Évaluation de méthodes de sélection de gènes pour le pronostic du cancer

Rémy JARDILLIER[1,2], Laurent GUYON[1], Florent CHATELAIN[2]

[1]University Grenoble Alpes, CEA, INSERM, Biology of Cancer Infection UMR_S 1036
17 rue des Martyrs 38054 Grenoble cedex 9, France

[2]University Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Institute of Engineering University Grenoble Alpes
11 Rue des Mathématiques, 38400 Saint-Martin-d'Hères, France
remy.jardillier@cea.fr, laurent.guyon@cea.fr
florent.chatelain@gipsa-lab.grenoble-inp.fr

Résumé – Cet article propose d'étudier des méthodes d'inférence et de sélection sur des données de survie clinique pour le cancer. Le modèle de Cox permet de prendre en compte les données censurées à droite, et nous l'utilisons pour modéliser les données de survie et les données d'expression génétique. Des variantes du Lasso sont étudiées en terme de stabilité, de sélection et de prédiction. Les résultats montrent que ces méthodes ne sont pas stables et que les performances de sélection sont mauvaises, mais que la prédiction de la survie globale est bonne pour les patients qui ont un indice pronostique élevé.

Abstract – This paper presents an evaluation of penalization methods to select genes in survival analysis for cancer. The Cox semi-parametric model aims at taking into account censorship and is used to model survival and genetic data. In this context, Lasso-like methods are studied in terms of stability, selection and prediction. Results show that these methods are unstable and the selection performances are disapointing, but prediction of overall survival is good for patients with high prognostic index.

1 Introduction

L'étude des données de survie est un enjeu majeur dans la compréhension biologique et la médecine personnalisée. Les jeux de données que nous étudions sont composés de données d'expression génétique et de temps de survie de patients atteints d'un cancer du rein. Ces données ont la particularité d'être censurées à droite : l'évènement étudié (le décès ici) n'est pas observé pour tous les patients. Le modèle de Cox est un modèle semi-paramétrique linéaire généralisé qui permet de modéliser ces données en prenant en compte à la fois les données censurées et les co-variables génétiques. Les méthodes classiques de pénalisation de la vraisemblance de type "Lasso" peuvent ainsi être appliquées pour sélectionner les co-variables.

En effet, sélectionner les gènes ayant un impact sur la survie des patients permet, d'une part, une meilleure compréhension biologique des mécanismes intervenant dans le développement et l'agressivité des cancers. D'autre part, obtenir un modèle parcimonieux permet se séquencer quelques gènes seulement et ainsi de réduire les coûts dans le but de faire de la prédiction.

Les données génétiques sont soumises à la "malédiction de la dimension" avec un nombre de patients n'excédant pas quelques centaines pour plus de 20 000 gènes dans certaines études. Une évaluation des méthodes classiques de sélection dans le contexte est donc nécessaire. Ainsi, nous étudierons les méthodes Lasso, Elastic Net et Adaptive Elastic Net en terme de stabilité, de sélection et de prédiction.

2 Inférence pour les données de survie

2.1 Données de survie

Une donnée de survie est un couple (τ, δ) où τ correspond au temps de suivi et δ à l'occurrence de l'évènement étudié ou non. On notera $\delta = 1$ si l'évènement est observé, et 0 sinon. Dans ce dernier cas, on parle de donnée censurée à droite.

Pour fixer le cadre théorique de l'étude des données de survie, nous noterons :

- T le temps de survie global : intervalle de temps entre le diagnostic et le décès du patient.
- C le temps de censure : intervalle de temps entre le diagnostic et la perte de vue du patient. T et C sont supposés indépendants.
- $\tau = \min(T, C)$ la variable qui modélise le temps de suivi, τ_i étant la $i$eme observation associée,
- δ le statut : 1 si le décès est observé, i.e. $T \leq C$, 0 sinon,
- $\boldsymbol{X} = (X_1, ..., X_p)^T$ le vecteur d'expression génétique, où p est le nombre de gènes.

