LASSO Bayésien
Projet de Statistiques Bayésiennes

Raphaël Huille & Mathis Linger

Janvier 2018

Table des matières

1 Présentation et analyse théorique **2**
- 1.1 Modèle Bayésien du LASSO . . . 2
- 1.2 Une pénalisation moindre avec la médiane . . . 5
- 1.3 Inférence avec Gibbs sampling . . . 6
- 1.4 Inférence du paramètre λ . . . 7

2 Implémentation et résultats empiriques **8**
- 2.1 Reproduction des résultats de l'article . . . 8
- 2.2 Choix des paramètres pour l'hyperprior de λ^2 . . . 11
- 2.3 Application en plus grande dimension . . . 12
- 2.4 Importance de la taille des données . . . 14

3 Références **15**

Introduction

La régression Lasso est une extension populaire de la méthode des moindres carrés qui peut être étudiée sous l'angle des statistiques bayésiennes. Ce nouveau point de vue permet alors d'exprimer un nouvel estimateur appelé "Lasso Bayésien". Nous définissons et expérimentons cet estimateur dans ce projet.
Le principal atout de la vision bayésienne est qu'elle offre une méthode de choix du paramètre d'estimation λ différente de la très (trop ?) populaire validation croisée.
Nous développons dans une première partie les aspect théoriques du Lasso Bayésien en nous appuyant sur l'article [1] de Park et Casella. Puis, dans la deuxième section, nous étudions expérimentalement le Lasso Bayésien sur les mêmes données que [1].
Nous nous sommes efforcés de développer l'analyse, notamment en section 1.2 où nous proposons une justification théorique à la constatation expérimentale "le Lasso pénalise plus que le Lasso Bayésien". En outre, nous avons comparé le Lasso Bayésien avec le Lasso sur un nouveau jeu de données, en plus grande dimension (73 prédicteurs) et où la sélection des prédicteurs est indispensable en raison de la sparsité de la base. C'est un cadre de l'utilisation de Lasso qui nous parait plus "traditionnel" pour le Lasso que le jeu de données de diabète.

1 Présentation et analyse théorique

On cherche à estimer $\beta \in \mathbb{R}^p$ dans le modèle linéaire suivant :

$$Y = \mu + X\beta + \varepsilon \qquad (\star)$$

On suppose d'emblée que : $\mu = 0$. Cela est sans perte de généralité car cela revient à remplacer Y par : $Y - \mu$ et y par : $y - \overline{y}$.

NOTATION

- $X = (X_1, ..., X_p)$ est un vecteur aléatoire sur $\mathbb{R}^p$
- Y est un vecteur aléatoire sur $\mathbb{R}$
- $x_j \in \mathbb{R}^n$ est un vecteur de n réalisations indépendantes de X_j i.e $x_{1j}, ..., x_{nj}$ sont i.i.d de même loi que X_j.
- $x = (x_{ij})_{(i,j)\in[1,n]\times[1,p]}$ la matrice de réalisation de X. Elle est de taille $n \times p$. x_{-j} représente la matrice x sans la colonne x_j.
- $y \in \mathbb{R}^n$ est un vecteur de n réalisations indépendantes de Y i.e $y_1, ..., y_n$ sont i.i.d de même loi que Y.
- $||a||_0 = \sum_i 1(ai \neq 0)$ la norme 0 correspond au nombre de coordonnées de a non nulles
- $||a||_1 = \sum_i |a_i|$ la norme 1 qui correspond à la somme des valeurs absolues des coordonnées
- $||a||_2 = \sqrt{\sum_i a_i^2}$ la norme euclidienne. NB : lorsque l'indice de la norme n'est pas renseigné, il s'agit par défaut de la norme euclidienne.
- $\mathcal{N}(\mu, \sigma)$ la loi normale ; $\mathcal{L}(\lambda)$ la loi de Laplace centrée ; $\mathcal{IG}(\alpha, \beta)$ la loi Inverse-Gamma ; $\mathcal{IN}(\mu, \lambda)$ la loi inverse-gaussienne.
- Dans les modèles présentés, lorsque la loi à priori n'est pas renseignée, elle est implicitement supposée non-informative constante.

1.1 Modèle Bayésien du LASSO

Rappels sur l'estimateur LASSO

L'estimateur LASSO peut se définir comme une pénalisation de l'estimateur des moindres carrés, qu'on rappelle ici :

Définition 1 : Estimateur des moindres carrés
L'estimateur des moindres carrés noté $\hat{\beta}^{LS}$ est défini par :

$$\hat{\beta}^{LS} \in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} ||y - x\beta||^2$$

Cet estimateur peut être compris comme une projection des données y, de dimension n sur $Vect\{x_j | j \in [1, p]\}$, espace de dimension inférieure ou égale à p. Lorsque $p < n$ il y a une contraction des données. L'inférence de Y est alors "synthétisée" en la connaissance de p régresseurs. Mais à l'inverse, si $p > n$, on comprend intuitivement qu'il n'est pas possible de "dilater" l'information. A partir de n observations, comment inférer correctement les p valeurs du vecteur $\hat{\beta}^{LS}$? Cela n'est possible qu'à la condition que l'hypothèse suivante soit vérifiée, ce que nous supposerons toujours par la suite :

Y ne dépend que d'un nombre $m < n$ de régresseurs. Le vecteur β est sparse.

Cette hypothèse est généralement vérifiée en grande dimension. Nous allons intégrer cette hypothèse dans l'estimateur Lasso via une pénalisation et dans le Lasso Bayésien via une prior judicieusement choisie.

Cette hypothèse impose que la norme 0 du vecteur β doit être petite. L'idée est alors de reprendre l'estimateur des moindres carrés en le forçant à avoir une petite norme 0. On introduit alors le paramètre λ qui va quantifier l'importance de la contrainte dans la minimisation de la fonction :

$$\beta \to ||y - x\beta||_2^2 + \lambda||\beta||_0 \quad (\star)$$

Si la fonction $\beta \to ||y - x\beta||_2^2$ est lisse et strictement convexe, ce qui facilite grandement sa minimisation, la fonction $\beta \to ||\beta||_0^2$ quant à elle, n'est ni lisse, ni convexe ce qui entraîne des (grosses) complications dans la minimisation de la fonction $(\star)$. L'estimateur LASSO contourne ce problème en utilisant la norme 1, qui est convexe, à la place de la norme 0. On parle de "convexification" du problème.

