Lasso pour données censurées à gauche : une comparaison par simulation d'algorithmes proposés dans la littérature

Perrine Soret [1,2,3,*] & Marta Avalos [1,2,3] & Linda Wittkop[1,2,3,4] & Rodolphe Thiebaut [1,2,3,4] & Daniel Commenges[1,2,3]

[1] *Univ Bordeaux, ISPED, INSERM, Centre INSERM U897, Bordeaux*
[2] *INRIA-SISTM, Bordeaux*
[3] *Vaccine Research Institute (VRI), Creteil*
[4] *CHU de Bordeaux*
[*] *perrine.soret@isped.u-bordeaux2.fr*

Résumé. Dans le cas de la recherche contre le VIH, lorsque la sensibilité d'une technique de dosage utilisée pour quantifier la charge virale, est faible, certaines valeurs sont censurées à gauche. Il existe un seuil de quantification analytique, en dessous duquel la valeur exacte de la mesure n'est pas connue, les concentrations sont dites indétectables. Cependant, même incomplètes, ces données apportent de l'information et méritent d'être conservées dans l'analyse. Nous proposons une comparaison par simulation de différents algorithmes proposés dans la littérature qui prennent en compte la censure dans une étude de grande dimension et dont les implémentations sont disponibles. Les méthodes ont été adaptées à l'hypothèse de données censurées gaussiennes.

Mots-clés. Grande dimension, Épidémiologie clinique, seuil de détectabilité

Abstract. In the case of the research against HIV, when the sensitivity of an assay used to quantify viral load is low, some values are left-censored. There is an analytical quantification threshold, below which the exact value of the measurement is not known, the concentrations are undetectable. However, even incomplete, these data give information and should be retained in the analysis. We propose a comparison by simulation of various algorithms proposed in the literature, which taking into account the censure in a high-dimensional study and whose the implementations are available. The methods have been adapted to the hypothesis of Gaussian censored data.

Keywords. High-dimensional, clinical Epidemiology, detectability threshold

1 Contexte

D'un point de vue biologique, on cherche à savoir quelles sont les mutations responsable de l'évolution de la charge virale chez des patients atteint du VIH. Après l'injection d'un traitement antirétroviral, on observe une chute de la charge virale, passant sous un seuil de détectabilité. En effet, il existe un seuil de quantification analytique en dessous duquel la valeur exacte de la mesure n'est pas connue, cette valeur est dite "indétectable". D'un point de vue statistique, ces données sont censurées à gauche, au seuil de détectabilité. Mais même incomplètes, ces données apportent de l'information et méritent d'être conservées dans l'analyse. Comment sont-elles pris en compte dans une étude de grande dimension ? Jusqu'à présent, la variable réponse était dichotomique : 0 pour le succès, c'est à dire que la charge virale n'était pas repassé au dessus du seuil de détectabilité et 1 pour l'échec thérapeutique, soit le cas où la charge virale est à nouveau détectable. Cette solution est cependant trop dure. Le but est donc de prendre en compte de façon simultanée la censure à gauche et la grande dimension tout en gardant la valeur réelle de la charge virale. Plusieurs méthodes et implémentations existent dans la littérature utilisant l'imputation des données censurées par l'estimateur Buckley-James [1]. De plus, la grande dimension est traitée par des méthodes de machine learning mais celles proposées dans la littérature pour ce type de problème [4, 5, 6, 9] se focalisent sur la prédiction alors qu'il est important pour le clinicien de comprendre quelles sont les mutations en cause.
L'objectif de notre étude est de comparer ces méthodes disponibles et d'évaluer leur performance en termes de sélection de variables et non de prédiction.

2 Modèle, méthode d'estimation et critère de sélection

Modèle : Nous considérons le modèle linéaire suivant :

$$Y_i = X_i\beta + \varepsilon_i, i = 1, \cdots, n \tag{1}$$

avec Y la variable réponse non censurée qui est dans notre cas la charge virale, X la matrice de design de taille $n \times p$ représentant les mutations et β de taille $p \times 1$ le vecteur de paramètres associés et ε suivant une $\mathcal{N}(0, \sigma^2)$ de façon indépendant et identiquement distribuées.

