p} \|\mathbf{x}_j\|_2$.

La condition de compatibilité permet essentiellement de vérifier que la matrice de design contient suffisamment d'information pour distinguer les différents vecteurs parcimonieux. En effet, lorsque $p > n$, la matrice $\mathbf{X}^\intercal\mathbf{X}$ n'est pas inversible et il est impossible d'identifier de manière unique un vecteur quelconque de $\mathbb{R}^p$ à partir de $\mathbf{X}\boldsymbol{\beta}$. En revanche, dans un contexte parcimonieux, nous ne cherchons pas à distinguer tous les vecteurs de $\mathbb{R}^p$, mais principalement ceux dont les coefficients non nuls sont concentrés sur un petit nombre de coordonnées. La compatibilité impose donc une forme d'inversibilité locale de la matrice de design.

Dans la définition précédente, la contrainte

$$\|\boldsymbol{\beta}_{\mathcal{S}^c}\|_1 \leq 7 \|\boldsymbol{\beta}_{\mathcal{S}}\|_1$$

restreint l'étude aux vecteurs dont la majorité de la norme $\ell_1$ est concentrée sur les coordonnées de $\mathcal{S}$. La constante 7 n'a pas de signification particulière et est simplement utilisée par les auteurs comme une constante numérique fixe. Lorsque $\phi(\mathcal{S}) > 0$, le modèle est compatible et nous avons alors

$$\|\mathbf{X}\boldsymbol{\beta}\|_2 \geq \phi(\mathcal{S}) \|\mathbf{X}\|_{\max} \frac{\|\boldsymbol{\beta}_{\mathcal{S}}\|_1}{\sqrt{|\mathcal{S}|}}.$$

Ainsi, un vecteur possédant des coefficients importants sur $\mathcal{S}$ ne peut pas être presque entièrement annulé par la transformation $\mathbf{X}\boldsymbol{\beta}$. Autrement dit, deux vecteurs suffisamment différents sur un petit ensemble de variables ne peuvent pas produire essentiellement les mêmes prédictions. Cette propriété est importante puisque la vraisemblance permet directement d'obtenir de l'information sur $\mathbf{X}\boldsymbol{\beta}$, alors que l'objectif est aussi de récupérer $\boldsymbol{\beta}$. La condition de compatibilité permet donc de transférer une bonne estimation de la composante prédictive vers une bonne estimation des coefficients.

Castillo, Schmidt-Hieber et Van der Vaart (2015) introduisent aussi des conditions uniformes sur la matrice de design afin d'étudier simultanément tous les vecteurs

suffisamment parcimonieux. En particulier, ils définissent la plus petite valeur singulière parcimonieuse normalisée par

$$\widetilde{\phi}(s) = \inf\left\{ \frac{\|\mathbf{X}\boldsymbol{\beta}\|_2}{\|\mathbf{X}\|_{\max}\|\boldsymbol{\beta}\|_2} : 0 < |\mathcal{S}_{\boldsymbol{\beta}}| \leq s \right\},$$

où $\mathcal{S}_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$ représente le support de $\boldsymbol{\beta}$. Cette quantité est similaire à la condition de compatibilité, mais compare maintenant directement les normes $\ell_2$ de $\mathbf{X}\boldsymbol{\beta}$ et de $\boldsymbol{\beta}$. Elle demande donc que la matrice $\mathbf{X}$ ne contracte pas excessivement un vecteur ayant au plus $s$ coefficients non nuls.

Cette condition peut aussi être interprétée à partir des valeurs propres. Pour un sous-ensemble $\mathcal{S}$ de variables, considérons la sous-matrice $\mathbf{X}_{\mathcal{S}}$ formée des colonnes de $\mathbf{X}$ dont les indices appartiennent à $\mathcal{S}$. Nous avons

$$\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2 = \boldsymbol{\beta}_{\mathcal{S}}^\intercal \mathbf{X}_{\mathcal{S}}^\intercal \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}.$$

Ainsi, la plus petite valeur propre de $\mathbf{X}_{\mathcal{S}}^\intercal\mathbf{X}_{\mathcal{S}}$ contrôle la plus petite valeur que peut prendre cette norme relativement à $\|\boldsymbol{\beta}_{\mathcal{S}}\|_2$. La condition $\widetilde{\phi}(s) > 0$ demande donc essentiellement que les plus petites valeurs propres des matrices $\mathbf{X}_{\mathcal{S}}^\intercal\mathbf{X}_{\mathcal{S}}$, pour les sous-modèles de taille au plus $s$, ne soient pas trop proches de 0. Il n'est donc pas nécessaire que la matrice complète $\mathbf{X}^\intercal\mathbf{X}$ soit inversible. Il suffit que les petites sous-matrices correspondant aux modèles parcimonieux considérés soient suffisamment bien conditionnées. Cette distinction est particulièrement importante lorsque $p > n$, puisque l'inversibilité de la matrice complète est alors impossible.

Les deux conditions jouent des rôles légèrement différents. La compatibilité est une condition relativement faible et est particulièrement utile pour contrôler l'erreur de prédiction et l'erreur d'estimation sous la norme $\ell_1$. La condition sur les valeurs singulières parcimonieuses est plus forte et permet notamment d'obtenir des résultats sous la norme $\ell_2$. Dans les deux cas, l'idée demeure la même : la parcimonie permet de remplacer une condition d'inversibilité globale impossible par des conditions d'inversibilité restreintes aux petits modèles qui nous intéressent.

Sous ces conditions, Castillo, Schmidt-Hieber et Van der Vaart (2015) obtiennent plusieurs résultats intéressants pour la loi *a posteriori*. Posons $\boldsymbol{\beta}^0$ comme étant le véritable vecteur de coefficients, $\mathcal{S}_0 = \mathcal{S}_{\boldsymbol{\beta}^0}$ son support et $s_0 = |\mathcal{S}_0|$ sa dimension. Un premier résultat montre que la loi *a posteriori* ne sélectionne généralement pas des modèles dont la dimension est beaucoup plus grande que $s_0$. Plus précisément, sous les conditions imposées à $\pi_p$, la masse *a posteriori* se concentre sur des modèles dont la taille demeure du même ordre que celle du véritable modèle. Le paramètre contrôlant la décroissance de $\pi_p$ peut même être utilisé afin de renforcer la pénalisation des modèles trop grands.