Définition 2 : Estimateur LASSO
L'estimateur des moindres carrés noté $\hat{\beta}^{LASSO}$ est défini par :

$$\hat{\beta}^{LASSO} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} ||y - x\beta||_2^2 + \lambda||\beta||_1$$

Vision bayésienne

Justifions maintenant la validité statistique de ces approches. **Il s'agit d'identifier les estimateurs définis dans la section précédente comme étant, le *mode*, la *médiane* ou la *moyenne* de la densité à posteriori d'un modèle bayésien.** On rappelle que dans les modèles présentés, lorsque la loi à priori n'est pas renseignée, elle est implicitement supposée non-informative constante.

Modèle 1 :

$$Y|X, \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma)$$

Densité à posteriori :

$$\pi(\beta|y, x, \sigma^2) \propto \exp\Big(-\frac{1}{\sigma^2}||y - x\beta||^2\Big)$$

Dans ce modèle, toutes les densités à priori sont plates. La densité à posteriori est donc égale à la vraisemblance. **L'estimateur des moindres carrés est égale au mode de la densité à posteriori du modèle 1.**

Pour obtenir l'estimateur Lasso, il suffit d'ajouter une loi à priori sur le paramètre β. A priori, comme on sait que β est sparse, les coordonnées β_j ont une forte probabilité d'être proche de 0. L'estimateur Lasso faisant intervenir la valeur absolue de ses coefficients, on pense à la loi de Laplace.

Modèle 2 :

$$Y|X,\beta,\sigma \sim \mathcal{N}(X\beta,\sigma)$$

$$\forall i,j \in [1,p]^2 \quad \beta_j \sim \mathcal{L}(\lambda) \quad \beta_j \perp\!\!\!\perp \beta_i$$

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

Densité à posteriori :

$$\pi(\beta|y,x,\sigma^2) \propto \exp\big(-\frac{1}{2\sigma^2}||y-x\beta||^2 - \lambda||\beta||_1\big)$$

Le mode de la densité à posteriori est similaire à l'estimateur Lasso. Il s'agit de la valeur de β en le minimum de la fonction :

$$\beta \to ||y-x\beta||^2 + 2\sigma^2\lambda||\beta||_1\big)$$

Cette vision de l'estimateur apporte un éclairage intéressant. Le paramètre de pénalisation, ici égal à $2\sigma^2\lambda$ est proportionnel à la variance du bruit du résidu. On retrouve ici une règle générale des modèles bayésiens : "plus les données sont bruitées, plus on doit faire confiance à la densité à priori". Si pour l'estimateur des moindres carrées, on peut exclure σ^2 de l'inférence, on comprend que ce paramètre jouera un rôle dans la pénalisation de l'estimateur Lasso Bayésien.
Dans l'estimateur Lasso de la définition 2, l'influence de σ^2 est simplement "absorbée" dans le paramètre λ. Le modèle bayésien approprié pour retrouver cet estimateur est le suivant.

Modèle 3 :

$$Y|X,\beta,\sigma \sim \mathcal{N}(X\beta,\sigma)$$

$$\forall i,j \in [1,p]^2 \quad \beta_j \sim \mathcal{L}(\frac{\lambda}{2\sigma^2}) \quad \beta_j \perp\!\!\!\perp \beta_i$$

Densité à posteriori :

$$\pi(\beta|y,x,\sigma^2) \propto \exp\big(-\frac{||y-x\beta||^2 + \lambda||\beta||_1}{2\sigma^2}\big)$$

On voit que : **l'estimateur LASSO est égal au mode de la densité à posteriori du modèle 3.**
Avoir une vision bayésienne de l'estimateur Lasso permet d'accéder à la grande variété des méthodes bayésiennes. Ainsi, au lieu de prendre le mode, Park et Casella suggèrent de prendre la médiane de la loi à posteriori comme nouvel estimateur. En outre, le modèle 2 suggère de prendre en compte l'influence de σ^2 sur la pénalisation. Pour cela, il faut considérer la loi jointe de (β,σ^2) Au lieu de supprimer cette influence comme dans le modèle 3, Park et Casella suggèrent de la prendre en compte en utilisant le modèle suivant :

Modèle 4 :

$$Y|X,\beta,\sigma \sim \mathcal{N}(X\beta,\sigma)$$

$$\forall i,j \in [1,p]^2 \quad \beta_j|\sigma^2 \sim \mathcal{L}(\frac{\lambda}{2\sqrt{\sigma^2}}) \quad \beta_j|\sigma^2 \perp\!\!\!\perp \beta_i|\sigma^2$$

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

Densité à posteriori :

$$\pi(\beta,\sigma^2|y,x) \propto \pi(\sigma^2)(\sqrt{\sigma^2})^{(n+p)/2}\exp\big(-\frac{1}{2\sigma^2}||y-x\beta||^2 - \frac{\lambda}{2\sqrt{\sigma^2}}||\beta||_1\big)$$

Définition 3 : Estimateur Lasso Bayésien
L'estimateur Lasso Bayésien est défini comme la médiane de β de la loi à posteriori du modèle 4.

La méthode d'inférence de cet estimateur suggérée par Park et Casella est présentée dans la section 1.3. Elle s'appuie sur l'algorithme Gibbs sampling qui nécessite, pour une meilleure convergence de la distribution, que $\pi(\beta,\sigma^2|y,x)$ soit unimodale, ce qui n'est pas vérifié dans le modèle 2 pour toutes les lois à priori sur σ^2 comme le montre Park et Casella.