On note Z la variable représentant les données censurées à gauche. Elle est défini comme suit :

$$Z_i = \begin{cases} Y_i & if \quad Y_i > c_i \\ c & if \quad Y_i \leq c_i \end{cases} \tag{2}$$

avec c le seuil de censure, pour simplifier les calculs on considère $c_i = c$ (le seuil de censure est le même pour chaque individu).

La variable réponse étant la charge virale, nous faisons l'hypothèse que les données censurées sont gaussiennes $Y_i|X_i \sim \mathcal{N}(X_i\beta, \sigma^2)$. La censure est donc modélisée de la façon suivante : [8]

$$\mathbb{P}(Y_i|Y_i < c, X_i) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(u-X_i\beta)^2}{2\sigma^2}} \, \mathrm{d}u \tag{3}$$

Méthode d'estimation : Buckley et James ont proposé dans les années soixante-dix [1] un estimateur compensant la perte d'information due à la censure à droite. On note Z^* l'estimation de la réponse non censurée. L'estimateur de la réponse, Z^*, correspond à l'imputation de la réponse censurée par son espérance conditionnelle.

$$Z_i^* = \delta_i Y_i + (1-\delta_i)\,\mathbb{E}\left(Y_i|Y_i \leq c, X_i\right) \quad avec \quad \delta_i = \begin{cases} 1 & if \quad Y_i > c \\ 0 & if \quad Y_i \leq c \end{cases} \tag{4}$$

Ce qui est équivalant à :

$$Z_i^* = \delta_i Z_i + \beta X_i + \int_{Y_i - \beta X_i}^{\infty} \frac{y dF(y)}{1 - F(Y_i - \beta X_i} \tag{5}$$

Où $F()$ est la distribution des résidus ε_i. Le plus souvent, $F()$ est l'estimateur de Kaplan Meier ou la fonction de poids de Gehan. Dans notre cas, notre variable réponse correspond à la charge virale, les résidus sont donc gaussiens. L'espérance conditionnelle peut donc être écrite de la façon suivante.

$$\mathbb{E}\left(Y_i|Y_i \leq c, X_i\right) = \left(\int_{-\infty}^{c} \frac{e^{\frac{(u-X_i\beta)^2}{-2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \, \mathrm{d}u \right)^{-1} \left(\int_{-\infty}^{c} \frac{u \; e^{\frac{(u-X_i\beta)^2}{-2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \, \mathrm{d}u \right) \tag{6}$$

[7]

Pour estimer la réponse non censurée, nous utilisons les coefficients de régression β obtenus à partir d'une estimation LASSO issue du package `glmnet`. Les deux estimations sont effectuées successivement puis itérativement dans une boucle. Les conditions d'arrêts doivent être choisies de manière à garder la variance faible et ainsi limiter les variations d'estimation. Cette méthode sera notée `BJ-Lasso` dans la suite du document.

Critère de sélection : Dans une estimation Lasso, l'un des paramètres important est le paramètre de régularisation λ qui est un paramètre positif et à choisir. Pour cela, on créé une grille de valeurs possible pour λ assez fine et une estimation Lasso est faite pour chacune de ces valeurs. Il existe plusieurs critère de choix de modèles. Dans les méthodes de machine learning, on cherche le plus souvent à prédire, c'est à dire que l'on va chercher le modèle qui minimise l'erreur de prédiction défini comme suit : $\|Z^* - X\beta\|^2$. Cependant,

comme précisé au dessus, l'objectif premier est la sélection donc nous avons utilisés des critères tels que l'AIC ou le BIC défini comme suit :

$$AIC = \min_{\ell}\{2\#\{j, \hat{\boldsymbol{\beta}}_{j,\ell} = 0\} - 2\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2_{\ell})\}$$

$$BIC = \min_{\ell}\{log(n)\#\{j, \hat{\boldsymbol{\beta}}_{j,\ell} = 0\} - 2\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2_{\ell})\}$$

Les trois critères ont été testés.

