Projet de Statistique Bayésienne

The Bayesian Lasso

Trevor Park and George Casella. Journal of the American Statistical Association, 103(482):681–686, 2008

Nicolas Schreuder, Sholom Schechtman, Amélie Segard

**Introduction**

Le Lasso est une méthode d’estimation des paramètres d’une régression linéaire permettant de réaliser à la fois une sélection de variables et une contraction des coefficients afin d’améliorer l’interprétation et la qualité de prédiction du modèle statistique obtenu. Le Lasso est notamment utilisé dans des domaines tels que la génétique ou le traitement de l’image où des problèmes de grande dimension (n<<p) se posent et rendent l’utilisation de techniques de régressions classiques difficile. Cette méthode, introduite par Robert Tibshirani en 1996, est fondé sur les travaux de Leo Breiman, *Better Subset Regression Using the Nonnegative Garrote* (1995).

Avant l’introduction du Lasso, la méthode la plus populaire pour améliorer la qualité de prédiction était la régression Ridge. Cette méthode permet de réduire le sur-apprentissage et donc l’erreur du modèle en introduisant un terme de régularisation, comme le Lasso, mais elle n’effectue pas de sélection de variable et ne permet donc pas de rendre le modèle plus interprétable. La sélection de variable était alors largement effectuée de façon itérative en utilisant des critères de sélection type AIC, BIC etc. Cependant ces méthodes améliorent la qualité de prédiction seulement dans certains cas, comme lorsque seuls certaines variables sont importantes.

Tout comme la régression Ridge, la régression Lasso peut s’interpréter en statistique Bayésienne : le Lasso peut être vu comme une régression linéaire dans laquelle les paramètres ont des distributions a priori de Laplace.

Cette approche bayésienne pourrait dans certains cas être plus performante que le modèle Lasso classique car elle pénalise et choisit les coefficients de régression en fonction des informations contenues dans les données mais aussi des connaissances a priori. Le *Lasso Bayésien* n’effectue pas en revanche une sélection des variables automatiquement mais fournit des régions crédibles pour tout les paramètres (y compris la variance de l’erreur) et ces régions crédibles peuvent guider la sélection des variables.

Comme la fonction objectif du Lasso n’est pas différentiable du fait de la forme de la pénalité l1, différents algorithmes ont été développés afin d'en trouver les solutions. Park et Casella ont montré que la formalisation Bayésienne, en utilisant un modèle hiérarchique avec une distribution à priori gaussienne pour les paramètres et une loi exponentielle indépendante pour leur variance, permet d’exploiter l’algorithme Gibbs sampling pour estimer le maximum a postériori.

D’autres algorithmes précédemment développés, tel que l’algorithme LARS par exemple, permettent de trouver des solutions au Lasso. Cependant Park et Casella soulignent que le *Lasso Bayésien* fournit automatiquement des écarts types pour tous les paramètres, contrairement aux autres méthodes pour lesquels les estimateurs des écart-types des paramètres estimés à 0 ne sont pas considérés comme entièrement satisfaisants. L’approche Bayésienne apporte par ailleurs une solution nouvelle au choix de la valeur du paramètre de régularisation qui permettrait d’aboutir à des estimations de Lasso plus stables.

Nous présentons ainsi ici, le *Lasso Bayésien* développé par Park et Casella en 2008. La première section expose le cadre théorique. La deuxième section présente les méthodes numériques employées dans la résolution du *Lasso Bayésien*. La troisième section présente des papiers associés. La dernière section développe une application sur des données réelles.

1. **Le cadre théorique**
   1. Le Lasso

La méthode du Lasso de Tibshirani est le plus souvent appliqué au modèle de régression linéaire. Considérons le modèle linéaire standard où l’on observe (yi,xi), i {1,…,n} avec yi variable réponse, xi p variables explicatives et

yi = + xi1i1 + … xipip + avec 1,…,p) et i inconnus.

On cherche à estimer le vecteur des coefficients de la régression.

Sous les hypothèses,

(H1) E(i) = 0

(H2) V(i) = 2 (homscédasticité)

(H3) i indépendant de j

(H4) i N(0,2)

Sous (H1),…,(H4) = n(0n,2In) avec In matrice identité de taille (n,n)

On suppose par ailleurs que les variables explicatives sont linéairement indépendantes.