1.2 Une pénalisation moindre avec la médiane

Dans cette section, nous étudions ce qu'implique de prendre la médiane de la loi à posteriori (Lasso Bayésien) au lieu du mode (Lasso classique). Ce changement va justifier l'observation empirique de Park et Casella dans la figure 1 de leur article : le Lasso Bayésien pénalise moins que le Lasso classique. Soulignons qu'on ne fait pas une démonstration mais une justification heuristique. **Soulignons également que cette justification théorique ne figure pas dans le papier de Park et Casella.**

Considérons la loi à posteriori d'une coordonnée de β dans le modèle 4 :

$$\beta_j|y,x,\beta_{-j},\sigma \propto \exp\big(-\frac{1}{2\sigma^2}||\tilde{y}-x_j\beta_j||^2 + -\frac{\lambda}{2\sqrt{\sigma^2}}|\beta_j|\big)$$

avec : $\tilde{y} = y - x_{-j}\beta_{-j}$
Dans le modèle 3, on obtient une fonction de la même forme, à cette différence près que le coefficient de pénalisation est $\frac{\lambda}{2\sigma^2}$ et non pas $\frac{\lambda}{2\sqrt{\sigma^2}}$. Mais cette différence est "artificielle" car en réalité, le paramètre λ ne sera pas choisi de la même manière dans les deux cas. On peut supposer que les λ choisis dans chacun des modèles feront coïncider le coefficient de pénalisation. Cette hypothèse est importante pour cette analyse car on veut comprendre l'influence du choix médiane/mode *à pénalisation égale*. Ainsi, β_j sera la médiane - dans le cas du Lasso Bayésien - et le mode - dans le cas du Lasso - d'une fonction de la forme :

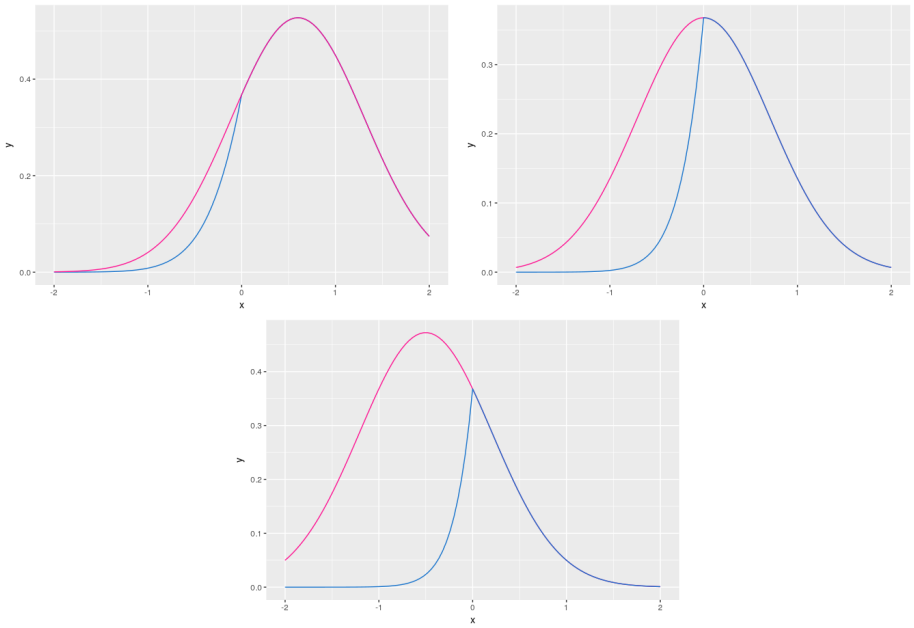
$$f : z \to \exp\big(-(a-z)^2 - b|z|\big)$$

Remarque : La norme $||\tilde{y}-x_j\beta_j||^2$ peut en effet s'écrire sous la forme : $(a-\beta_j)^2 + cst$, grâce à la méthode de réduction des formes quadratiques de Gauss. La constante est ensuite éliminée, car elle ne constitue qu'un simple facteur multiplicatif de la fonction f.

On peut montrer que le mode d'une telle fonction est toujours inférieur ou égale à la médiane. Cela justifie que le Lasso pénalise plus que le Lasso Bayésien. La figure 1 montre la forme d'une telle

fonction. La "cassure" en $x = 0$ vient de l'irrégularité de la fonction valeur absolue en 0. Ce résultat repose sur la comparaison de $g : z \to \exp\big(-(a-z)^2 - bz\big)$ et de f. Ces fonctions coïncident sur les valeurs positives, mais f est inférieur à g sur les valeurs négatives. g étant symétrique par rapport à une certaine valeur plus petite que a, le mode et la médiane de g sont identiques. En revanche, la médiane de f va être écartée du mode dans la direction où les valeurs sont les plus grandes, c'est à dire dans la direction opposée à la "cassure", i.e. la médiane sera un peu plus grande (en valeur absolue) que le mode. Notez que si le mode de g est négatif ou nul, le mode de f sera nul, donc l'estimateur Lasso sera nul, mais l'estimateur Lasso Bayésien sera non nul. Conclusion : le Lasso est nul plus facilement que le Lasso Bayésien.

FIGURE 1 – Tracé de $f : z \to \exp\big(-(a-z)^2 - b|z|\big)$ (en bleu) et de $g : z \to \exp\big(-(a-z)^2 - bz\big)$ (en rose) avec : $a = 1$, $b = 0.8$ (à gauche) ; $a = 1$, $b = 2$ (à droite) ; $a = 1$, $b = 3$ (en bas)



1.3 Inférence avec Gibbs sampling

Rappelons que l'estimateur Lasso est généralement utilisé en grande dimension, c'est à dire lorsque le nombre de régresseurs est grand. Dans ce cadre, β est un vecteur en grande dimension et l'algorithme à préconiser pour l'inférence du Lasso Bayésien est donc naturellement Gibbs Sampling. L'algorithme repose sur l'idée qu'il est plus facile de simuler des distributions de valeur dans $\mathbb{R}$ que des distributions de vecteurs dans $\mathbb{R}^p$. Soulignons que Gibbs sampling présente des similitudes avec l'algorithme à gradient de descente qui est très utilisé pour l'estimateur Lasso. Pour ces deux algorithmes il s'agit de résoudre un problème compliqué sur un vecteur dans $\mathbb{R}^p$ en résolvant p fois le même problème séparément sur les coordonnées du vecteur.