3 Protocole de simulation

Nous souhaitons vérifier l'efficacité de l'algorithme en terme de sélection de variable pour de la petite et grande dimension. Nous avons donc fixé le nombre d'individu $N = 100$ et fait varier le nombre de variables $p \in \{50, 200\}$. La matrice X est simulée de façon à ce que la probabilité d'avoir une mutation soit de 0,25 et que sa matrice de variance-covariance Σ soit de la forme : $\Sigma_{ij} = 0.4^{|i-j|}$. Le cas sans corrélation a également été testé. Le nombre de coefficients non nuls varie entre 10% ou 20% du nombre total de variables. Pour finir, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ avec σ^2 choisi pour avoir un rapport signal/bruit égal à 1. Nous avons donc 4 scénarios différents, 2 pour la petite dimension et 2 pour la grande dimension avec la sparsité qui varie. Un autre élément que nous avons souhaité contrôler est le taux de censure. L'algorithme a été testé pour 20% ou 30% de taux de censure.

4 Résultats

Trois méthodes ont été comparées : `BJ_Lasso` que nous avons implémenté à l'aide de code déjà existant et disponible par les auteurs, le package `bujar` proposé par Wang & Wang [9]. Leur méthode repose sur de l'analyse de survie et utilise des méthodes de boosting. Nous l'avons juste adaptées pour les données censurées à gauche. Pour finir, la méthode des tests multiples avec et sans correction a également été testée.

Les résultats de la figure 1 sont données pour 100 simulations et dans le cas des données corrélées pour 20% de taux de censure.

5 Conclusion

En petite dimension (Scénario 1 et 2), les trois méthodes ont des taux de faux positifs ou faux négatifs entre 0 et 10% et sont du même ordre. Les méthodes `BJ-Lasso` et `Bujar` ont un écart inter-quartile plus faible que la méthode des tests multiples sur l'ensemble des simulations. Dans le cas de la grande dimension (Scénario 3 et 4), le nombre de faux