Sous les hypothèses ci-dessus, en notant la moyenne des éléments, on a (a)

D’où en notant on a (b)

Il est donc standard de centrer les variables.

En passant à l’écriture matricielle avec y = , X=, = , on obtient,

= + (1)

Ainsi dans le modèle décrit ci-dessus, nous considérons que les variables explicatives X=(X1,…,Xp) sont centrées et également réduites afin que les coefficients estimés ne dépendent pas de l’échelle de mesure des variables.

Dans le cadre du modèle linéaire standard, la méthode des moindres carrés ordinaires est généralement utilisée pour estimer les paramètres : les coefficients sont obtenus par minimisation de la somme des carrés des résidus.

Avec la méthode Lasso, le vecteur des coefficients est également obtenu en minimisant la somme des carrés des résidus mais sous une contrainte supplémentaire :

22 sous la contrainte 1

sous la contrainte (2)

La contrainte est une pénalisation de la norme l1 des coefficients. Le paramètre t contrôle le niveau de régularisation.

Ainsi, le Lasso est souvent considéré comme une méthode d’estimation par moindre carrés pénalisés par la norme l1 et le problème (2) est réécrit sous la forme suivante, avec :

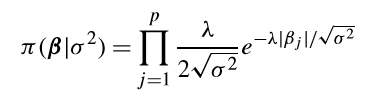
(3)

Cette pénalité va contracter la valeur des coefficients et contraindre certains coefficients à valoir zéro en forçant la somme de la valeur absolue des coefficients de régression à être intérieur à une valeur fixe. Cette idée est donc similaire à la régression Ridge dans laquelle la somme des carrées des coefficients doit être inférieure à une valeur fixe, mais dans le cas de la régression Ridge, cela ne fait que réduire la taille des coefficients. C’est ainsi la nature de la contrainte exercée sur les coefficients dans le cadre du Lasso qui permet de sélectionner un sous-ensemble de variables.

* 1. Interprétation Bayésienne du Lasso

Tout comme la régression Ridge qui peut être vu comme une régression linéaire dans laquelle les paramètres ont des distributions a priori de Gaussiennes, la régression Lasso peut s’interpréter en statistique Bayésienne.

L’expression (3) suggère, en effet que l’estimation par la méthode du Lasso peut s’interpréter comme l’estimation du maximum a posteriori lorsque les paramètres de la régression, conditionnellement à 2, ont une loi a priori i.i.d de Laplace L(0,).

 (4)

En effet, notons (2) la loi marginale à priori de 2.

La loi jointe à priori s’écrit alors

(c) (5)

Sous (H1),…,(H4), n(,2In) et la vraisemblance du modèle est (d) (6)

La loi jointe à postériori est donnée par

(e) (7)

Le log de la loi jointe à postériori est donc

(f) (8)

Ainsi pour tout 2 > 0, en maximisant cette expression par rapport à β on obtient l’estimateur Lasso. Par conséquent, l’estimateur du maximum à postériori, correspond à l’estimateur Lasso. Cet estimateur dépendra du choix de et de la loi à priori de 2. Nous compléterons cette spécification avec la loi à priori (impropre) invariante par changement d’échelle (2) = pour 2 mais une loi inverse-gamma conviendrait également. Notons que le conditionnement par rapport à2 dans (4) permet de garantir que la loi à postériori (7) est unimodale.

Nous pouvons également noter que nous aurions obtenu ceci également en conditionnant par rapport à y si nous avions intégré par rapport à μ avec une loi à priori

Enfin, notons que la distribution de Laplace présente un pic en zéro (sa dérivée première est discontinue) et concentre sa masse de probabilité plus près de zéro que la distribution Gaussienne. C’est une autre façon de comprendre comment le Lasso tend à fixer certains coefficients à zéro et permet d’effectuer une sélection des variables contrairement à la régression Ridge.