Implémenter Gibbs sampling pour le Lasso bayésien n'est pas direct. En effet, on ne sait pas simuler directement la forme de la loi conditionnelle des coordonnées de β. Afin de surmonter ce problème, Park et Casella utilise la représentation de la loi de Laplace suivante (équation (4) dans [1]) :

Lemme 1 : Représentation de la loi de Laplace
Si $z|r, s \sim \mathcal{N}(0, sr)$ et s $\sim \mathcal{E}(a^2/2)$ alors : $z \sim \mathcal{L}(\frac{a}{2\sqrt{r}})$

En posant $z = \beta_j$, $s = \tau_j$ et $a = \lambda^2$, $r = \sigma^2$ on peut réécrire le modèle 4 en :

Modèle 5 : (Modèle 4 réécrit)

$$Y|X,\beta,\sigma^2 \sim \mathcal{N}(X\beta,\sigma^2)$$

$$\forall i,j \in [1,p]^2 \quad \beta_j|\sigma^2,\tau_j \sim \mathcal{N}(0,\sigma^2\tau_j) \quad \beta_j|\sigma^2,\tau_j \perp\!\!\!\perp \beta_i|\sigma^2,\tau_j$$

$$\forall i,j \in [1,p]^2 \quad \tau_j \sim \mathcal{E}(\lambda^2/2) \quad \tau_j \perp\!\!\!\perp \tau_i$$

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

Remarque : le modèle 5 correspond à (5) dans [1] avec quelques différences mineures dues à des choix que nous avons fait. D'abord, par souci de simplicité, nous avons éliminé le paramètre μ. cf. équation $(\star)$. Ensuite, ce que [1] appelle "τ_j^2" nous l'avons écrit "τ_j"

A l'inverse du modèle 4, le modèle 5 fait intervenir des lois conjuguées, ce qui rend le calcul des postérieures conditionnelles faciles. Dans la liste ci-dessous, nous avons indiqué par une (*) les conjugaisons. Dans ce modèle, la seule postérieure conditionnelle qui ne soit pas conjuguée à la prior est celle des τ_j. En effet, les τ_j sont les variances des β_j qui suivent une loi normale. La prior conjuguée est donc la loi gamma inverse, mais nous sommes contraint par le lemme 1. Les lois conditionnelles à simuler à chaque itération du Gibbs Sampler sont :

-(*) $\beta \sim \mathcal{N}((x^Tx + D_\tau^{-1})x^Ty, \sigma^2(x^Tx + D_\tau^{-1}))$

- $\forall j,i \quad \dfrac{1}{\tau_j} \sim \mathcal{IN}(\sqrt{\lambda^2\sigma^2/\beta_j^2}, \lambda^2) \quad \tau_j \perp\!\!\!\perp \tau_i$

-(*) $\sigma^2 \sim \mathcal{IG}(n/2 + p/2, (y - x\beta)^T(y - x\beta)/2 + \beta^T D_\tau^{-1}\beta/2)$

En notant : $D_\tau = \text{diag}(\tau_1, ...\tau_p)$

Soulignons que dans ce Gibbs Sampler, β est mis à jour à chaque itération "en bloc". Pas besoin de le simuler coordonnée par coordonnée.

1.4 Inférence du paramètre λ

L'un des principaux défis du Lasso est d'être capable de choisir λ. Le Lasso Bayésien ouvre la voie à de multiples méthodes bayésiennes pour un tel problème. Ces méthodes offrent des solutions alternatives à la validation croisée habituellement utilisée avec le Lasso. Park et Casella proposent deux méthodes d'estimation de λ que nous présentons ici. Ces deux méthodes ont l'avantage de simplement compléter l'algorithme Gibbs sampling, décrit dans la section précédente, pour obtenir une estimation de λ

Estimation avec l'algorithme Espérance-Maximisation

La première manière pour estimer λ repose sur la maximisation de la vraisemblance marginale. La vraisemblance marginale en λ s'écrit : (il s'agit de la loi postérieure jointe de toutes les variables, à laquelle on "retire" tous les termes n'incluant pas λ)

$$\text{L}(\lambda;\tau_j) \propto \prod_j \frac{\lambda^2}{2}\exp(\frac{\lambda^2}{2}\tau_j)$$

Soulignons qu'il s'agit de la vraisemblance d'une loi exponentielle de paramètre $\lambda^2/2$. Le maximum d'une tel vraisemblance a une forme analytique (que l'on calcule en annulant la dérivée de $log(\text{L})$) qui nous donne :

$$\lambda^* = \sqrt{\frac{2p}{\sum_j^p \tau_j}}$$

Problème : les (τ_j) sont des variables latentes, leurs valeurs sont inconnues. L'algorithme Espérance-Maximisation est une méthode populaire pour ce genre de maximisation. Il s'agit de maximiser itérativement (comme Gibbs Sampling) l'espérance de la vraisemblance sur des variables latentes. Dans ce point de vue, les variables latentes sont toutes celles qui ne ne sont pas directement liées à λ dans le modèle, les données étant (τ_j). Park et Casella montrent qu'il suffit de rajouter à chaque étape k du Gibbs sampling le calcul d'une nouvelle valeur de λ dépendant des valeurs des τ_j simulées au cours des étapes précédentes :

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_j^p \sum_i^k \tau_j^{(i)}}}$$

Mettre une prior sur λ^2

Jusqu'à présent nous avons considéré λ comme un hyperparamètre. Park et Casella suggèrent de simplement l'intégrer dans le modèle en lui donnant une prior. Le modèle 5 faisant intervenir λ^2 comme le paramètre d'une loi Exponentielle, il est judicieux de choisir la loi conjuguée pour λ^2, c'est à dire la famille des loi Gamma. Cela permet de compléter simplement le modèle 5 en :

Modèle 6 :

$$Y|X, \beta, \sigma \sim \mathcal{N}(X\beta, \sigma)$$

$$\forall i, j \in [1, p]^2 \quad \beta_j|\sigma, \tau_j \sim \mathcal{N}(\sigma^2\tau_j) \quad \beta_j|\sigma, \tau_j \perp\!\!\!\perp \beta_i|\sigma, \tau_j$$

$$\forall i, j \in [1, p]^2 \quad \tau_j|\lambda \sim \mathcal{E}(\lambda^2/2) \quad \tau_j|\lambda \perp\!\!\!\perp \tau_i|\lambda$$

$$\lambda^2 \sim \mathcal{G}(r, \delta)$$

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

Avec r et δ deux nouveaux hyperparamètres à choisir. La loi conditionnelle de λ^2 qui s'ajoute à la liste des variables à simuler à chaque itération du Gibbs Sampler est alors :

- $\lambda^2 \sim \mathcal{G}(r + p, \delta + \sum_j \tau_j/2)$

Comparée à la première méthode - maximiser la vraisemblance - cette deuxième méthode - mettre une prior sur lambda - présente le désavantage de reporter le problème du choix de λ sur le choix des hyperparamètres r et δ.

