Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation.

Audrey Lavenu[1], Juliette Murris[2*], Alexis Mareau[3], Timothé Rouzé[4], Magalie Fromont[5], Valérie Gares[6] & Sandrine Katsahian[7]

[1] *CIC1414 Inserm, IRMAR, Université de Rennes 1. audrey.lavenu@univ-rennes1.fr,*
[2,3,4,7] *CIC1418 Inserm, HeKA-Inria, Université de Paris. juliette.murris-ext@aphp.fr, alexis.mareau5@gmail.com, timothe.rouze1@gmail.com, sandrine.katsahian@aphp.fr,*
[5] *IRMAR, Université de Rennes 2. magalie.fromont@univ-rennes2.fr,*
[6] *IRMAR, INSA, Rennes, France. valerie.gares@insa-rennes.fr,*
* *Subvention de l'ANRT et Pierre Fabre (CIFRE 2020/1701).*

Résumé. Avec l'augmentation du nombre de données sur les patients dans les domaines de l'imagerie médicale ou encore de la génomique, les méthodes d'analyses classiques sont souvent inadéquates dans les cas où il y a moins d'observations que de variables. L'objectif de notre travail est d'étudier différents critères de performance et leur estimation de la méthode Cox Boost pour analyser des données de survie en grande dimension sur petits échantillons. Nous nous intéressons à la prédiction et la discrimination des variables pronostiques, mais aussi à la "tunabilité" du modèle Cox Boost (gain par optimisation des hyperparamètres). Nous simulons les temps de survie avec une loi exponentielle et les temps de censure avec une loi uniforme. Pour fixer le taux de censure à un taux prédéfini, nous montrons comment calculer le paramètre de la distribution de censure. Avec un schéma de simulation faisant varier : la taille d'effet des covariables, la taille de l'échantillon, et le taux de variables actives, nous comparons différents critères de performance de la méthode (C de Harrell et mesure d'importance de variable) estimés par validation croisée à 2 et 5 blocs, avec trois méthodes de choix des hyperparamètres. Nous montrons la difficulté d'optimiser les hyperparamètres pour de petits échantillons, et les lacunes des mesures d'importance des variables à détecter les variables simulées actives, même quand la performance du modèle en termes de prédiction est correcte.
Mots-clés. Méthodes d'apprentissage supervisé, Survie, Grande dimension.

Abstract. With an increasing amount of patient data in the fields of medical imaging and genomics, conventional analysis methods are often inadequate for modeling cases with fewer samples than features. The objective of the present work is to study different performance criteria of the Cox Boost method for analysing high-dimensional survival data on small sample sizes. Here we focus on prediction and discrimination of prognostic features, as well as the tunability of the model (potential score increase by optimization of hyperparameters). We simulate survival times with exponential distribution, and censoring times with uniform distribution. To fix the censoring rate at a predefined value, we show how to derive the parameter of the censoring distribution. The simulation scheme modifies the effect size of the covariates, the sample size, and the rate of active features. We compare

the different performance criteria (Harrell's C and variable importance measure) by 2 and 5-fold cross-validation, with three ways of selecting the hyperparameters. We show the complexity of optimising hyperparameters for small sample size data. We also illustrate that the feature importance measure does not always allow detection of active simulated features, although the predictive performance of the model is correct.
Keywords. Supervised machine learning, Survival, High-dimensional.

1 Introduction

L'utilisation des mégadonnées dans la recherche et la pratique en santé publique nécessite de nouvelles compétences pour gérer et analyser ces données. La particularité des données de survie est principalement la présence de réponses éventuellement censurées. Sur la base d'une revue de la littérature des méthodes statistiques couramment utilisées et des techniques d'apprentissage automatique développées pour l'analyse de survie, une taxonomie détaillée des méthodes existantes a été proposée (Wang, 2019).

Dans les études pronostiques, nous avons deux objectifs : la prédiction (c'est-à-dire la précision de l'estimation du risque) et la détection des facteurs de risque (c'est-à-dire l'interprétation des résultats en fonction de l'importance des covariables). Cette dernière est nécessaire pour développer un meilleur diagnostic et des stratégies de traitement optimales (Pittman, 2004). Ces deux objectifs sont rarement étudiés simultanément.

