Réseaux de neurones pour l'analyse de survie en grande dimension

Mathilde Sautreuil [1], Sarah Lemler [1] and Paul-Henry Cournède [1]

[1] *Laboratoire MICS, CentraleSupélec, Université Paris-Saclay, 9 rue Joliot Curie, 91190 Gif-sur-Yvette, France; prenom.nom@centralesupelec.fr.*

Résumé. L'analyse de survie est l'étude du temps écoulé jusqu'à la survenue d'un événement d'intérêt qui peut correspondre par exemple au décès ou à la rémission dans une étude médicale. Dans ce contexte, l'objectif de notre travail est de prédire la durée de survie de patients à partir de données génomiques et cliniques. L'approche par réseaux de neurones pour l'analyse de survie n'est pas récente, mais seulement des données d'entrées de faible dimension ont été considérées par le passé. Cependant, depuis l'arrivée du séquençage à haut-débit le nombre de covariables potentiellement intéressantes pour les modèles de prévision est devenu très important. Le cadre statistique a donc changé. Nous présenterons et testerons donc quelques approches récentes des réseaux de neurones, adaptées à l'analyse de survie en grande dimension.

Mots-clés. Analyse de survie, Réseau de neurones, Grande dimension, Cancer, Transcriptomique.

Abstract. Survival analysis consists in studying the elapsed time until an event of interest, which can be for example the death or recovery of a patient in the medical studies. In this context, the objective of this work is to explore the potential of neural networks in survival analysis from clinical and genomic data. If the approach is not recent, past approaches only considered in small numbers of input data. But with the emergence of high-throughput sequencing data, the number of covariates of interest became very important, with new statistical issues to consider. We present and test a few recent approaches for neural-network based survival analysis adapted to high-dimensional inputs.

Keywords. Survival analysis, Neural networks, High-dimension, Cancer, Transcriptomics.

1 Introduction

Le but de notre travail est de comparer des méthodes pour prédire la durée de survie des patients à partir de données cliniques et génomiques. Le modèle de Cox (Cox, 1972) est le modèle de référence dans le domaine de l'analyse de survie. Celui-ci permet de relier la durée de survie d'un individu pour une maladie donnée aux variables explicatives. Ce modèle permet également de prendre en compte les données censurées à droite.

Les données de survie sont dites censurées à droite quand le temps observé est plus petit que le temps de survie. Cela se produit quand l'étude se termine avant que l'événement d'intérêt (décès, rémission) ait lieu mais cela peut être également dû à d'autres raisons (décès du patient pour une autre maladie que celle étudiée,...). De plus, avec l'arrivée des techniques de séquençage à haut-débit, les données transcriptomiques sont de plus en plus utilisées comme covariables dans l'analyse de survie. L'ajout de ces covariables amène de nouvelles problématiques car on se trouve dans le cadre de la grande dimension, c'est-à-dire lorsque le nombre de covariables est plus important que le nombre d'individus. Dans ce cadre, la procédure d'estimation classique du modèle de Cox ne peut plus être appliquée directement. Des méthodes ont alors été développées pour réduire le nombre de covariables, notamment le Lasso (Tibshirani *et al.*, 1996) et autres variantes, en minimisant un critère d'estimation pénalisé qui rend nuls certains coefficients. Ces méthodes ont d'abord été implémentées dans le cadre linéaire et elles ont été ensuite adaptées au domaine de l'analyse de survie en utilisant la vraisemblance partielle de Cox (Tibshirani, 1997; Zou and Hastie, 2005). La vraisemblance partielle de Cox (Cox, 1975) correspond à une partie de la vraisemblance totale qui est connue pour permettre d'obtenir un bon estimateur du paramètre de régression du modèle de Cox en la maximisant. Cependant, des problèmes de stabilité ont été montrés (Fan *et al.*, 2009). Pour résoudre cela, d'autres méthodes comme SIS *(Sure Independence Screening)* et ses dérivées ont été proposées (Fan and Song, 2010). SIS peut être résumée en deux étapes. La première étape consiste à réduire le nombre de covariables en gardant seulement celles dont le score calculé sur le modèle de Cox est supérieur à un seuil. La seconde étape consiste à exécuter une procédure Lasso afin de sélectionner les variables importantes parmi les variables restantes. Toutes ces méthodes sont applicables avec le modèle de Cox, mais elles ne résolvent pas complètement les problèmes de stabilité. Le *deep learning* est un domaine de recherche très populaire depuis quelques années et le domaine biomédical s'y intéresse de plus en plus. Des réseaux de neurones ont été développés en analyse de survie avec deux stratégies différentes. La première est un réseau de neurones basé sur la log-vraisemblance partielle de Cox comme celui développé par Faraggi and Simon en 1995 ou Ching *et al.* en 2018 et dernièrement par Katzman *et al.* en 2018. Faraggi a été le premier à développer cette stratégie en analyse de survie et Ching *et al.* l'a ensuite appliquée dans le cadre de la grande dimension. La seconde stratégie consiste à développer un réseau de neurones basé sur un modèle de survie à temps discret (Brown *et al.*, 1997; Biganzoli *et al.*, 1998, Rodrigo and Tsokos, 2017; Gensheimer and Narasimhan, 2018). Comme décrit précédemment, le modèle de Cox est très populaire en analyse de survie car il prend en compte les données censurées. Mais avec ce modèle, il est plus difficile d'estimer la fonction de risque directement. La plupart des études utilise la vraisemblance partielle de Cox et se concentre donc sur l'estimation des coefficients de régression sans prendre en compte le risque de base, même si certaines procédures en deux étapes existent (Lemler, 2016). L'avantage des modèles à temps discret est qu'ils permettent d'estimer directement la fonction de risque. On se concentrera donc sur l'étude d'un réseau de neurones développé par Biganzoli *et al.*