2 Implémentation et résultats empiriques

2.1 Reproduction des résultats de l'article

Dans leur article, Park et Casella réalisent une expérience sur la célèbre base de donnée 'diabetes data' d'Efron et al. (2004). Cette base est constituée de 442 observations et 10 prédicteurs (âge, sexe, body mass index, pression sanguine moyenne, et six mesures de sérum sanguin). La variable que l'on cherche à estimer est continue et représente la progression du diabète 1 an après avoir mesuré les variables de contrôle. Les données sont standardisées de manière à ce que la moyenne

de chaque prédicteur soit zéro, et la variance 1.

Nos implémentations des deux méthodes proposées pour l'inférence du paramètre de régularisation λ sont justes puisque nous trouvons les mêmes régularisations que Park et Casella. En effet, en utilisant l'algorithme EM Monte Carlo proposé, nous trouvons un $\lambda = 0.237$, après 30 itérations nécessitant chacune 1,000 simulations (exactement égal à celui trouvé dans l'article). De la même manière, en intégrant dans le modèle hiérarchique une loi a priori de type gamma pour λ^2, et en utilisant la distribution conditionnelle complète associée (avec les paramètres suggérés dans l'article [1]) dans le Gibbs sampler, nous trouvons que la médiane a posteriori pour λ est égale à 0.275. Ce résultat obtenu en simulant 10,000 λ^2 est équivalent à celui trouvé dans l'article (0.279).

Maintenant que nous savons que nos implémentations des méthodes d'inférence du paramètre de régularisation λ sont justes, nous pouvons reproduire les résultats empiriques sur lesquels les conclusions de Park et Casella reposent. Comme eux, comparons les coefficients suivants :

— les coefficients obtenus en prenant la médiane a posteriori pour le Lasso bayésien (et l'intervalle crédible associé à 95% (equal-tailed)), avec la régularisation λ sélectionnée par la méthode du maximum de vraisemblance

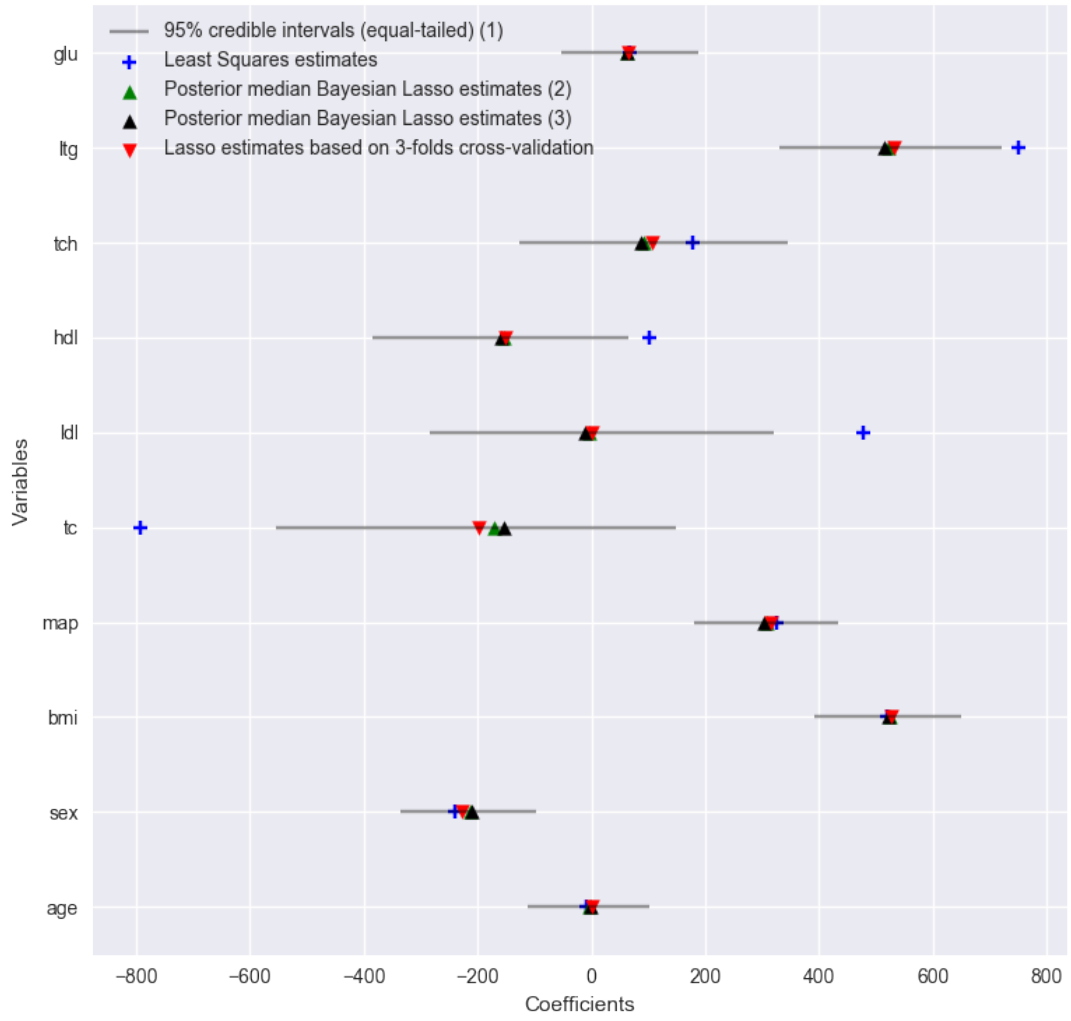
— les coefficients obtenus en prenant la médiane a posteriori pour le Lasso bayésien avec la régularisation λ sélectionnée en utilisant une loi a priori sur la régularisation λ