L'évènement étudié dans ce papier est le décès des patients atteints d'un cancer du rein. Le modèle de Cox introduit dans la partie suivante est largement utilisé pour lier les temps de survie à des covariables explicatives (le niveau d'expression des gènes dans cet article) en prenant en compte les censures [1].

2.2 Estimateurs pour les données censurées

Estimateur de Kaplan-Meier. Cet estimateur [2] permet de modéliser la fonction de survie $S(t) = P(T > t)$ de manière non-paramétrique en prenant en compte les données censurées. Il est défini comme

$$\hat{S}(t) = \prod_{i:\tau_i<t} \frac{n_i - d_i}{n_i}, \quad (1)$$

où, pour $i = 1, \ldots, n$, τ_i est le temps de suivi pour le $i$ème patient, n_i est le nombre de survivants (patients non décédés, et non censurés, i.e. non perdus de vue) avant cet instant τ_i, et d_i est le nombre de patients décédés à l'instant τ_i. Cet estimateur simple ne permet cependant pas de relier directement le temps de survie avec les covariables telles que les données d'expression génétiques.

Le modèle de Cox : estimation paramétrique. Le modèle de Cox (ou modèle à risque proportionnel) [3] lie les données d'expression génétiques aux données de survie à travers la fonction de hasard $h(t) = \lim_{h\to 0} \frac{P(t\leq T<t+h|T\geq t)}{h}$ pour tout $t > 0$. Cette fonction représente la probabilité de décès instantanée par unité de temps. Cox propose de modéliser cette fonction comme suit :

$$h(t; \boldsymbol{X}) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{X}), \quad (2)$$

où $t > 0$ est la variable temporelle, $h_0(\cdot)$ le risque de base commun à tous les patients, et $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ les coefficients du modèle à estimer. Si on observe un échantillon $\tau_1, \ldots, \tau_n$ des temps de suivi de n patients, la pseudo-vraisemblance de Cox pour estimer le vecteur $\boldsymbol{\beta}$ est définie commme

$$L(\boldsymbol{\beta}) = \prod_{i=1\ldots n\ :\delta_i=1} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{X}^i)}{\sum_{l\in R_i} \exp(\boldsymbol{\beta}^T \boldsymbol{X}^l)}, \quad (3)$$

où $R_i = \{k \in \{1, ..., n\} : t_k \geq \tau_i\}$ est l'ensemble des survivants à risque, i.e. non perdus de vue, au temps τ_i. Le modèle est dit "semi-paramétrique" car le risque de base $h_0(\cdot)$ n'apparaît pas dans cette pseudo-vraisemblance et n'a pas besoin d'être inférée afin d'estimer le vecteur $\boldsymbol{\beta}$. Néanmoins ce risque peut également être spécifié afin de caractériser entièrement la fonction de survie. Le modèle classique de Weibull-Cox consiste à supposer que $h_0(\cdot)$ est la fonction de hasard d'une loi de Weibull $\mathcal{W}(\lambda, k)$:

$$h_0(t) = \lambda^{-k} k t^{k-1}, \quad (4)$$

où $\lambda > 0$ et $k > 0$ sont respectivement le paramètre d'échelle et de forme de la loi de Weibull.

2.3 Sélection de variables

Trois variantes classiques d'inférence et de sélection de variables par pénalisation de la log pseudo-vraisemblance de Cox, $l(\boldsymbol{\beta}) \equiv \log L(\boldsymbol{\beta})$, sont étudiées dans ce papier. La méthode Lasso [4] où l'estimateur des coefficients $\hat{\boldsymbol{\beta}}$ est défini comme

$$\hat{\boldsymbol{\beta}}(\text{lasso}) = \underset{\boldsymbol{\beta}}{\text{argmax}}\ l(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_1, \quad (5)$$

la variante *Elastic Net* (EN) [5] :