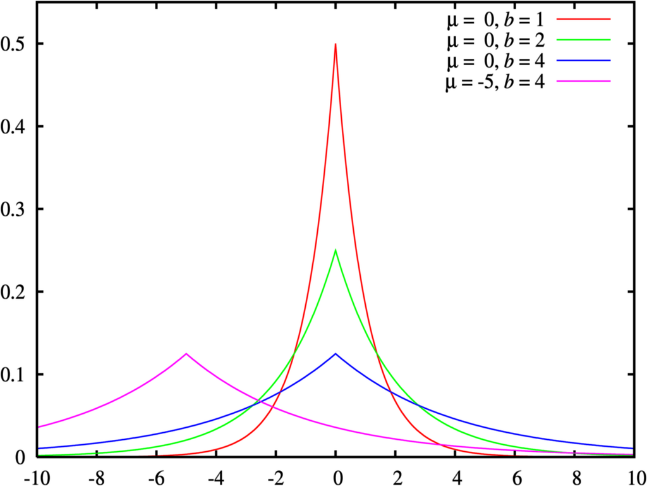


Fig. 1. Distribution de Laplace pour différents paramètres ()

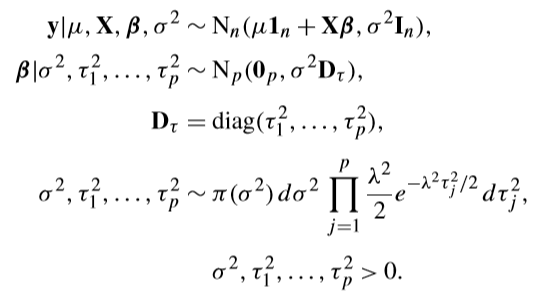
Cette interprétation du Lasso était déjà suggérée par Tibshirani en 1996 et a motivé plusieurs auteurs à utiliser des lois a priori de type Laplace.

Finalement l’estimateur du maximum à postériori, n’étant pas un véritable estimateur Bayésien, l’analyse Bayésienne suggère plutôt d'utiliser la moyenne ou la médiane à postériori pour estimer β. Ces estimateurs, ne permettent plus une sélection automatique des variables mais les coefficients sont « rétrécis » de façon similaire et les régions crédibles obtenues pour les estimateurs peuvent guider la sélection de variables.

* 1. Ecriture du Lasso sous la forme d’un modèle Bayésien hiérarchique

Comme la fonction objectif du Lasso n’est pas différentiable (car la norme l1 n’est pas différentiable en zéro), différents algorithmes fondés sur l’optimisation convexe ont été développées pour résoudre (3). On peut évoquer les algorithmes de sous-gradient (généralisation naturelle des méthodes de descente de gradient au cas où la fonction objectif n’est pas différentiable en tout point), proximal gradient et l’algorithme LARS (Least-Angle Regression)

Park et Casella proposent d’utiliser une autre méthode d’estimation fondée sur l’interprétation bayésienne du Lasso. Il s’agit d’utiliser l’algorithme Gibbs Sampling avec la représentation hiérarchique suivante du modèle :

(9) nb : mettre plutôt

Le modèle hiérarchique ci-dessus correspond au modèle bayésien présenté dans la section précédente. En exploitant la représentation de la distribution de Laplace en tant que mélange de normales (avec une densité de mélange exponentielle) nous obtenons, en effet, en intégrant (|2, ) par rapport à la loi à priori de sachant 2 voulue (3).

En effet nous avons, (g)

On reconnait la fonction génératrice des moments de la loi Laplace L(0,).

Nous avons ainsi montré que, (h)

La distribution de Laplace peut s’exprimer en tant que mélange de gaussienne et (|2) est de la forme (3) désirée. On conclut que nous obtenons le modèle bayésien souhaité.

Il est aisé de déterminer les lois conditionnelles |2,  ; 2|, et |2,  j {1,…p}.

En effet, on a (i)

Ainsi

(j) (10)

* 1. Extensions à rédiger

Le lasso possède d’intéressantes variantes et avec une autre distribution pour τ2 le modèle bayésien hiérarchique présenté (9) peut être utilisé pour produire des versions bayésiennes de ces méthodes

* 1. Le choix fondamental du paramètre de régularisation

Le choix du paramètre de régularisation () est fondamental pour obtenir une bonne performance du Lasso car il contrôle la contraction des coefficients et la sélection des variables, ce qui peut améliorer la qualité de prédiction et l’interprétation du modèle.