— les coefficients obtenus en utilisant le Lasso basique (et en sélectionnant la régularisation λ par validation croisée)

— les coefficients obtenus par moindres carrés ordinaire

1. Voir la section 2.3 Choix des paramètres pour l'hyperprior de λ^2

FIGURE 2 – Estimates of Lasso and non Lasso regressions on Diabete data



(1) with λ selected according to marginal maximum likelihood
(2) λ selected according to marginal maximum likelihood
(3) λ selected using hyperpriors for the Lasso Parameter

Quatre conclusions ressortent particulièrement : (i) le vecteur des coefficients obtenus en prenant la médiane a posteriori pour le Lasso bayésien n'est pas sparse au sens strict du terme ; (ii) les coefficients obtenus en prenant la médiane a posteriori pour le Lasso bayésien sont très similaires à ceux obtenus par le Lasso simple ; (iii) tous les coefficients du Lasso basique sont dans les intervalles crédibles à 95% (associés aux médianes a posteriori des coefficients du Lasso bayésien) ; (iv) 4 des coefficients des moindres carrés ordinaire (dont l'un est significatif) ne sont pas dans les intervalles crédibles à 95%.

Cette étude, met en exergue les qualités du Lasso bayésien. Toutefois, elle est réalisée sur une base assez petite (442 observations et 10 prédicteurs). Or, du fait qu'il fonctionne en grande dimension (cas où le nombre d'individus est inférieur au nombre de variables) et qu'il permette de sélectionner un sous-ensemble restreint de variables (retourner un modèle sparse), la technique du Lasso est souvent utilisée pour des problèmes en grande dimension. De ce fait, nous pensons qu'il serait plus pertinent de reproduire cette étude sur une base plus conséquente, afin de voir si le Lasso bayésien se comporte de la même manière.

2.2 Choix des paramètres pour l'hyperprior de λ^2

Comme montré dans l'article, une alternative à l'utilisation de l'algorithme Monte Carlo EM est l'ajout d'une loi a priori sur le carré de l'hypermaramètre λ dans le modèle hierarchique. La loi a priori (sur λ^2) suggérée est une Gamma de la forme :

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2} \text{ avec } \lambda^2 > 0 \ , \ r > 0 \ , \ \delta > 0$$

La distribution conditionnelle complète pour λ^2 ajoutée au Gibbs sampler est par suite une Gamma :

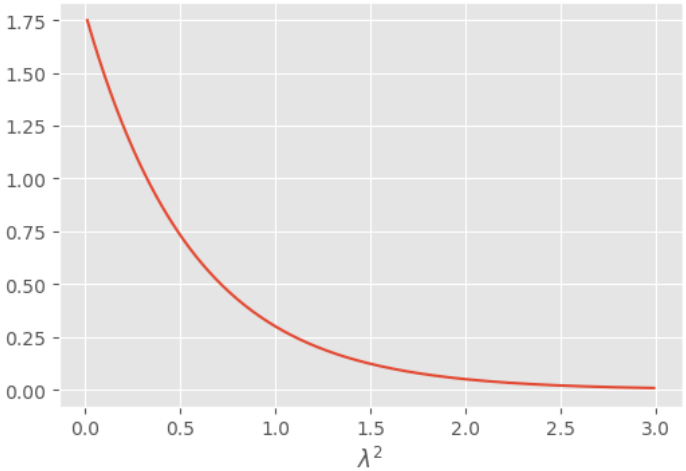
$$\mathcal{G}amma(p + r, \frac{\sum_{j=1}^{p} \tau_j^2}{2} + \delta)$$

où p est le nombre de prédicteurs.

Même si l'ajout d'une hyperprior sur λ^2 nous permet de ne pas avoir à choisir la régularisation λ, nous avons tout de même à choisir avec précaution les paramètres (shape et rate) de la loi a priori pour λ^2. Afin de fixer ces paramètres, nous pouvons utiliser la valeur du λ donnée par l'algorithme Monte Carlo EM. En effet, deux caractéristiques souhaitables de la distribution a priori pour λ^2 sont : (i) qu'elle approche assez vite 0 quand $\lambda^2 \to \infty$; (ii) qu'elle place une haute probabilité aux alentours de la valeur de λ donnée par l'algorithme Monte Carlo EM.

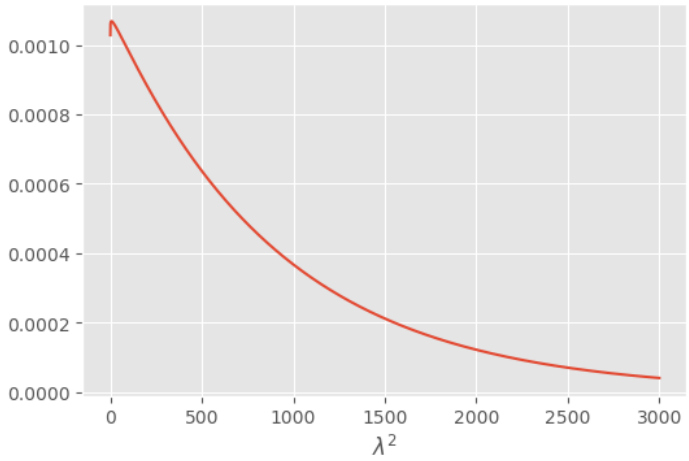
Par exemple, pour la base de donnée diabete utilisée dans l'article, les auteurs proposent de prendre une *shape* $= 1$ et un *rate* $= 1.78$, afin d'obtenir une distribution a priori sur λ^2 qui soit exponentielle et de moyenne égale à 10 fois le paramètre inféré par maximum de vraisemblance (i.e. par l'algorithme Monte Carlo EM). La Figure 2 représente cette distribution. Nous voyons en effet qu'elle décroit rapidement vers 0, qu'elle porte plus de poids aux alentours de 0.237 (valeur de λ inférée par maximum de vraisemblance), et que son espérance ($\frac{1}{1.78} \approx 0.561$) est très proche de 10 fois le paramètre inféré par maximum de vraisemblance ($0.237^2 \times 10 \approx 0.561$).