$$\hat{\boldsymbol{\beta}}(\text{EN}) = \underset{\boldsymbol{\beta}}{\text{argmax}}\ l(\boldsymbol{\beta}) - \lambda(\alpha||\boldsymbol{\beta}||_1 + \tfrac{1-\alpha}{2}||\boldsymbol{\beta}||_2^2), \quad (6)$$

et la méthode *Adaptive Elastic Net* (AEN), procédure en deux étapes définie dans [6]. La pénalisation Lasso a tendance à sélectionner une seule variable aléatoirement parmi un groupe de variables corrélées, et n'est pas stable en grande dimension. La norme l_2 de la pénalisation Elastic Net permet de pallier à ces problèmes, et il a été montré empiriquement que les capacités de prédiction avec la pénalisation ridge (norme l_2 uniquement) sont meilleurs que le Lasso [5]. Les propriétés théoriques de sélection de l'Adaptive Elastic Net sont meilleures que l'Elastic Net en grande dimension, et les hypothèses sont plus souples.

3 Évaluation des méthodes Lasso

On définit dans ce travail des procédures afin d'évaluer empiriquement les variantes du Lasso présentées au paragraphe précédent à la fois en terme de sélection (stabilité, performance de sélection), et de prédiction.

Stabilité de la sélection. En notant $D = \{((\tau_i, \delta_i), \mathbf{X}_i), i \in \{1; ..., n\}\}$ l'ensemble des données de survie et d'expression génétique, et $\hat{G}$ l'ensemble des gènes sélectionnés, la méthode étudiée est dite instable [7] si et seulement si de faibles modifications dans D entraînent de grands changements dans $\hat{G}$. Les modifications de D ont été effectuées en remplaçant une partie des données par celles provenant d'autres patients. Ceci permet de conserver un échantillon i.i.d. de taille constante afin de sélectionner les gènes.

Performance de sélection. Les performances de sélection sont étudiées en fonction du taux de censure, en terme 1) de taux de vrais positifs (TPR) pour mesurer la puissance de la procédure et 2) de taux de fausses découvertes (FDR) parmi les détections pour mesurer les erreurs de type I. Afin de disposer d'une vérité terrain, des données synthétiques sont générées selon un modèle Weibull-Cox (voir pex [8] pour les détails de la méthode de simulation). Pour conserver des données réalistes, les niveaux d'expression génétiques $\boldsymbol{X}$ sont repris de jeux de données réels. De plus, les valeurs des coefficients β_i et les paramètres k, λ du risque de base correspondent aux valeurs estimées sur des jeux de données réelles.

Performance de prédiction. La prédiction est étudiée individuellement pour chaque patient à partir de l'indice pronostique

$$P = \boldsymbol{\beta}^T \boldsymbol{X}. \quad (7)$$

Un inconvénient du modèle de Cox est qu'il ne fournit pas de modèle direct entre le temps de survie T et cet indice pronostique. L'objectif ici est de pouvoir représenter et quantifier cette relation.

Proposition 3.1. *Sous l'hypothèse du modèle Weibull-Cox, le temps de survie T et l'indice pronostique $P \equiv \boldsymbol{\beta}^T \boldsymbol{X}$ sont reliés par l'expression :*

$$\ln T = -\frac{1}{k} P + \log \epsilon, \tag{8}$$

où ϵ suit la loi de Weibull $\mathcal{W}(\lambda, k)$ [9].

Démonstration. Si (8) est vérifiée, alors

$$S(t) = \Pr(T > t) = \Pr\left(e^{-\frac{1}{k}\boldsymbol{\beta}^T \boldsymbol{X}} \times \epsilon > t\right),$$
$$= 1 - F_\epsilon\left(te^{+\frac{1}{k}\boldsymbol{\beta}^T \boldsymbol{X}}\right) = \exp\left(-\lambda^{-k} t^k e^{\boldsymbol{\beta}^T \boldsymbol{X}}\right),$$

où F_ϵ est la FdR de la loi de Weibull $\mathcal{W}(\lambda, k)$. Par construction, la fonction de hasard est l'opposée de la dérivée logarithmique de la fonction de survie. Il vient donc que $h(t) = -(\ln S(t))' = \lambda^{-k} k t^{k-1} e^{\boldsymbol{\beta}^T \boldsymbol{X}}$. D'après (4), on reconnaît bien l'expression du modèle Weibull-Cox, ce qui démontre le résultat. □

La proposition précédente montre qu'il existe, sous l'hypothèse du modèle Weibull-Cox, une relation linéaire entre le logarithme des temps de survie et l'indice pronostique. La réprésentation des données de survie dans le plan $(\ln T, P)$ permet donc d'analyser l'influence des indices pronostiques sur les temps de survie.