Tous les algorithmes d'apprentissage impliquent un plus ou moins grand nombre d'hyperparamètres (ou "paramètres de réglage"). La quantité de gain de performance qui peut être obtenue en optimisant les hyperparamètres par rapport aux valeurs par défaut définit la "tunabilité" de ces algorithmes ou modèles. Il a été montré que la méthode d'optimisation des hyperparamètres et le moment où elle se fait peut engendrer des biais sur les petits échantillons, par exemple un fort taux de bien classés d'une variable binaire alors que les covariables étaient simulées sans effet (Vabalas, 2019).

L'objectif de notre travail est d'étudier les performances des méthodes pour analyser des données de survie de grande dimension, donc souvent sur petits échantillons avec beaucoup de variables. Nous étudierons ici les performances de la méthode Cox Boost (Binder, 2008) en termes de prédiction et de discrimination des variables pronostiques, mais aussi la tunabilité en fonction des tailles d'échantillon et du taux de covariables informatives.

2 Critère de performance en analyse de survie et critère d'importance des variables

En analyse de survie, nous notons $T_1, \ldots, T_n$ les temps auxquels surviennent l'évènement (ou plus précisément les durées d'apparition de l'évènement) pour n individus. Ces variables aléatoires sont supposées positives, indépendantes et de même densité f, telle que f est continue, bornée et strictement positive sur $\mathbb{R}^+$. La spécificité des analyses de survie est que nous ne pouvons pas observer $(T_1, ..., T_n)$ mais $(T_1^S, \delta_1), (T_2^S, \delta_2), \ldots, (T_n^S, \delta_n)$, où

$$\text{pour tout } 1 \leq i \leq n, \quad T_i^S = \min(T_i, C_i), \quad \delta_i = \mathbf{1}_{\{T_i \leq C_i\}},$$

et C_i est une variable aléatoire positive qui représente le temps de censure. Soit $X_i = (X_i^1, \ldots, X_i^p)$ un vecteur de p covariables.

Risque. Soit g : $\mathbb{R}^d \to \mathbb{R}$ une règle de décision. La performance d'une règle de décision est définie par le risque ou l'erreur de généralisation $\mathcal{R}(g)$. L'objectif est de construire une règle de décision qui soit aussi performante que la règle optimale $g^* \in \text{argmin}_g \mathcal{R}(g)$.
Le **C de Harrell** est le plus souvent utilisé en analyse de données de survie pour mesurer le pouvoir discriminant et la capacité prédictive des modèles. Le risque considéré ici est :

$$\mathcal{R}(g) = 1 - \mathbb{P}[M_i^g < M_j^g | T_i > T_j, T_j < \min(C_i, C_j)],$$

où M_i^g représente la probabilité d'avoir eu l'évènement d'intérêt à la fin de l'étude pour l'individu i prédite par la règle g. $\mathbb{P}[M_i^g < M_j^g | T_i > T_j, T_j < \min(C_i, C_j)]$ est estimée par le **C de Harrell** qui est une mesure de concordance :

$$\text{CIndex}^n(g) = \frac{\sum_{i=1}^n \sum_{j=1, i \neq j}^n \mathbf{1}_{\left\{T_i^S > T_j^S\right\}} . \mathbf{1}_{\left\{M_j^g > M_i^g\right\}} . \delta_j}{\sum_{j=1, i \neq j}^n \mathbf{1}_{\left\{T_i^S > T_j^S\right\}} \delta_j}.$$