Ce résultat illustre directement l'intérêt d'introduire une loi *a priori* sur la dimension. Il existe $\binom{p}{s}$ modèles différents de taille $s$, de sorte que le nombre de modèles possibles augmente extrêmement rapidement avec la dimension. Sans mécanisme pénalisant les grandes

valeurs de $s$, la loi *a posteriori* pourrait attribuer une masse importante à de nombreux modèles inutilement complexes. Les auteurs proposent notamment des lois dites de complexité de la forme

$$\pi_p(s) \propto c^{-s} p^{-as}, \qquad a, c > 0,$$

qui accordent progressivement moins de masse aux modèles de grande dimension. La loi $\pi_p$ agit donc comme une pénalisation de la complexité du modèle.

Cette formulation offre également l'avantage de séparer deux rôles qui sont davantage liés dans le Lasso bayésien. La sélection des variables est principalement contrôlée par $\pi_p$ et par la masse ponctuelle en 0, tandis que la distribution $g_{\mathcal{S}}$ contrôle l'estimation des coefficients une fois ceux-ci inclus dans le modèle. Il n'est donc pas nécessaire d'utiliser une forte contraction des coefficients actifs vers 0 afin d'obtenir un modèle parcimonieux. Le *spike* permet de mettre exactement à 0 les variables jugées inactives, alors que le *slab* peut demeurer relativement diffus afin d'estimer les coefficients réellement non nuls. Cette séparation permet de limiter le biais qui pourrait apparaître si le même mécanisme devait simultanément éliminer les variables inutiles et rétrécir les variables actives.

Les auteurs montrent ensuite que cette loi *a posteriori* se concentre autour du véritable paramètre. Sous les conditions précédentes sur la matrice de design, les ordres de convergence obtenus sont comparables à ceux retrouvés dans la littérature fréquentiste sur la régression parcimonieuse. Lorsque les colonnes de $\mathbf{X}$ ont une norme de l'ordre de $\sqrt{n}$, les résultats correspondent essentiellement à une erreur d'estimation sous la norme $\ell_2$ de l'ordre

$$\sqrt{\frac{s_0 \log p}{n}},$$

à des constantes dépendant de la matrice de design près. Ainsi, même lorsque $p$ est très grand, une bonne estimation de $\boldsymbol{\beta}^0$ demeure possible lorsque $s_0$ est suffisamment petit relativement à $n$ et que la matrice de design respecte les conditions précédentes.

Finalement, la méthodologie possède aussi de bonnes propriétés de sélection de variables. La loi *a priori* sur la dimension évite asymptotiquement d'ajouter inutilement des variables au véritable modèle. Inversement, une variable dont le véritable coefficient est extrêmement près de 0 ne peut généralement pas être distinguée d'une variable inactive. Les auteurs imposent donc une condition minimale sur l'amplitude des coefficients non nuls. Sous cette condition minimale, ainsi que sous les hypothèses supplémentaires imposées par les auteurs sur la matrice de design et la loi *a priori*, la probabilité *a posteriori* attribuée au véritable support $\mathcal{S}_0$ tend vers 1. La procédure permet alors d'effectuer une sélection consistante du véritable modèle.

En plus de ces résultats de sélection et d'estimation, Castillo, Schmidt-Hieber et Van der Vaart (2015) montrent que, sous certaines conditions, la loi *a posteriori* peut être approximée par un mélange de lois normales associées aux différents sous-modèles plausibles. Lorsque la sélection du véritable support devient consistante, ce mélange se réduit asymptotiquement à une loi normale centrée autour de l'estimateur des moindres

carrés calculé uniquement sur les variables du véritable support. Ainsi, contrairement à une procédure produisant seulement une estimation ponctuelle, cette approche permet à la fois de sélectionner des variables, d'estimer leurs coefficients et de conserver une mesure de l'incertitude à travers la loi *a posteriori*.

## 3.3 Exemples bayésiens

De façon similaire à la section 2.4, nous testons deux techniques de régression bayésienne sur un jeu de données simulé ainsi que sur des données réelles. Le Lasso bayésien ainsi qu'une méthode inspirée des techniques *Spike and Slab* sont étudiés. Cette seconde approche utilise une masse de probabilité en 0 comme composante *spike* et une loi normale plus diffuse comme composante *slab*, conjointement à des indicateurs d'inclusion permettant d'effectuer la sélection des variables.

### 3.3.1 Comparaison sur un jeu de données simulé

Afin de comparer nos deux méthodes bayésiennes avec les approches fréquentistes explorées précédemment, le même jeu de données simulé est utilisé. Ainsi, nous fixons le vecteur des coefficients de régression aux mêmes valeurs qu'en (17), avec $n = 72$ et $p = 16$. Les prédicteurs sont générés aléatoirement selon une loi $\Gamma(2, 2)$ avec des erreurs normales de moyenne nulle et de variance unitaire. Pour comparer l'efficacité des différentes méthodes, nous réutilisons le même ensemble test comprenant $N = 10\,000$ observations pour calculer l'erreur quadratique moyenne de prédiction de test (18). La mesure de distance quadratique (19) est aussi utilisée afin de comparer les coefficients estimés aux véritables coefficients.

Dans cette simulation, nous utilisons la hiérarchie (27) pour le Lasso bayésien et nous incluons le paramètre de régularisation $\lambda$ directement dans l'échantillonneur de Gibbs en attribuant une loi *a priori* à $\lambda^2$ selon (29). Les coefficients de régression sont ensuite estimés par leur moyenne *a posteriori*. Contrairement au Lasso fréquentiste, cette estimation ne produit généralement pas de coefficients exactement nuls et n'effectue donc pas directement une sélection de variables.

Pour l'approche *Spike and Slab*, nous utilisons une combinaison d'une composante *spike* discrète en 0 avec une loi normale plus diffuse pour le *slab*. Pour chaque coefficient $\beta_j$, nous introduisons un indicateur d'inclusion $\gamma_j \in \{0, 1\}$ tel que

$$\beta_j|\gamma_j \sim \begin{cases} \delta_0, & \gamma_j = 0, \\ \mathcal{N}(0, \tau^2), & \gamma_j = 1, \end{cases}$$

où $\delta_0$ représente une masse ponctuelle en 0 et $\tau^2$ contrôle la dispersion de la composante *slab*. Nous attribuons également aux indicateurs la loi

$$\gamma_j \sim \text{Bernoulli}(q),$$

où $q$ représente la probabilité *a priori* qu'une variable soit incluse dans le modèle. Dans notre implémentation, nous fixons $q = 0.5$ et $\tau^2 = 100$, de manière à ne pas favoriser initialement l'inclusion ou l'exclusion d'une variable et à conserver une composante *slab* relativement diffuse. Pour la variance du terme d'erreur, nous utilisons la loi *a priori* impropre invariante d'échelle

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Conditionnellement aux coefficients de régression, sa distribution conditionnelle complète est alors

$$\sigma^2|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X} \sim \Gamma^{-1}\left(\frac{n}{2}, \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2}\right).$$

Cette implémentation diffère du SSVS gaussien présenté précédemment puisque la composante *spike* est ici une masse ponctuelle en zéro plutôt qu'une loi normale de faible variance.