Cependant les données sont en pratique largement censurées (taux de censure supérieur à 50%), ce qui a pour effet de considérablement biaiser les relations entre les indices pronostiques et les temps de survie (les grands temps de survie sont les plus censurés). Nous proposons ici d'analyser la fonction de répartition (FdR) déduite de l'estimateur de Kaplan-Meier (1) prise au temps des décès t_i, i.e. $\hat{F}_{KM}(t_i) \equiv 1 - \hat{S}(t_i)$, en fonction de la FdR empirique des indices pronostiques prise au point P_i, notée $\hat{F}_P(P_i)$. En effet si la relation (8) est vérifiée et que l'effet dû aux indices pronostiques est significatif par rapport au bruit du modèle Weibull, les points $(\hat{F}_{KM}(t_i), \hat{F}_P(P_i))$ de ce diagramme proba-proba devraient se concentrer autour de la droite $y = 1 - x$. De plus, l'information apportée par les données censurées est maintenant efficacement prises en compte à l'aide de l'estimateur de Kaplan-Meier.

4 Résultats

Les co-variables utilisées sont des niveaux d'expression de gènes dans des tumeurs du rein après ablation provenant de la base de données TCGA [1]. Deux jeux de données sont disponibles :

— les niveaux d'expression de gènes codant pour les protéines (ARNm) : $n = 464$ patients, $p = 20525$ gènes.
— les niveaux d'expression de gènes non-codant (miARN) : $n = 504$ patients, $p = 462$ gènes

On se concentre ici principalement sur le problème plus simple des données miARN, où le nombre de covariables reste similaire au nombre d'échantillons. Afin de sélectionner les gènes et estimer l'indice pronostique, les hyperparamètres des méthodes étudiées sont classiquement estimés par validation croisée.

1. The Cancer Genome Atlas, `http://cancergenome.nih.gov/`

La figure 1 met en évidence la grande instabilité pour ce problème des méthodes de sélection, qui sont pourtant classiquement employées dans ces applications [1]. Le nombre de gènes communs décroît en fonction du pourcentage de patients remplacés dans le jeu de données d'apprentissage. Elastic Net obtient les meilleures performances en termes de stabilité. Mais le nombre de gènes décroît rapidement, et seulement 20% de gènes en commun sont sélectionnés pour deux jeux de données indépendants pour les gènes non-codants (miARN). Les résultats ne sont pas représentés pour les gènes codants (ARNm), mais il n'apparaît presque aucun gène en commun dans ce cas de très grande dimension.