Estimation du risque. Le risque estimé sur les données servant à l'apprentissage conduit à une sous-estimation du risque. Plusieurs solutions existent pour construire des estimations sans biais du risque telles que le partage de l'échantillon en deux (apprentissage/test) ou validation croisée holdout, où le risque est alors estimé sur les données test. Une autre méthode utilisée est la validation croisée par blocs. Sur de petits échantillons, la première est difficile à envisager, et les méthodes de validation croisée à K blocs sont alors privilégiées. On considère une partition $\{\mathcal{I}_1, \ldots \mathcal{I}_K\}$ de $\{1, \ldots, n\}$. Pour chaque $k = 1, \ldots, K$, la réglè g est construite avec l'échantillon d'aprentissage $\left\{(X_i, T_i^S, \delta_i), i \in \{1, \ldots, n\} \setminus \mathcal{I}_k\right\}$, on la note $\widehat{g}^{n,k}$. Le C de Harrell $\text{CIindex}_{n,k}(\widehat{g}^{n,k})$ est calculé sur l'échantillon test $\left\{(X_i, T_i^S, \delta_i), i \in \mathcal{I}_k\right\}$. On considère ensuite le risque moyen sur les K blocs $\mathcal{R}_n(g) = 1 - \frac{1}{K} \sum_{k=1}^K \text{CIindex}_{n,k}(\widehat{g}^{n,k})$. Plus le nombre de blocs augmente, plus la taille de l'échantillon d'apprentissage est grande, plus celle de l'échantillon test est petite, et dans le cas du "leave-one-out" avec un seul individu par bloc, le C de Harell ne

peut pas être calculé.

Choix des hyperparamètres. Les règles de décision (méthodes) peuvent dépendre d'hyperparamètres $h \in \mathcal{H}$. Soit $\mathcal{G} = \{g_h, h \in \mathcal{H}\}$, la famille des règles g dépendant de l'hyperparamètre h. L'objectif est de sélectionner la règle g_h qui minimise le risque estimé par validation croisée $\mathcal{R}_n(g_h)$. Nous avons choisi deux approches d'optimisation des hyperparamètres. La méthode HalvingGridSearch (Jamieson, 2016) consiste en l'évaluation à partir d'un sous-ensemble des données plutôt que sur toutes les données. A chaque étape du processus, les combinaisons d'hyperparamètres les plus performantes sont évaluées et la quantité de données utilisées augmente en se concentrant sur les meilleures combinaisons. On dit qu'il s'agit d'halving successifs. La méthode Optuna (Akiba, 2019) est un système d'optimisation automatique par recherche dynamique des hyperparamètres. A chaque itération, cet algorithme permet de tirer au hasard une combinaison d'hyperparamètres en se basant sur les performances de la meilleure combinaison des itérations précédentes, et stoppe les combinaisons les moins prometteuses.

Critère d'importance des variables. Pour mesurer l'importance de la variable j, on considère la partition $\{\mathcal{I}_1, \dots \mathcal{I}_K\}$. Pour chaque $k = 1, \dots, K$, $\mathcal{I}_k^j$ est construit à partir de l'échantillon $\mathcal{I}_k$ où les valeurs de la variable j sont permutées aléatoirement. La règle $\widehat{g}^{n,k}$ est définie précédemment. On calcule le C de Harrell sur l'échantillon dont les valeurs de la variable j sont perturbées $\left\{(X_i, T_i^S, \delta_i), i \in \mathcal{I}_k^j\right\}$ que l'on note $\text{CIndex}_{n,k}^j(\widehat{g}^{n,k})$. La mesure d'importance de la variable j est donnée par $\text{Imp}(j) = \frac{1}{K}\sum_{k=1}^{K}(\text{CIndex}_{n,k}(\widehat{g}^{n,k}) - \text{CIndex}_{n,k}^j(\widehat{g}^{n,k}))$.

3 Méthode CoxBoost et hyperparamètres associés

L'ajustement des modèles de survie par la vraisemblance partielle de Cox est la méthode la plus largement utilisée pour la régression des risques proportionnels linéaires. Le modèle de Cox exprime la fonction de risque instantané de décès λ en fonction du temps t et des covariables $X_1, \cdots, X_n$. On a alors : $\lambda\left(t, X_1, \cdots, X_n\right) = \lambda_0\left(t\right) exp\left(\Sigma_{i=1}^n \beta_i X_i\right)$.
Le boosting s'adapte aussi aux modèles de régression non linéaires et aux modèles de Cox en construisant une famille d'estimateurs qui sont ensuite agrégés. Les 3 hyperparamètres que nous optimiserons pour cette méthode sont : le taux d'apprentissage (learning rate, la valeur par défaut est à 0,1), le nombre d'estimateurs (100 par défaut), et le taux d'abandon (dropout rate, 0 par défaut).