FIGURE 3 – Prior Gamma density for λ^2 with shape 1 and rate 1.78



En procédant de la même manière, nous décidons choisissons pour notre loi gamma a priori les paramètres *shape* $= 1.007$ et *rate* $= 0.001117$. La densité correspondante est représentée par la Figure 3. Nous observons un pic aux alentours de 9 (valeur de λ inférée par maximum de vraisemblance), et l'espérance de cette densité est de $\frac{1.007}{0.001117} = 901.52 \approx 9.49^2 \times 10 = 901.39$.

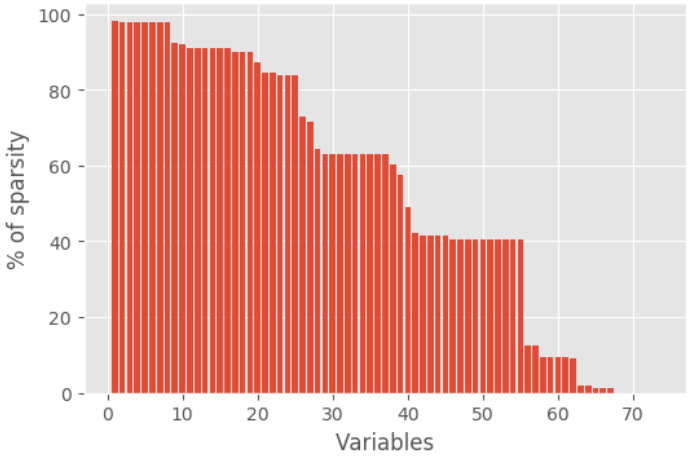
FIGURE 4 – Prior Gamma density for λ^2 with shape 1.007 and rate 0.001117



2.3 Application en plus grande dimension

Pour l'étude en grande dimension, nous utiliserons une base de données de physique. Cette base de données est composée de 4,000 observations et 73 prédicteurs. La variable à prédire est ici aussi une variable continue. Nous avons choisi ce jeu de données car il est très sparse. Cette spécificité du jeu de données nous servira entre autre à évaluer la première conclusion qui est que le vecteur des coefficients obtenus en prenant la médiane a posteriori pour le Lasso bayésien n'est pas sparse au sens strict du terme. En effet, plus de la moitié des variables ayant plus de 50% de zéros (Figure 4), nous espérons obtenir un modèle sparse en utilisant le Lasso.

FIGURE 5 – Sparsity of each variable in the dataset



Cette remarque sur la sparsité des données faite, reproduisons les résultats empiriques sur lesquels les conclusions de Park et Casella reposent (Figure 5). Cela nous permet de nuancer les conclusions faites par Park et Casella.

Tout d'abord, leur affirmation concernant la non sparsité (au sens strict du terme) du modèle obtenu en prenant la médiane a posteriori pour le Lasso bayésien semble s'avérer vraie. En effet,

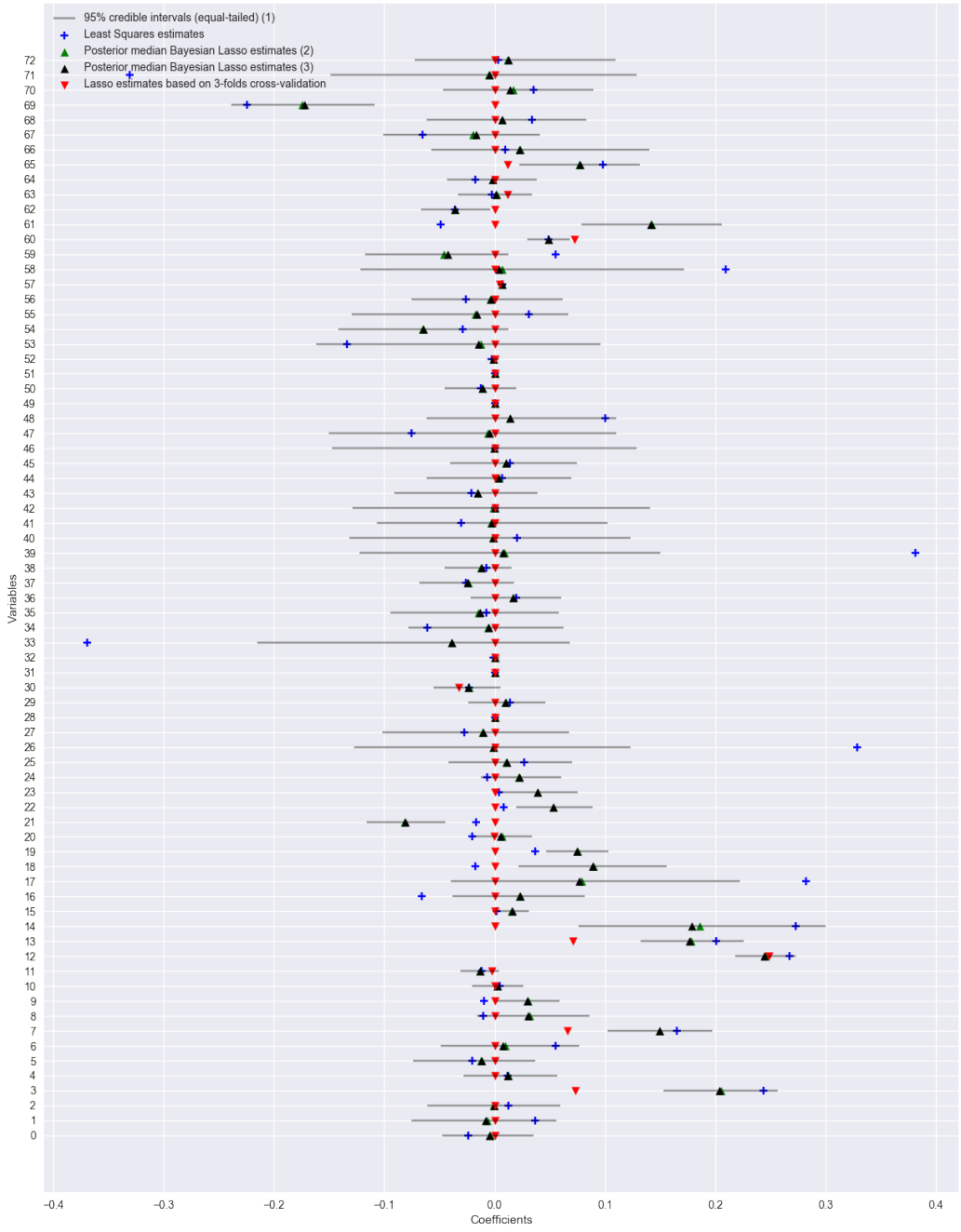
le Lasso bayésien renvoie un vecteur de coefficients qui n'est pas sparse. En effet, la plupart des coefficients du modèle bayésien sont compris entre la valeur du Lasso basique et la valeur des coefficients de la régression par moindres carrés ordinaire. Le lasso basique (avec le paramètre de régularisation choisi par 3-folds cross validation) renvoie un vecteur de coefficients sparse à 86% alors que le Lasso bayésien renvoie un vecteur de coefficients sparse à 17%. L'avantage premier du Lasso étant de renvoyer un vecteur de coefficients sparse, nous voyons ici que l'utilisation de l'approche bayésienne réduit fortement cet avantage.