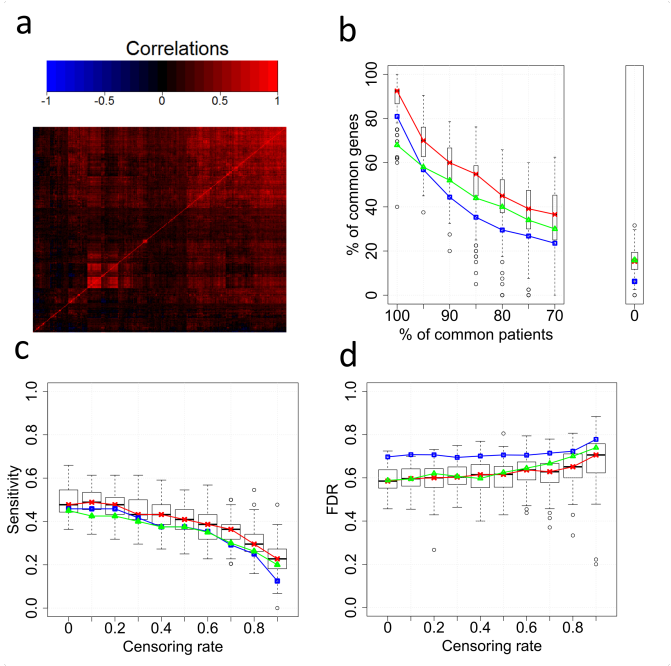

FIGURE 1 – Données miRNA (a) Heatmap des corrélations entre gènes (b) Nombre de gènes en commun entre les gènes sélectionnés et la liste de référence en fonction du nombre de patients en commun avec le jeu de données d'apprentissage. (b) TPR et (c) FDR en fonction du taux de censure pour des données synthétiques générées selon les données miARN, $N = 100$ simulations sont faites pour chaque taux de censure. Les boîtes sont tracées pour l'Elastic Net, et les médianes sont tracées pour l'Elastic Net (rouge), le Lasso (bleu) et l'Adaptive Elastic Net (vert)

Le TPR, Fig.1(b), décroît en fonction du taux de censure et cette décroissance s'accélère à partir d'un taux de censure de 60%. Le FDR, Fig. 1(c), reste quant à lui relativement stable à un niveau élevé supérieur à 50%. Ces figures montrent que les performances de sélection sont mauvaises. Il est difficile de déterminer quels gènes sont impliqués dans l'agressivité des cancers, ce qui est cohérent avec l'instabilité mise en évidence sur la figure 1(a). Une explication réside dans les fortes corrélations entre les covariables, ce qui rend difficile la sélection des vrais biomarqueurs. Ce phénomène est exacerbé en grande di-

mension, les performances étant encore dégradées dans le jeu de données ARNm lorsque la dimension augmente.

Les résultats pour évaluer maintenant la capacité prédictive de la survie du modèle de Cox pénalisé sont présentés sur la figure 2. L'indice pronostique P_i est estimé par validation croisée *leave-one-out*. Ainsi, pour chaque patient le vecteur des coefficients $\boldsymbol{\beta}$ est estimé sur les $n-1$ autres patients afin d'éviter d'utiliser les mêmes données pour estimer le modèle et le valider. Les figures 2(a) et 2(b) représentent les nuages de points