4 Simulation

Génération de données. Rappelons nos notations : nous considérons un échantillon de n individus, pour chaque inidividu i, soit X_i vecteur de p covariables, T_i temps de

survie, C_i temps de censure. Nous observons : $T_i^S = \min(T_i, C_i)$ et $\delta_i = \mathbf{1}_{T_i \leq C_i}$. Soit $X_i \sim \mathcal{N}_p(\mu, \Sigma)$ où $\mu = (a, \ldots, a)$ et Σ ayant une structure de corrélation autoregressive telle que $\text{corr}(X_i^k, X_i^{k'}) = \rho^{|k-k'|}$ ($\rho = 0.7$) avec $1 \leq i, j \leq p$. Basé sur un modèle de Cox avec composantes linéaires, nous associons les coefficients $\beta = (\beta_1, \ldots, \beta_p)$ aux p covariables. p' coefficients sont fixés à b (indicés $k_1, \ldots, k_{p'}$) et $p - p'$ sont fixés à 0. Le taux appelé $\text{sr} = \frac{p'}{p}$ (sparse rate) est alors le taux de variables actives. Prenons T un temps d'évènement suivant une distribution exponentielle $\mathcal{E}(\lambda_x)$ où $\lambda_x = \alpha \exp\left(\sum_{j=1}^{p'} b x_i^{k_j}\right)$. Si nous fixons m_0 le temps de survie médian de base souhaité et m_1 celui pour une augmentation de une unité d'une covariable active, nous pouvons alors calculer $b = \ln\left(\frac{m_0}{m_1}\right)$ et $\alpha = \frac{\ln(2)}{m_0}$. Nous supposons que la censure C suit une distribution uniforme sur $[0, \theta]$, $C \sim \mathcal{U}([0, \theta])$ et est indépendante de T conditionnellement à X. Soit $\tau = \mathbb{P}(T < C) = \mathbb{P}(\delta = 1)$ le taux de censure, θ est solution de l'équation adaptée de la méthode de Wan (2017) :

$$\gamma(\theta) = \int_0^{+\infty} \frac{1}{\theta u} \exp(-u\theta) \frac{1}{u\sqrt{2\pi}\sqrt{(\beta)^T \Sigma \beta}} \exp\left(-\frac{(\ln(u) - \ln(\alpha) - p'ba)^2}{2(\beta)^T \Sigma \beta}\right) du - \tau = 0.$$

Schémas de simulation. Pour chaque scénario, nous simulerons $M = 100$ jeux de données avec un taux de censure de $\tau = 50\%$. Nous choisirons $a = 0$ et $m_0 = 10$. Nous ferons varier : le gain de médiane pour l'augmentation de une unité : $m_1 - m_0 \in \{0.5, 1\}$; la taille de l'échantillon : $n \in \{50, 100, 150, 200, 500\}$; et le taux de variables actives : $\text{sr} \in \{0, 0.25, 0.5\}$.

Evaluation de la méthode. Pour les M jeux de données, nous estimerons le C de Harrell de la méthode CoxBoost par la validation croisée à 2 et 5 blocs (par la moyenne des C de Harrell dans chaque bloc). Nous évaluerons trois méthodes de choix des hyperparamètres (sans optimisation donc avec les valeurs par défaut, et les deux méthodes précédemment citées : HalvingGridSearch et Optuna).

Sur des jeux de données simulés, nous connaissons le vrai statut actif ou passif des p covariables, nous notons Y la variable aléatoire associée. Soit $Y^j = 1$ si la variable X^j est active (et a un effet sur la durée de vie) et $Y^j = 0$ sinon. Pour $s \in [0, 1]$ on définit :

$$\widehat{x}(s) = \frac{\sum_{j=1}^p \mathbf{1}_{\{\text{Imp}(j) > s\}}(1 - Y^j)}{\sum_{j=1}^p (1 - Y^j)} \text{ et } \widehat{y}(s) = \frac{\sum_{j=1}^p \mathbf{1}_{\{\text{Imp}(j) \geq s\}} Y^j}{\sum_{j=1}^p Y^j}, \quad \forall s \in [0, 1].$$

Nous appelons AUC^{Imp} l'aire sous la courbe paramétrée définie par $(\widehat{x}(s), \widehat{y}(s)), \forall s \in [0, 1]$. Si AUC^{Imp} est proche de 1, nous conclurons que la méthode de prédiction aura plutôt utilisé les variables actives sur ce jeu de données.