L'échantillonneur de Gibbs alterne alors entre l'inclusion ou l'exclusion de chaque variable et l'estimation des coefficients actifs. Conditionnellement aux autres paramètres et aux données, une probabilité d'inclusion est calculée pour chaque variable. L'indicateur $\gamma_j$ est ensuite tiré selon cette probabilité. Lorsque $\gamma_j = 0$, le coefficient correspondant est fixé exactement à $\beta_j = 0$, alors que lorsque $\gamma_j = 1$, sa nouvelle valeur est simulée à partir de sa loi conditionnelle normale. Ainsi, contrairement au Lasso bayésien, cette approche permet aux variables de sortir complètement du modèle au cours de l'échantillonnage.

Après le temps de chauffe, la proportion des itérations pour lesquelles $\gamma_j = 1$ permet d'estimer la probabilité *a posteriori* d'inclusion $\mathbb{P}\{\gamma_j = 1|\mathbf{y}\}$. Nous retenons finalement une variable lorsque cette probabilité est supérieure ou égale à 0.5. Les coefficients des variables ne respectant pas ce seuil sont fixés à 0, tandis que les coefficients retenus sont estimés à partir de leur moyenne *a posteriori*. Cette procédure permet ainsi d'obtenir simultanément une estimation des coefficients et une sélection explicite des variables. Le seuil de 0.5 correspond à la règle du modèle de probabilité médiane, qui possède des propriétés d'optimalité prédictive sous les conditions étudiées par Barbieri et Berger (2004). Le tableau 3 présente les performances des deux méthodes bayésiennes.

Table 3 – Performances prédictives et erreur d'estimation des coefficients bayésiens sur les données simulées.

| | $EQMP_{ent}$ | $EQMP_{test}$ | Distance | Nombre non nuls |
|---|---|---|---|---|
| Lasso bayésien | 0.7968 | 1.2494 | 0.2147 | 16 |
| Spike and Slab | 0.9653 | 1.0720 | 0.0794 | 4 |

Le Lasso bayésien possède une EQMP d'entraînement de 0.7968, soit une valeur très similaire à celles obtenues avec les MCO et la régression Ridge, respectivement de 0.7965 et 0.7966. Son EQMP de test de 1.2494 est également proche de celles de ces deux méthodes.

Dans cette réalisation, le Lasso bayésien améliore donc légèrement la performance prédictive des MCO, mais demeure moins performant que le Lasso fréquentiste et l'*Elastic Net*, dont les EQMP de test sont respectivement de 1.1516 et 1.0971.

Une situation similaire est observée pour l'estimation des coefficients. Le Lasso bayésien obtient une distance quadratique de

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 = 0.2147,$$

comparativement à 0.2216 pour les MCO et à 0.2271 pour la régression Ridge. Cette amélioration demeure toutefois moins importante que celle obtenue avec le Lasso et l'*Elastic Net*, pour lesquels les distances quadratiques sont respectivement de 0.1389 et 0.0937. Les estimations individuelles présentées en annexe montrent également que les coefficients obtenus avec le Lasso bayésien sont relativement proches de ceux des MCO et de la régression Ridge.

Comme prévu, la moyenne *a posteriori* du Lasso bayésien ne produit aucun coefficient exactement nul. Les 16 variables sont donc conservées, même si plusieurs coefficients associés aux variables réellement inactives sont estimés très près de 0. Cette observation illustre une différence importante avec le Lasso fréquentiste. Bien que la loi de Laplace utilisée par le Lasso bayésien favorise le rétrécissement des coefficients vers 0, la loi *a posteriori* demeure continue et la moyenne *a posteriori* n'est généralement pas exactement nulle. Le Lasso bayésien permet donc ici de réduire l'amplitude de certains coefficients sans effectuer directement une sélection de variables.

Le comportement du *Spike and Slab* est sensiblement différent. Cette méthode possède l'EQMP d'entraînement la plus élevée parmi les méthodes considérées, soit 0.9653. Cette augmentation est cohérente avec la sélection plus importante effectuée par le modèle. En imposant exactement certains coefficients à 0, le modèle possède moins de flexibilité pour s'ajuster aux observations utilisées pour l'estimation. Cependant, cette situation s'inverse lorsque nous considérons les données de test. Le *Spike and Slab* obtient une EQMP de test de 1.0720, soit la plus faible parmi toutes les méthodes étudiées. Elle est notamment inférieure à celle de l'*Elastic Net*, qui était de 1.0971. Puisque la variance du terme d'erreur est égale à 1, ces deux valeurs demeurent relativement proches de l'erreur de prédiction minimale.

Le *Spike and Slab* obtient également la plus faible erreur d'estimation des coefficients avec

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 = 0.0794.$$

Cette valeur est inférieure à 0.0937 pour l'*Elastic Net* et à 0.1389 pour le Lasso. Dans cette réalisation, la sélection explicite effectuée par le *Spike and Slab* permet donc à la fois d'améliorer la prédiction sur les nouvelles observations et de se rapprocher davantage du véritable vecteur de coefficients.

La sélection obtenue permet cependant de nuancer ce résultat. Le véritable modèle contient cinq coefficients non nuls, soit $\beta_1$, $\beta_2$, $\beta_4$, $\beta_5$ et $\beta_8$. Le *Spike and Slab* sélectionne

exactement les quatre premiers, mais élimine $\beta_8 = 0.2$. En revanche, les 11 coefficients réellement nuls sont tous éliminés. Comparativement, le Lasso conserve 12 coefficients et l'*Elastic Net* en conserve 10. Le *Spike and Slab* produit donc un modèle beaucoup plus parcimonieux, mais ne parvient pas à détecter le coefficient non nul de plus faible amplitude. Ce comportement est raisonnable puisque l'effet associé à $\beta_8 = 0.2$ est relativement faible par rapport aux autres coefficients du modèle.