Ensuite, une seconde conclusion était que les coefficients obtenus en prenant la médiane a posteriori pour le Lasso bayésien sont très similaires à ceux obtenus par le Lasso simple. Cela reste vrai dans l'ensemble, même si parfois les coefficients générés par le Lasso bayésien sont plus proches de ceux des moindres carrés ordinaires que ceux du Lasso simple.

Par ailleurs, Park et Casella concluaient que tous les coefficients du Lasso basique sont dans les intervalles crédibles à 95% (associés aux médianes a posteriori des coefficients du Lasso bayésien). Il est clair que cela n'est plus le cas ici. En effet, 16% des coefficients générés par le Lasso basique ne sont pas compris dans l'intervalle crédible à 95% des coefficients du Lasso bayésien.

Enfin, dans l'étude empirique réalisée sur les données diabete, 4 des 10 coefficients des moindres carrés ordinaire n'étaient pas dans les intervalles crédibles à 95%. Dans notre exemple, seulement 20% des coefficients des moindres carrés ordinaire ne sont pas dans les intervalles crédibles à 95%. La proportion annoncée est donc réduite de moitié dans notre exemple.

FIGURE 6 – Estimates of Lasso and non Lasso regressions on large data



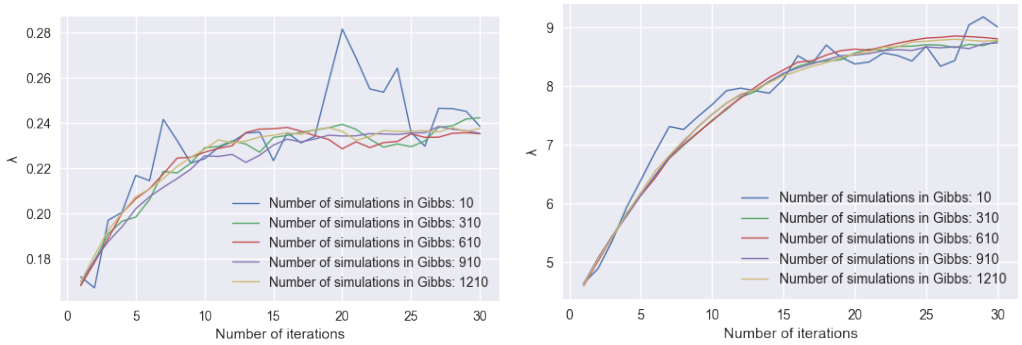
(1) with λ selected according to marginal maximum likelihood
(2) λ selected according to marginal maximum likelihood
(3) λ selected using hyperpriors for the Lasso Parameter

2.4 Importance de la taille des données

L'approche bayésienne, même si elle permet de ne pas devoir choisir le paramètre de régularisation directement (par validation croisée par exemple), n'est pas gratuite. En effet, l'algorithme Monte

Carlo EM est relativement coûteux car il nécessite un bon nombre d'itérations, et au sein de chaque itération la réalisation de simulations à partir d'un Gibbs sampler. L'autre solution qui consiste à mettre une loi gamma comme a priori sur les λ^2 est plus rapide mais nécessite le choix de deux paramètres (shape et rate). Les auteurs choisissant ces paramètres à partir de la valeur du λ inférée à l'aide de l'algorithme Monte Carlo EM, nous voyons bien que dans tous les cas un coût de calcul devra être assuré à un moment donné.

FIGURE 7 – Evolution of the lambda depending on the number of iterations for the Monte Carlo EM algorithm (left : on diabete data, right : on the physic data)



D'après la Figure 6, nous pouvons remarquer que plus le nombre de simulations réalisées au sein de chaque itération est grand, plus la variance du lambda inféré est petite. Cela est naturel et découle de la loi forte des grands nombres. Une seconde remarque est que le λ final semble être atteint à partir de la 15ème itération pour la base de donnée diabete alors qu'il semble n'être atteint qu'à partir de la 25ème itération pour le jeu de données plus grand. A noter que pour le jeu de données plus grand, le λ initial est plus éloigné du λ final que pour le jeu de données diabete. On voit bien ici une intuition se dessiner, qui est que plus le nombre de variables dans une base de données sera élevé, plus le paramètre de régularisation λ devrait être élevé, et plus l'algorithme Monte Carlo EM nécessitera d'itérations pour trouver le λ final (ce qui est coûteux).

Conclusion

Utiliser une base de donnée plus grande et plus sparse, nous a amené à nuancer certaines conclusions de l'article. En particulier, il n'est pas si évident que les coefficients du Lasso bayésien soient si alignés que cela avec ceux du Lasso basique.

D'autres part, même si l'approche bayésienne permet de ne pas avoir à choisir le paramètre de régularisation par validation croisée, la promesse d'une méthode beaucoup plus rapide et moins coûteuse en calculs n'est pas forcément tenue.

3 Références

[1] Trevor Park and George Casella (2008) "The bayesian lasso", *Journal of the American Statistical Association*, 103(482) :681–686

[2] Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression", *Annals of Statistics*, 407-499