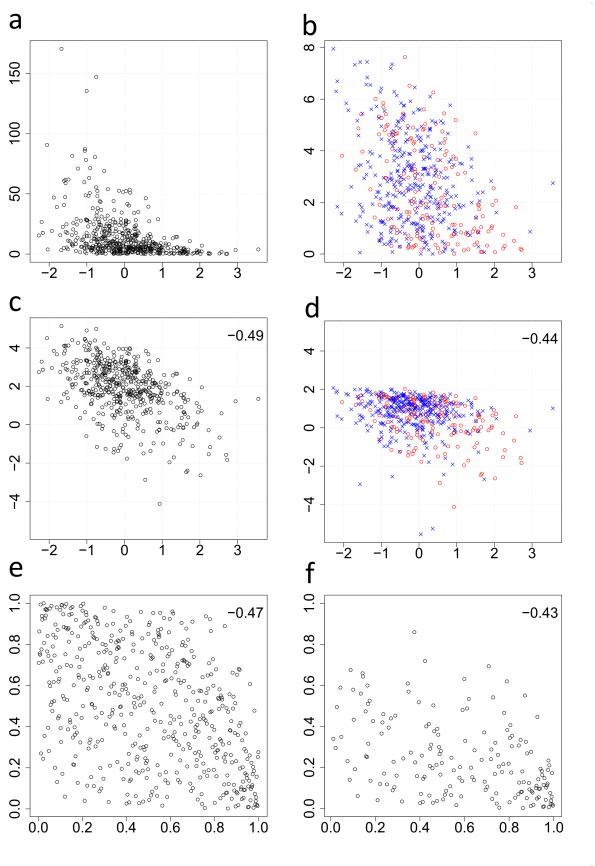

FIGURE 2 – Nuages de points pour des données synthétiques générées d'après les données miARN, $n = 504$, $p = 462$
(a) Temps de survie t_i vs indice pronostique P_i
(b) Temps de suivi τ_i vs P_i, bleu : censure, rouge : décès
(c) $\ln(t_i)$ vs indice pronostique P_i
(d) $\ln(\tau_i)$ vs P_i, bleu : censure, rouge : décès
(e) $\hat{F}_{KM}(t_i)$ vs $\hat{F}_P(P_i)$, voir paragraphe 3
(f) Idem que (e) en supprimant les données censurées

des temps de survie et de suivi, respectivement, en fonction des indices pronostiques. Par comparaison, la figure 2(c) illustre que le logarithme du temps de survie est une fonction linéaire de l'indice pronostique (proposition 3.1). Cette tendance linéaire devient moins évidente à partir des seuls temps de suivi montrés sur la figure 2(d). Néanmoins les diagrammes proba-proba 2(e) et 2(f) pour les temps de survie et les temps de suivi non censurés, respectivement, mettent en évidence cette tendance linéaire. Il apparaît ainsi que les grands indices pronostiques sont corrélés à de faibles probabilités de survie. De plus un coefficient de corrélation empirique $\hat{r} = -0.43$ est obtenu pour la figure 2(f). Afin de mesurer la significativité de cette corrélation, un test de permutation est finalement effectué : 10^8 permutations des indices pronostiques sont effectuées afin de simuler selon l'hypothèse nulle où les indices pronostiques et les temps de survie/suivi ne sont pas corrélés. Ceci permet d'estimer la FdR empirique des coefficients de corrélation et d'estimer la probabilité d'obtenir une corrélation aussi extrême sous l'hypothèse nulle. Il en ressort une p-valeur inférieure à 10^{-8}, ce qui confirme la capacité prédictive de l'indice P estimé.

5 Conclusion

Le problème de sélection de biomarqueurs génétiques s'avère en pratique instable et mal posé, notamment en raison des fortes corrélations entre les expressions des gènes. Il apparaît cependant possible, sur des données réalistes, d'estimer un indice qui permette d'améliorer la prédiction de la survie, tout particulièrement pour les patients qui ont un indice pronostique élevé.

Références

[1] Jardillier, Rémy and Chatelain, Florent and Guyon, Laurent. *Bioinformatics Methods to Select Prognostic Biomarker Genes from Large Scale Datasets : A Review*. Biotechnol J., 13(12), 2018.

[2] Kaplan, E.L. and Meier, Paul. *Nonparametric Estimation from Incomplete Observations*. American statistical Association, 53(282) :457-481, 1958.

[3] Cox, David R. *Regression Models and Life-Tables*. Journal of the Royal Statistical Society. Series B : Statistical Methodology, 34(2) :187-220 1972.

[4] Tibshirani, Robert. *The lasso method for variable selection in the cox model*. Statistics in Medicine, 16 :385-395, 1997.

[5] Zou, Hui and Hastie, Trevor. *Regularization and variable selection via the elastic-net*. Journal of the Royal Statistical Society, 67(2) :301-320, 2005.

[6] Zou, Hui and Zhang, Hao Helen. *On the adaptive elastic-net with a diverging number of parameters*. Annals of Statistics, 37(4) :1733-1751, 2009.

[7] Breiman, Leo *Heuristics of instability and stabilization*. Annals of Statistics, 24(6) :2350-2383, 1996.

[8] Bender, Ralf and Augustin, Thomas and Blettner, Maria *Generating survival times to simulate Cox proportional hazards models*. Statistics in Medicine, 24 :1713-1723, 2005.

[9] Van Houwelingen, Hans C. *Validation, calibration, revision and combination of prognostic survival models* Statistics in Medicine, 19 :3401-3415, 2000.