Un avantage supplémentaire des approches bayésiennes est qu'elles permettent de quantifier directement l'incertitude associée à l'estimation des paramètres. Pour le Lasso bayésien, les intervalles de crédibilité de niveau 95% sont obtenus à partir des réalisations conservées de l'échantillonneur de Gibbs après le temps de chauffe. Pour chaque coefficient $\beta_j$, nous calculons les quantiles empiriques de niveau 0.025 et 0.975 des réalisations

$$\beta_j^{(b+1)}, \ldots, \beta_j^{(M)},$$

afin d'obtenir l'intervalle de crédibilité

$$\big[q_{0.025}(\beta_j|\mathbf{y}), q_{0.975}(\beta_j|\mathbf{y})\big].$$

Les résultats du Lasso bayésien montrent que les intervalles associés aux quatre coefficients de grande amplitude, soit $\beta_1$, $\beta_2$, $\beta_4$ et $\beta_5$, excluent tous la valeur 0. À l'inverse, les intervalles associés aux coefficients réellement nuls contiennent tous 0. Le coefficient $\beta_8 = 0.2$ constitue encore une fois le cas le plus difficile. Son estimation *a posteriori* est de 0.2736, mais son intervalle de crédibilité est approximativement donné par $[-0.0001, 0.5514]$, de sorte qu'il n'est pas possible de l'identifier clairement comme différent de 0 à partir de cet intervalle. Ce résultat est cohérent avec la difficulté rencontrée par le *Spike and Slab* pour sélectionner cette même variable.

Pour le *Spike and Slab*, les intervalles sont obtenus de la même manière à partir des quantiles des réalisations *a posteriori* des coefficients. Leur interprétation est toutefois légèrement différente puisque la loi *a posteriori* de chaque coefficient peut contenir une masse ponctuelle en 0. Ainsi, lorsqu'une variable est exclue dans une grande proportion des itérations de l'échantillonneur, une grande partie des réalisations de son coefficient est exactement nulle. Les quantiles peuvent alors eux-mêmes être égaux à 0, ce qui explique les intervalles dégénérés $[0, 0]$ observés pour plusieurs variables inactives.

La probabilité *a posteriori* d'inclusion apporte une information supplémentaire pour cette méthode. Les coefficients $\beta_1$, $\beta_2$, $\beta_4$ et $\beta_5$ possèdent tous une probabilité d'inclusion égale à 1 dans notre simulation. Les variables réellement nulles ont quant à elles des probabilités très faibles, généralement de l'ordre de 0.01 à 0.03. Le coefficient $\beta_8$, bien qu'il soit réellement non nul, possède seulement une probabilité d'inclusion de 0.0746. Puisque notre règle de sélection conserve une variable seulement lorsque

$$\mathbb{P}\big\{\gamma_j = 1|\mathbf{y}\big\} \geq 0.5,$$

cette variable est éliminée du modèle final.

Il est toutefois important de souligner que le seuil de 0.5 constitue une règle de décision et non une contrainte imposée par le modèle *Spike and Slab*. L'utilisateur peut choisir un seuil différent selon le compromis désiré entre parcimonie et détection des effets de faible amplitude. Par exemple, une règle moins restrictive pourrait consister à éliminer une variable uniquement lorsque sa probabilité *a posteriori* d'inclusion est inférieure à 0.05. Dans notre simulation, un tel choix permettrait de conserver $\beta_8$, puisque

$$\mathbb{P}\{\gamma_8 = 1|\mathbf{y}\} = 0.0746 > 0.05.$$

Une diminution du seuil favorise donc l'inclusion des variables pour lesquelles l'évidence *a posteriori* est plus faible, au prix d'un risque accru de conserver certaines variables réellement inactives. À l'inverse, un seuil plus élevé produit généralement une sélection plus parcimonieuse. Le choix du seuil dépend ainsi du compromis recherché entre l'élimination des variables inutiles et la détection des effets de faible amplitude. Dans cette simulation, nous conservons tout de même le seuil de 0.5 conformément à la règle du modèle de probabilité médiane de Barbieri et Berger (2004) et à ses propriétés prédictives sous certaines conditions.

Ainsi, dans cette réalisation, le *Spike and Slab* présente les meilleures performances parmi les méthodes étudiées en termes d'EQMP de test et de distance quadratique au véritable vecteur de coefficients. Il permet également une sélection beaucoup plus parcimonieuse que le Lasso et l'*Elastic Net*. Avec la règle de sélection adoptée, son principal défaut dans cette simulation est l'élimination du faible coefficient $\beta_8 = 0.2$. Cette conclusion dépend toutefois du seuil de probabilité *a posteriori* d'inclusion retenu, puisqu'un seuil moins restrictif pourrait permettre de conserver cette variable. Le Lasso bayésien offre, pour sa part, des performances proches de celles des MCO et de la régression Ridge, mais l'information contenue dans sa loi *a posteriori* permet d'obtenir des intervalles de crédibilité et d'évaluer l'incertitude associée à chaque coefficient. Comme pour les résultats fréquentistes, ces observations correspondent toutefois à une seule réalisation du modèle simulé et ne permettent pas, à elles seules, de conclure à la supériorité générale d'une méthode.

### 3.3.2 Comparaison sur un jeu de données réel

Dans cette seconde comparaison, nous appliquons les deux méthodes bayésiennes au même jeu de données sur le cancer de la prostate que celui utilisé dans la section 2.4.2. Ces données, provenant de Stamey et al. (1989), comprennent 97 observations et huit variables explicatives permettant d'expliquer le logarithme de la quantité de l'antigène prostatique spécifique (`lpsa`). Nous conservons également la même séparation des données, soit 67 observations pour l'entraînement et 30 observations pour le test. Les variables explicatives sont standardisées à partir des moyennes et écarts-types calculés uniquement sur les données d'entraînement, puis les mêmes transformations sont appliquées aux observations de test.

Nous utilisons les mêmes deux méthodes bayésiennes que pour les données simulées. Le Lasso bayésien est estimé à partir de la hiérarchie (27), avec une loi *a priori* sur $\lambda^2$, alors

que la seconde méthode utilise une composante *Spike* ponctuelle en 0 et une composante *Slab* normale. Pour cette dernière, nous conservons les mêmes valeurs $q = 0.5$ et $\tau^2 = 100$ que dans la simulation et une variable est sélectionnée lorsque sa probabilité *a posteriori* d'inclusion est supérieure ou égale à 0.5.