En notant $j_1, \ldots, j_p$ les indices des covariables tels que $\text{Imp}_{j_1} \geq \ldots \geq \text{Imp}_{j_p}$ on calculera en outre $\sum_{k=1}^{p'} \mathbf{1}_{\left\{Y^{j_k} = 1\right\}}$ le nombre de variables actives dans les p' variables dont la mesure d'importance est la plus grande, avec $p' = p.sr$ le nombre de variables simulées actives.

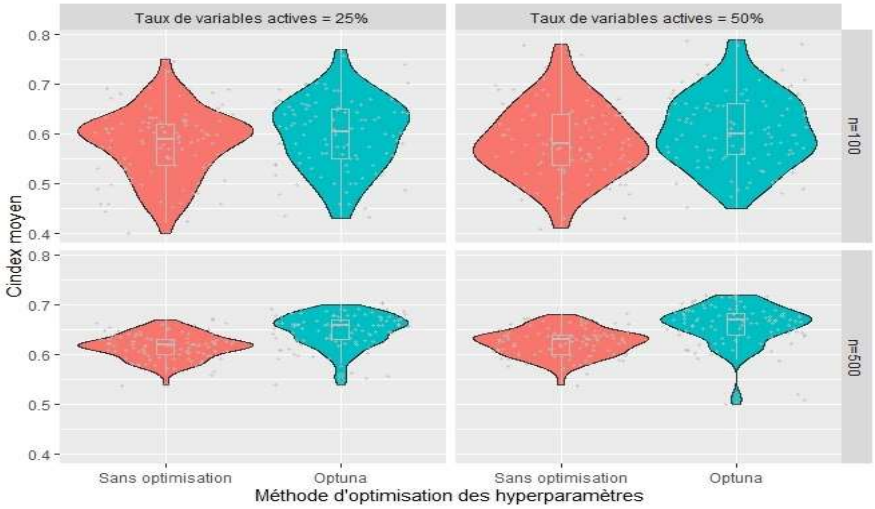

FIGURE 1 – Distribution des CIndex moyens sur les 5 blocs de la validation croisée pour les $M = 100$ jeux de données simulées, $n \in \{100, 500\}$, $m_1 - m_0 = 1$, sr $\in \{0.25, 0.5\}$.

Résultats. La figure 1 montre un exemple de simulations où $n \in \{100, 500\}$, $m_1 - m_0 = 1$, sr $\in \{0.25, 0.5\}$. Les validations croisées sont réalisées en 5 blocs et le choix des hyperparamètres se fait soit sans optimisation soit avec Optuna. L'optimisation des hyperparamètres est efficace pour $n = 500$ avec un C de Harrell estimé au dessus de 0.65.
Nous montrerons la difficulté d'optimiser les hyperparamètres pour de petits échantillons, et que l'importance des variables dans le modèle utilisé ne permet pas toujours de détecter les variables simulées actives, même quand la performance du modèle en termes de prédiction est correcte.

Bibliographie

Akiba T et al. (2019). Optuna : A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining.*
Binder H and Schumacher M (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics.* 9 :14.
Jamieson K and Talwalkar A (2016). Non-stochastic best arm identification and hyperparameter optimization. *Artificial intelligence and statistics.* 240-248.
Pittman J, et al. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc Natl Acad Sci USA.* 101(22) :8431-6.
Vabalas A, Gowen E, Poliakoff E, Casson A. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE.* 14(11) :e0224365
Wan F. (2016). Simulating survival data with predefined censoring rates for proportional hazards models. *Stat Med.* 36(5) :838-854.
Wang P, Li Y, Reddy C.K. (2017). Machine Learning for Survival Analysis : A Survey. *ACM Computing Surveys* 51(6)