Si le paramètre de régularisation est trop grand, des variables importantes peuvent être exclues du modèle et les coefficients peuvent être réduits de façon excessive, ce qui peut nuire à la capacité prédictive et aux inférences tirées. En pratique, le Lasso est testé pour différentes valeurs de . Un chemin de solution représentant l’évolution des coefficients en fonction de est ainsi obtenu. L’algorithme LARS permet de calculer le chemin de solution de façon efficace, ce qui rend la détermination de la valeur optimale du paramètre de régularisation beaucoup plus simple. En pratique, la validation croisée est également souvent utilisée pour déterminer la valeur de ce paramètre.

Park et Casella proposent deux alternatives afin de choisir la valeur de dans le cadre du *Lasso Bayésien*. …

1. **La méthode d’estimation du Lasso Bayésien**
   1. Les méthodes numériques employées à rédiger

L’estimation du Lasso Bayésien exploite les méthodes statistiques de Monte Carlo (MCMC) afin de simuler la loi à postériori et estimer les paramètres du modèle.

* + 1. L’algorithme Gibbs sampling
    2. L’algorithme Monte Carlo EM
    3. L’algorithme Importance Sampling

1. Ajustement du modèle : méthodes d’estimation et implémentation des algorithmes de résolution du Lasso Bayésien
   * 1. L’estimation des coefficients de régression par l’algorithme Gibbs Sampling

Nous détaillons ici l’implémentation du Gibbs sampler associée au modèle hiérarchique (9). Rappelons que la loi à postériori est unimodale ce qui permettra d’assurer une convergence rapide de l’algorithme Gibbs sampling exposé ci-après et d’obtenir des estimateurs ponctuels significatifs.

Un échantillon de la loi à postériori ,2,) est générée à partir de l’algorithme Gibbs sampling fondé sur les lois conditionnelles de (10). L’algorithme répète itérativement les 3 étapes suivantes :

(k) (11)

L’algorithme de Chhikara et Folks (1989), section 4.5, permet de simuler une variable aléatoire de loi Inverse-Gaussienne. Une variante numériquement stable de cet algorithme est par ailleurs implémentée dans le package statmod de R (Smyth, 2005).

* + 1. La détermination du paramètre de régularisation par l’algorithme Monte Carlo EM

En considérant le modèle hiérarchique (9) comme un modèle paramétrique, le paramètre λ a une fonction de vraisemblance qui peut être maximisée pour obtenir une estimation empirique de Bayes.

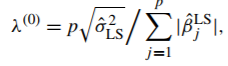
Casella (2001) a proposé un algorithme Monte Carlo EM qui complémente l’implémentation du Gibbs sampler. Les étapes de cet algorithme sont les suivantes :

1. Générer un échantillon de la loi à postériori ,2,) en utilisant le Gibbs sampler ci-dessus, avec λ= λ(k)
2. Evaluation de l’espérance (étape E) : Approximer l’espérance de la log vraisemblance des données complétées pour λ conditionnellement à λ(k) en substituant les moyennes de l’échantillon du Gibbs sampler à tous les termes comportant ,2 et
3. Maximisation (étape M) : déterminer λ(k+1) en maximisant l’espérance de la log vraisemblance des données complétées par rapport à λ (12)

Ces étapes sont répétées jusqu’à ce que le niveau de convergence souhaité soit obtenu.

Notons que (l)

Les auteurs suggèrent de prendre pour valeur initiale de λ,

 (13)

où et sont les estimateur des moindres carrés ordinaires.

/!\ Compléter avec dernier paragraphe sur Importance sampling.

* + 1. La détermination du paramètre de régularisation en utilisant une loi à priori sur

Park et Casella propose une autre solution pour le choix de la valeur de λ. Il s’agit de donner à λ2 une loi à priori diffuse. Les auteurs proposent d’utiliser la loi gamma suivante

 (14)

Cette loi à priori a pour avantage d’assurer que la distribution à postériori est unimodale et non impropre.

En utilisant (14) dans le modèle hiérarchique (9) la loi conditionnelle ,2,) est une loi gamma de paramètre (p+r, ).

En effet,

(m)

Les lois conditionnelles des autres paramètres (10) restent inchangées. Il est dès lors simple d’ajouter ce paramètre à l’algorithme Gibbs sampling (11).

/!\ Compléter avec le dernier paragraphe : The prior density for λ2 should approach 0 sufficiently fast as λ2 → ∞ (to avoid mixing problems) but should be relatively flat and place high probability near the maximum likelihood estimate.