Comme pour les méthodes fréquentistes, les performances sont évaluées à l'aide de l'EQMP d'entraînement et de l'EQMP de test. La distance quadratique au véritable vecteur de coefficients utilisée dans la simulation ne peut toutefois pas être calculée puisque les véritables coefficients de régression sont inconnus pour ces données. Le nombre de coefficients non nuls est également présenté afin de comparer la parcimonie des modèles obtenus. Les estimations individuelles des coefficients ainsi que les intervalles de crédibilité des méthodes bayésiennes sont présentés en annexe.

TABLE 4 – Performances d'ajustement, performances prédictives et parcimonie des modèles sous les méthodes bayésiennes sur les données du cancer de la prostate.

| | $EQMP_{ent}$ | $EQMP_{test}$ | Nombre non nuls |
|---|---|---|---|
| Lasso bayésien | 0.4425 | 0.4943 | 8 |
| Spike and Slab | 0.5609 | 0.4824 | 2 |

Les résultats du Lasso bayésien sont tout d'abord particulièrement proches de ceux du Lasso fréquentiste. Son EQMP d'entraînement est de 0.4425, comparativement à 0.4423 pour le Lasso fréquentiste, alors que leur EQMP de test sont respectivement de 0.4943 et 0.4949. Les estimations individuelles des coefficients sont également très similaires entre les deux méthodes. Par exemple, les coefficients associés à `lcavol`, `lweight`, `age`, `lbph`, `svi`, `lcp` et `pgg45` sont pratiquement identiques. La principale différence demeure que le Lasso fréquentiste fixe le coefficient associé à `gleason` exactement à 0, tandis que le Lasso bayésien l'estime à $-0.0004$.

Cette situation est cohérente avec les propriétés des deux méthodes. Le Lasso fréquentiste peut produire des coefficients exactement nuls en raison de la pénalisation $\ell_1$, alors que la moyenne *a posteriori* du Lasso bayésien est généralement différente de 0. Le Lasso bayésien conserve ainsi les huit coefficients dans cette application, même si certains d'entre eux sont estimés extrêmement près de 0. En termes de prédiction, il offre une performance légèrement supérieure aux MCO, dont l'EQMP de test est de 0.5213, et à la régression Ridge, dont l'EQMP est de 0.5000. Sa performance demeure cependant inférieure à celle de l'*Elastic Net*, qui obtient l'EQMP de test la plus faible avec une valeur de 0.4357.

Les intervalles de crédibilité permettent d'obtenir davantage d'information sur l'incertitude associée aux estimations du Lasso bayésien. Comme dans la simulation, ces intervalles sont construits à partir des 24 500 réalisations de l'échantillonneur de Gibbs conservées après le temps de chauffe de 500 itérations. Pour chaque coefficient $\beta_j$, nous calculons les quantiles empiriques 0.025 et 0.975 des réalisations de sa loi *a posteriori*, ce qui donne un

intervalle de crédibilité de niveau 95% de la forme

$$\big[q_{0.025}(\beta_j|\mathbf{y}), q_{0.975}(\beta_j|\mathbf{y})\big].$$

Les intervalles associés aux variables `lcavol`, `lweight`, `lbph` et `svi` excluent la valeur 0. À l'inverse, ceux associés à `age`, `lcp`, `gleason` et `pgg45` contiennent 0. Ainsi, bien que la moyenne *a posteriori* du Lasso bayésien ne permette pas directement d'éliminer des variables, les intervalles de crédibilité permettent d'identifier les coefficients pour lesquels les données apportent davantage d'évidence en faveur d'un effet différent de 0.

Le *Spike and Slab* produit pour sa part un modèle beaucoup plus parcimonieux. Seulement deux variables sont retenues, soit `lcavol` et `lweight`. Leurs probabilités *a posteriori* d'inclusion sont respectivement de 1 et de 0.7489, alors que les probabilités associées aux six autres variables demeurent toutes inférieures à 0.1. Avec la règle de sélection $\mathbb{P}\{\gamma_j = 1|\mathbf{y}\} \geq 0.5$, seules ces deux variables sont donc conservées dans l'estimateur final.

Cette sélection importante se traduit par une EQMP d'entraînement de 0.5609, soit la plus élevée parmi toutes les méthodes considérées. Le modèle possède effectivement beaucoup moins de flexibilité pour s'ajuster aux observations d'entraînement puisqu'il ne contient que deux variables explicatives. Toutefois, son EQMP de test diminue à 0.4824. Le *Spike and Slab* obtient ainsi une meilleure performance prédictive que les MCO, la régression Ridge, le Lasso fréquentiste et le Lasso bayésien. Il demeure cependant moins performant que l'*Elastic Net*, dont l'EQMP de test est de 0.4357.

Les intervalles de crédibilité du *Spike and Slab* sont également obtenus à partir des quantiles empiriques des réalisations de l'échantillonneur de Gibbs. Leur comportement diffère toutefois de ceux du Lasso bayésien puisque la loi *a posteriori* des coefficients peut contenir une masse ponctuelle en 0. Lorsqu'une variable est exclue pendant une grande proportion des itérations, plusieurs réalisations de son coefficient sont donc exactement égales à 0. Ceci explique notamment pourquoi certains intervalles sont réduits à $[0, 0]$ ou possèdent une borne égale à 0.

Il est également important de distinguer l'intervalle de crédibilité de la règle utilisée pour effectuer la sélection des variables. Par exemple, le coefficient associé à `lweight` possède un intervalle de crédibilité approximativement donné par [0, 0.5338], mais sa probabilité *a posteriori* d'inclusion est de 0.7489. La variable est donc sélectionnée puisque cette probabilité dépasse le seuil de 0.5. Dans le cas du *Spike and Slab*, c'est ainsi la probabilité d'inclusion qui détermine la sélection et non l'exclusion de 0 de l'intervalle de crédibilité.

Les deux méthodes bayésiennes proposent donc des comportements assez différents sur ces données. Le Lasso bayésien produit des estimations et des performances prédictives presque identiques à celles du Lasso fréquentiste, mais permet en plus de quantifier l'incertitude associée aux coefficients à partir de leur loi *a posteriori*. Le *Spike and Slab* produit quant à lui un modèle beaucoup plus parcimonieux en ne conservant que `lcavol` et `lweight`. Malgré cette réduction importante du nombre de variables, son EQMP de test demeure relativement faible et est inférieure à celles du Lasso fréquentiste et du Lasso bayésien. Dans

cette application, l'*Elastic Net* conserve toutefois la meilleure performance prédictive avec une EQMP de test de 0.4357.