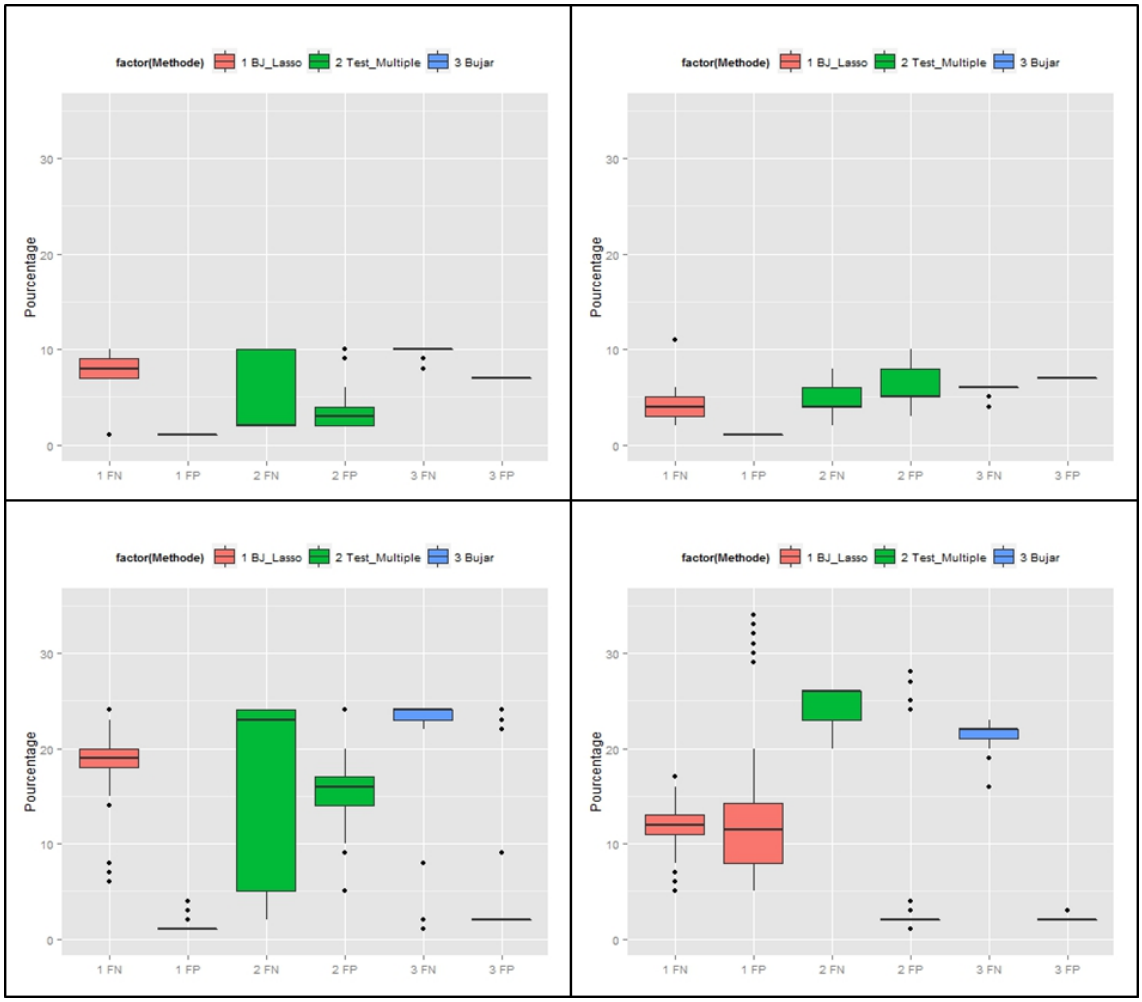

FIGURE 1 – Résultats pour les 4 scénarios avec 20% de taux de censure pour 100 simulations (Haut-Gauche : Scénario 1, Haut-Droite : Scénario 2, Bas-Gauche : Scénario 3, Bas-Droite : Scénario 4, FP = Faux Positifs, FN = Faux Négatifs)

positifs et faux négatifs ne sont plus du même ordre et augmente (entre 0 et 30%). Dans le scénario 3, c'est à dire 10% de sparsité, la méthode `BJ-Lasso` a son taux de faux négatifs qui augmente (20%) par rapport à la petite dimension alors que le taux de faux positifs reste inchangé(2%). En revanche, quand la sparsité est plus importante, on observe une forte augmentation des faux positifs. Pour les méthodes `bujar` et tests multiples, le nombre de faux négatifs augmente par rapport à la petite dimension et est généralement plus haut que la méthode `BJ-Lasso`. En revanche, le nombre de faux positifs reste bas.
`bujar` et les tests multiples ont tendance à trop supprimer de variables mais sélectionne des variables pertinentes. Alors que la méthode `BJ-Lasso` conserve plus de variables mais pas les plus pertinentes.

Références

[1] J Buckley and I James. Linear regression with censored data. *Biometrika*, 66, 1979.

[2] N Fouret, M Avalos, L Wittkop, R Thiebaut, and D Commenges. Prise en compte de la censure à gauche dans la modélisation de données de grande dimension. In *Journées de la Statistique*, 2014.

[3] S Gaiffas and A Guilloux. High-dimensional additive hazrds models and the lasso. *EJS*, 6, 2012.

[4] B.A Johnson. Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society*, 70, 2008.

[5] B.A Johnson. On lasso for censored data. *Electronic Journal of Statistics*, 3, 2009.

[6] B.A Johnson. Rank-based estimation in the l1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics*, 1, 2009.

[7] J Zhu S Wang, B Nan and D Beer. Doubly penalized buckley-james method for survival data with high-dimensional covariates. *Biometrics*, 64, 2008.

[8] R Thiebaut and al. Estimation of dynamical model parameters taking into account undectable marker values. *BMC Medical Research Methodology*, 1, 2006.

[9] Z Wang and C.Y Wang. Buckley-james boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology*, 9, 2010.

[10] L Dicker Y Li and S Zhao. The dantzig selector for censored linear regression models. *Stat Sin*, 24, 2014.

[11] D.Y Lin Z Jin and Z Ying. On least-squares regression with censored data. *Biometrika*, 93, 2006.

[12] M Zhou and G Li. Empirical likelihood analysis of buckley-james estimator. *Journal of multivariate analysis*, 99, 2008.