1. **Discussion** (sorte de placement dans la littérature) à rédiger
   1. Les algorithmes antérieurs

<http://www.stat.washington.edu/courses/stat527/s13/readings/bayeslasso.pdf> page 7.

* 1. Les alternatives proposées (sur la base d’articles plus récents)

Minjung Kyung , Jeff Gill , Malay Ghosh and George Casella, Penalized Regression, Standard Errors, and Bayesian Lassos, 2010

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.296.8397&rep=rep1&type=pdf>

Bala Rajaratnam, Steven Roberts, Doug Sparks et and Onkar Dalal, Lasso Regression: Estimation and Shrinkage via Limit of Gibbs Sampling, 2015

La plupart des propriétés statistique des estimateurs Lasso en grande dimension sont souvent prouvé sous l’hypothèse que la corrélation entre les prédicteurs est bornée. D’autre part, la vitesse computationnelle des algorithmes se dégrade lorsque le vrai vecteur de solution n’est pas creux (non sparsité) et que et que la multi-colinéarité augmente. Or, le manque de parcimonie et une multi-colinéarité élevée peut être assez courante dans les applications contemporaines, la sélection de modèle est toujours une nécessité dans de tels contextes. Motivé par les limitations des algorithmes de *descente de coordonnées circulaire* dans ce contexte, Bala Rajaratnam, Steven Roberts, Doug Sparks et and Onkar Dalal ont proposé en 2015 un nouvel algorithme fondé sur le *Lasso Bayésien* et le Gibbs sampler associé nommé *Le Lasso Bayésien déterministe*. Leur étude montre notamment que cet algorithme présente de meilleures performances computationnelles comparé à la *descente de coordonnées circulaire,* lorsque le vecteur des solutions n’est pas creux et que la multi-colinéarité augmente.

Geometric Ergodicity of Gibbs Samplers in Bayesian Penalized Regression Models <https://arxiv.org/pdf/1609.04057.pdf>

Inference with normal-gamma prior distributions in regression problems <https://projecteuclid.org/download/pdf_1/euclid.ba/1340369797>

1. **Une application sur des données réelles** à faire
   1. Les données exploitées : présentation et analyse exploratoire
   2. L’ajustement du modèle
   3. La présentation des résultats

**Conclusion** à faire

**Annexes**

**Références**

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

[2] Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88.

[3] Andrews, D. F., and Mallows, C. L. (1974), “Scale Mixtures of Normal Distributions,” Journal of the Royal Statistical Society, Ser. B, 36, 99–102.

[4] Christian P. Robert and George Casella. Monte Carlo Statistical Methods, Springer, 2014

[5] George Casella, Empirical Bayes Gibbs Sampling, 2001

Truc que j’ai pris dans un livre d’apprentissage statistique pour la prédiction, il y a peut être des idées intéressantes à intégrer :

La variance d’un modèle augmente et son biais diminue avec l’accroissement de sa complexité. Il est nécessaire de trouver le bon niveau de complexité, assurant un juste équilibre entre la variance et le biais, et évitant dans un modèle de régression, les inconvénients liés à la multi-colinéarité des prédicteurs.

La régression Lasso est plus adaptés aux cas où de nombreux coefficients à estimer sont faibles.

On peut voir le paramètre comme un « curseur » qui permet de choisir continûment plus de biais et moins de variance, de façon à optimiser la prédiction.

Avec la pénalisation l2 (Ridge), la pénalisation la plus couramment employée en régression est la pénalisation l1 (Lasso). Elles possèdent des propriétés différentes, adaptés à des situations et des objectifs différents.

Le lasso est une méthode assez prisée, notamment lorsque le nombre p de variables est très grand. En effet, lorsque p est très grand, une bonne généralisation du modèle exige une pénalisation de plus en plus forte pour diminuer sa complexité. Si de plus fortes colinéarités existent entre les variables, la régression Ridge tend à rétrécir les coefficients de variables corrélées vers des valeurs proches. Il est alors plus pertinent et plus lisible pour le modèle de supprimer certaines variables, ce que fait le Lasso. Cela n’est pas gênant lorsque l’on peut supposer nuls un assez grand nombre de coefficients (hypothèse de sparsité). Mais dans certains cas, comme en génomique, il peut être souhaitable de conserver plus de variables.