Contrairement à la simulation, il n'est cependant pas possible de déterminer quelles variables devraient réellement appartenir au modèle puisque le véritable vecteur de coefficients est inconnu. Le fait que le *Spike and Slab* ne retienne que deux variables ne permet donc pas de conclure qu'il effectue une meilleure sélection que le Lasso ou l'*Elastic Net*. Ces résultats montrent plutôt qu'un modèle très parcimonieux peut, dans cette réalisation, conserver une bonne capacité de prédiction sur les données laissées de côté.

# 4 Conclusion

Dans ce rapport, nous avons étudié différentes méthodes d'estimation et de sélection de variables dans le modèle de régression linéaire, sous des approches fréquentistes et bayésiennes. Le point de départ de cette étude était l'idée qu'un estimateur biaisé peut parfois présenter de meilleures propriétés qu'un estimateur non biaisé lorsque l'objectif est notamment de diminuer l'erreur de prédiction. Les méthodes de régularisation permettent d'exploiter ce compromis en introduisant un rétrécissement des coefficients de régression.

D'un point de vue fréquentiste, l'estimateur Ridge permet de stabiliser l'estimation des coefficients, mais ne produit généralement pas de modèles parcimonieux. Le Lasso ajoute à cet effet une propriété de sélection de variables grâce à sa pénalisation $\ell_1$. L'*Elastic Net* combine finalement les pénalisations du Ridge et du Lasso afin de conserver cette capacité de sélection tout en améliorant notamment le traitement des variables corrélées. Les algorithmes LARS et LARS-EN permettent alors de construire efficacement les chemins de solutions de ces deux dernières méthodes.

L'approche bayésienne permet de considérer le même problème sous un angle différent en attribuant des lois *a priori* aux paramètres. Le Lasso bayésien montre notamment le lien entre la pénalisation $\ell_1$ et une loi *a priori* de Laplace. L'utilisation d'un échantillonneur de Gibbs permet ensuite d'étudier l'ensemble de la loi *a posteriori* plutôt que seulement une estimation ponctuelle. Cette approche offre ainsi l'avantage supplémentaire de quantifier l'incertitude entourant les coefficients à l'aide d'intervalles de crédibilité. En revanche, les moyennes *a posteriori* du Lasso bayésien ne produisent généralement pas de coefficients exactement nuls. Les lois *a priori* de type *Spike and Slab* offrent une solution naturelle à cette difficulté en séparant explicitement les variables actives et inactives.

Les applications numériques illustrent bien les compromis entre ces différentes méthodes. Sur les données simulées, le *Spike and Slab* obtient la plus faible EQMP de test et la plus faible distance quadratique au véritable vecteur de coefficients, tout en retrouvant quatre des cinq variables réellement actives. L'*Elastic Net* présente toutefois des performances très proches et conserve l'ensemble des variables réellement actives. Sur les données du cancer de la prostate, l'*Elastic Net* obtient plutôt la meilleure performance prédictive, alors que le *Spike and Slab* produit le modèle le plus parcimonieux en ne conservant que deux variables.

Le Lasso fréquentiste et le Lasso bayésien fournissent quant à eux des estimations et des erreurs de prédiction particulièrement similaires.

Ainsi, aucune méthode ne domine nécessairement sous tous les critères considérés. Le choix d'une technique dépend autant de la qualité de prédiction recherchée que de l'importance accordée à la parcimonie, à l'interprétabilité et à la quantification de l'incertitude. Finalement, les résultats numériques présentés demeurent limités à une réalisation simulée et à un seul jeu de données réel. Une étude basée sur plusieurs répétitions de la simulation, différents niveaux de corrélation entre les prédicteurs et des contextes de plus grande dimension permettrait de comparer plus solidement les performances moyennes des différentes approches.

# A Estimation des coefficients de régression dans une simulation

TABLE 5 – Estimations des coefficients de régression obtenues sur les données simulées.

| | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ |
|---|---|---|---|---|---|---|---|---|
| Bêta réel | 25.0000 | -10.0000 | 0 | 2.0000 | -5.0000 | 0 | 0 | 0.2000 |
| MCO | 24.9913 | -9.7905 | -0.1312 | 2.1085 | -5.0010 | -0.1203 | -0.1735 | 0.2856 |
| Ridge | 24.9816 | -9.7855 | -0.1326 | 2.1084 | -4.9980 | -0.1202 | -0.1758 | 0.2860 |
| Lasso | 24.9579 | -9.7611 | -0.0705 | 2.0921 | -4.9603 | -0.0554 | -0.1208 | 0.2033 |
| Elastic Net | 25.1079 | -9.8015 | -0.0226 | 2.0859 | -4.9616 | 0 | -0.0846 | 0.1392 |
| Lasso bayésien | 24.9860 | -9.7860 | -0.1273 | 2.1056 | -4.9961 | -0.1165 | -0.1714 | 0.2763 |
| Spike and Slab | 25.0446 | -9.9031 | 0 | 2.1589 | -5.0528 | 0 | 0 | 0 |

| | $\beta_9$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{13}$ | $\beta_{14}$ | $\beta_{15}$ | $\beta_{16}$ |
|---|---|---|---|---|---|---|---|---|
| Bêta réel | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MCO | -0.1246 | -0.1189 | -0.0065 | -0.2505 | -0.0357 | -0.0472 | -0.0057 | -0.0270 |
| Ridge | -0.1232 | -0.1202 | -0.0065 | -0.2541 | -0.0354 | -0.0480 | -0.0059 | -0.0288 |
| Lasso | -0.0622 | -0.0735 | 0 | -0.1929 | 0 | -0.0299 | 0 | 0 |
| Elastic Net | 0 | -0.0236 | 0 | -0.1479 | 0 | -0.0037 | 0 | 0 |
| Lasso bayésien | -0.1192 | -0.1167 | -0.0085 | -0.2460 | -0.0328 | -0.0492 | -0.0050 | -0.0257 |
| Spike and Slab | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# B Estimation des coefficients de régression sur les données du cancer de la prostate

TABLE 6 – Estimations des coefficients de régression obtenues sur les données réelles du cancer de la prostate.

| | `lcavol` | `lweight` | `age` | `lbph` | `svi` | `lcp` | `gleason` | `pgg45` |
|---|---|---|---|---|---|---|---|---|
| MCO | 0.7164 | 0.2926 | -0.1425 | 0.2120 | 0.3096 | -0.2890 | -0.0209 | 0.2773 |
| Ridge | 0.6428 | 0.2891 | -0.1210 | 0.2055 | 0.2928 | -0.1987 | 0.0039 | 0.2257 |
| Lasso | 0.6786 | 0.2848 | -0.1150 | 0.1984 | 0.2834 | -0.2093 | 0 | 0.2197 |
| Elastic Net | 0.6347 | 0.2515 | 0 | 0.1080 | 0.1844 | 0 | 0 | 0.0674 |
| Lasso bayésien | 0.6756 | 0.2843 | -0.1158 | 0.1967 | 0.2811 | -0.2068 | -0.0004 | 0.2209 |
| Spike and Slab | 0.7885 | 0.2635 | 0 | 0 | 0 | 0 | 0 | 0 |

# C Intervalles de crédibilité de niveau 95% pour les estimations bayésiennes des coefficients de régression dans la simulation

TABLE 7 – Intervalles de crédibilité pour les coefficients par la technique du Lasso bayésien

| | Bêta réel | Estimation | Borne inférieure | Borne supérieure | IC exclut 0 |
|---|---|---|---|---|---|
| $\beta_1$ | 25.0000 | 24.9861 | 24.7271 | 25.2435 | Oui |
| $\beta_2$ | -10.0000 | -9.7848 | -10.0654 | -9.5041 | Oui |
| $\beta_3$ | 0 | -0.1259 | -0.4160 | 0.1572 | Non |
| $\beta_4$ | 2.0000 | 2.1079 | 1.8386 | 2.3798 | Oui |
| $\beta_5$ | -5.0000 | -4.9955 | -5.2805 | -4.7119 | Oui |
| $\beta_6$ | 0 | -0.1159 | -0.3953 | 0.1556 | Non |
| $\beta_7$ | 0 | -0.1707 | -0.4594 | 0.1144 | Non |
| $\beta_8$ | 0.2000 | 0.2736 | -0.0001 | 0.5514 | Non |
| $\beta_9$ | 0 | -0.1198 | -0.4076 | 0.1630 | Non |
| $\beta_{10}$ | 0 | -0.1156 | -0.3918 | 0.1575 | Non |
| $\beta_{11}$ | 0 | -0.0090 | -0.2802 | 0.2589 | Non |
| $\beta_{12}$ | 0 | -0.2454 | -0.5413 | 0.0485 | Non |
| $\beta_{13}$ | 0 | -0.0309 | -0.3221 | 0.2583 | Non |
| $\beta_{14}$ | 0 | -0.0509 | -0.3260 | 0.2228 | Non |
| $\beta_{15}$ | 0 | -0.0033 | -0.2747 | 0.2664 | Non |
| $\beta_{16}$ | 0 | -0.0264 | -0.2907 | 0.2426 | Non |

TABLE 8 – Intervalles de crédibilité pour les coefficients par la technique *Spike and Slab*

| | Bêta réel | Estimation | Borne inférieure | Borne supérieure | Probabilité d'inclusion | Sélection |
|---|---|---|---|---|---|---|
| $\beta_1$ | 25.0000 | 25.0434 | 24.8022 | 25.2835 | 1.0000 | Oui |
| $\beta_2$ | -10.0000 | -9.9038 | -10.1478 | -9.6609 | 1.0000 | Oui |
| $\beta_3$ | 0 | 0 | 0.0000 | 0.0000 | 0.0145 | Non |
| $\beta_4$ | 2.0000 | 2.1587 | 1.9124 | 2.4022 | 1.0000 | Oui |
| $\beta_5$ | -5.0000 | -5.0529 | -5.2983 | -4.8103 | 1.0000 | Oui |
| $\beta_6$ | 0 | 0 | 0.0000 | 0.0000 | 0.0142 | Non |
| $\beta_7$ | 0 | 0 | 0.0000 | 0.0000 | 0.0216 | Non |
| $\beta_8$ | 0.2000 | 0 | 0.0000 | 0.2866 | 0.0746 | Non |
| $\beta_9$ | 0 | 0 | 0.0000 | 0.0000 | 0.0265 | Non |
| $\beta_{10}$ | 0 | 0 | 0.0000 | 0.0000 | 0.0181 | Non |
| $\beta_{11}$ | 0 | 0 | 0.0000 | 0.0000 | 0.0149 | Non |
| $\beta_{12}$ | 0 | 0 | -0.0635 | 0.0000 | 0.0313 | Non |
| $\beta_{13}$ | 0 | 0 | 0.0000 | 0.0000 | 0.0126 | Non |
| $\beta_{14}$ | 0 | 0 | 0.0000 | 0.0000 | 0.0221 | Non |
| $\beta_{15}$ | 0 | 0 | 0.0000 | 0.0000 | 0.0117 | Non |
| $\beta_{16}$ | 0 | 0 | 0.0000 | 0.0000 | 0.0119 | Non |

# D Intervalles de crédibilité de niveau 95% pour les estimations bayésiennes des coefficients de régression sur les données du cancer de la prostate

TABLE 9 – Intervalles de crédibilité pour les coefficients par la technique du Lasso bayésien

| | Estimation | Borne inférieure | Borne supérieure | IC exclut 0 |
|---|---|---|---|---|
| `lcavol` | 0.6756 | 0.4215 | 0.9321 | Oui |
| `lweight` | 0.2843 | 0.0777 | 0.4878 | Oui |
| `age` | -0.1158 | -0.3107 | 0.0685 | Non |
| `lbph` | 0.1967 | 0.0042 | 0.3965 | Oui |
| `svi` | 0.2811 | 0.0408 | 0.5207 | Oui |
| `lcp` | -0.2068 | -0.4991 | 0.0693 | Non |
| `gleason` | -0.0004 | -0.2481 | 0.2465 | Non |
| `pgg45` | 0.2209 | -0.0514 | 0.5115 | Non |

TABLE 10 – Intervalles de crédibilité pour les coefficients par la technique *Spike and Slab*

| | Estimation | Borne inférieure | Borne supérieure | Probabilité d'inclusion | Sélection |
|---|---|---|---|---|---|
| `lcavol` | 0.7885 | 0.5432 | 1.0228 | 1.0000 | Oui |
| `lweight` | 0.2635 | 0.0000 | 0.5338 | 0.7489 | Oui |
| `age` | 0 | 0.0000 | 0.0000 | 0.0133 | Non |
| `lbph` | 0 | 0.0000 | 0.3267 | 0.0981 | Non |
| `svi` | 0 | 0.0000 | 0.3206 | 0.0880 | Non |
| `lcp` | 0 | 0.0000 | 0.0000 | 0.0158 | Non |
| `gleason` | 0 | 0.0000 | 0.0000 | 0.0151 | Non |
| `pgg45` | 0 | 0.0000 | 0.1527 | 0.0431 | Non |

## Références


ALLEN, David M (1974). « The relationship between variable selection and data augmentation and a method for prediction ». In : *Technometrics* 16.1, p. 125-127.

ANDREWS, David F et Colin L MALLOWS (1974). « Scale mixtures of normal distributions ». In : *Journal of the Royal Statistical Society: Series B (Methodological)* 36.1, p. 99-102.

BARBIERI, Maria Maddalena et James O. BERGER (2004). « Optimal predictive model selection ». In : *The Annals of Statistics* 32.3, p. 870-897.

Bédard, Mylène (2025). *Modèles hiérarchiques et échantillonneur de Gibbs*. STT 2105, Université de Montréal. Accès: 19 décembre 2025.

Booth, James G et James P Hobert (1999). « Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm ». In : *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.1, p. 265-285.

Boyd, Stephen et Lieven Vandenberghe (2004). *Convex Optimization*. Cambridge university press.

Castillo, Ismaël, Johannes Schmidt-Hieber et Aad Van der Vaart (2015). « Bayesian linear regression with sparse priors ». In : *The Annals of Statistics* 43.5, p. 1986-2018.

Dempster, Arthur P, Nan M Laird et Donald B Rubin (1977). « Maximum likelihood from incomplete data via the EM algorithm ». In : *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, p. 1-22.

Efron, Bradley et al. (2004). « Least angle regression ». In : *The Annals of Statistics* 32.2, p. 407-499.

El Ghaoui, L, V Viallon et T Rabbani (2010). « Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems ». In : *Journal of Machine Learning Research*. arXiv preprint arXiv:1009.4219.

Frank, ILdiko E et Jerome H Friedman (1993). « A statistical view of some chemometrics regression tools ». In : *Technometrics* 35.2, p. 109-135.

Gelman, Andrew et al. (1995). *Bayesian Data Analysis*. Chapman et Hall/CRC.

George, Edward I et Robert E McCulloch (1993). « Variable selection via Gibbs sampling ». In : *Journal of the American Statistical Association* 88.423, p. 881-889.

Golub, Gene H, Michael Heath et Grace Wahba (1979). « Generalized cross-validation as a method for choosing a good ridge parameter ». In : *Technometrics* 21.2, p. 215-223.

Hastie, Trevor, Robert Tibshirani et Martin Wainwright (2015). « Statistical learning with sparsity ». In : *Monographs on Statistics and Applied Probability* 143.143, p. 8.

Hoerl, Arthur E et Robert W Kennard (1970). « Ridge regression: Biased estimation for nonorthogonal problems ». In : *Technometrics* 12.1, p. 55-67.

Hoff, Peter D (2009). *A First Course in Bayesian Statistical Methods*. T. 580. Springer.

Ishwaran, Hemant et J Sunil Rao (2005). « Spike and slab variable selection: Frequentist and Bayesian strategies ». In : *The Annals of Statistics* 33.2, p. 730-773.

James, William, Charles Stein et al. (1961). « Estimation with quadratic loss ». In : *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. T. 1. 1961. University of California Press, p. 361-379.

Lee, Peter M (1989). *Bayesian Statistics*. Oxford University Press London.

Leng, Chenlei, Minh-Ngoc Tran et David Nott (2014). « Bayesian adaptive Lasso ». In : *Annals of the Institute of Statistical Mathematics* 66.2, p. 221-244.

McCulloch, Charles E (1997). « Maximum likelihood algorithms for generalized linear mixed models ». In : *Journal of the American Statistical Association* 92.437, p. 162-170.

Mitchell, Toby J et John J Beauchamp (1988). « Bayesian variable selection in linear regression ». In : *Journal of the American Statistical Association* 83.404, p. 1023-1032.

Montgomery, Douglas C, Elizabeth A Peck et G Geoffrey Vining (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Osborne, Michael R, Brett Presnell et Berwin A Turlach (2000). « A new approach to variable selection in least squares problems ». In : *IMA Journal of Numerical Analysis* 20.3, p. 389-403.

Park, Trevor et George Casella (2008). « The bayesian lasso ». In : *Journal of the American Statistical Association* 103.482, p. 681-686.

Stamey, Thomas A et al. (1989). « Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients ». In : *The Journal of Urology* 141.5, p. 1076-1083.

Stein, Charles (1956). « Inadmissibility of the usual estimator for the mean of a multivariate normal distribution ». In : *Proceedings of the third Berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics*. T. 3. University of California Press, p. 197-207.

Tibshirani, Robert (1996). « Regression shrinkage and selection via the Lasso ». In : *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, p. 267-288.

Tibshirani, Robert et al. (2012). « Strong rules for discarding predictors in Lasso-type problems ». In : *Journal of the Royal Statistical Society Series B: Statistical Methodology* 74.2, p. 245-266.

Tibshirani, Ryan J (2013). « The Lasso problem and uniqueness ». In : *Electronic Journal of Statistics* 7, p. 1456-1490.

— (2023). *High-Dimensionnal Regression: Lasso*. Technical Report. University of California, Berkeley.

Tibshirani, Ryan J et Larry Wasserman (2017). *Sparsity, the Lasso, and Friends*. Technical Report. University of California, Berkeley.

Weisberg, Sanford (2005). *Applied Linear Regression*. T. 528. John Wiley & Sons.

Wu, Tong Tong et Kenneth Lange (2008). « Coordinate descent algorithms for Lasso penalized regression ». In : *The Annals of Applied Statistics* 2.1, p. 224-244.

Zou, Hui (2006). « The adaptive Lasso and its oracle properties ». In : *Journal of the American Statistical association* 101.476, p. 1418-1429.

Zou, Hui et Trevor Hastie (2005). « Regularization and variable selection via the elastic net ». In : *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, p. 301-320.