en 1998 basé sur la seconde stratégie et qui n'a jamais été étudié en grande dimension. On comparera cette procédure à d'autres méthodes : estimation classique des paramètres du modèle de Cox par maximisation de la log-vraisemblance pénalisée et la procédure Cox-nnet qui est un réseau de neurones basé sur la vraisemblance partielle de Cox.

2 Modèles

2.1 Modèle de Cox

Le modèle de Cox est un modèle couramment utilisé dans l'analyse de survie. Ce modèle est défini à partir de la fonction de risque λ qui est fonction du temps conditionnellement aux covariables $X_{i.} = (X_{i1}, \ldots, X_{ip})^T$:

$$\lambda(t|X_{i.}) = \alpha_0(t) \exp(\beta X_{i.}), \tag{1}$$

avec $\alpha_0(t)$ la fonction du risque de base et β est le vecteur des coefficients de régression. La fonction de base $\alpha_0(t)$ dépend du temps et la seconde partie de (1) dépend seulement des covariables. Cela peut être utile quand on s'intéresse aux rapports de la fonction de risque pour deux individus car le rapport ne dépendera pas de $\alpha_0(t)$. En grande dimension, la procédure d'estimation classique qui consiste à maximiser la vraisemblance ne fonctionne plus. Comme mentionné dans la Section 1, la procédure Lasso est une méthode classique en grande dimension afin de réduire le nombre de covariables. Celle-ci va mettre les variables non informatives à zéro. Cette procédure adaptée au modèle de Cox utilise la vraisemblance partielle de Cox, qui est une partie de la log-vraisemblance totale ne dépendant pas du paramètre $\alpha_0(t)$ définie de la façon suivante :

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} (\beta X_{i.}) - \sum_{i=1}^{n} \delta_i \log \left(\sum_{l \in R_i} \exp(\beta X_l) \right),$$

avec R_i les individus à risque au temps t_i et δ_i l'indicateur de censure.

2.2 Réseau de neurones

Les réseaux de neurones considérés ici sont de type *multi-layer perceptron*. Un réseau est constitué de plusieurs couches avec au moins une couche d'entrée, une couche cachée et une couche de sortie. Chaque couche est constituée de plusieurs neurones et chaque neurone joue le rôle d'une régression non-linéaire entre ses entrées et sa variable de sortie. Les coefficients pondérant les entrées dans ces régressions sont appelés poids, et la fonction non-linéaire de transformation de cette combinaison pour donner la sortie est appelée fonction d'activation. La sortie d'un neurone va être l'entrée d'un neurone en aval et c'est cette association qui va constituer le réseau. De plus, les neurones de la même couche n'ont pas de connexions entre eux. Ils possèdent seulement des connexions avec les neurones des couches précédentes et suivantes.

2.2.1 Cox-nnet

Cox-nnet est un réseau de neurones développé par Ching *et al.* en 2018, basé sur le modèle de Cox. Avant celui-ci, un réseau de neurones a été développé (Faraggi and Simon, 1995) basé sur le modèle à risques proportionnels mais il n'avait pas été appliqué sur des données en grande dimension. Le principe de Cox-nnet est que sa couche de sortie correspond à une régression de Cox. Dans le modèle Cox-nnet, $X_{i.}$ de l'équation (1) est remplacé par la sortie de la couche cachée qui est $G(WX_{i.} + b)^T\beta$ avec W la matrice des poids, b est le terme de biais pour chaque nœud caché et G est la fonction d'activation. Dans ce réseau, la fonction d'activation *tanh* est utilisé et ils ont ajouté un terme de régularisation *ridge* dans la log-vraisemblance partielle de Cox. Ils emploient aussi un *dropout* pour réduire le sur-apprentissage.

2.2.2 Réseau de neurones basé sur un modèle à temps discret

Comme décrit précédemment, plusieurs stratégies existent pour modéliser les données de survie. Biganozli *et al.* a développé en 1998 un réseau de neurones basé sur un modèle à temps discret. On introduit L intervalles de temps $A_l =]t_{l-1}, t_l]$ auxquels appartiennent les temps de survie. Dans ce contexte, la fonction de risque discrète est définie comme la probabilité conditionnelle de survie :

$$h_{il} = P(t_i \in A_l | T_i > t_{l-1}),$$

avec T_i le temps de survie de l'individu i. Biganozli *et al.* introduit l'indicateur de censure d_{il} égale à 1 dans l'intervalle A_l contenant l'événement d'intérêt pour les individus non censurés et égale à 0 sinon. La vraisemblance totale s'écrit :

$$L = \prod_{i=1}^{n} \prod_{l=1}^{l_i} h_{il}^{d_{il}} (1 - h_{il})^{(1-d_{il})}. \quad (2)$$

Pour le réseau de neurones, la fonction d'erreur de *cross-entropy* correspond à l'opposé de la log-vraisemblance. La fonction logistique est utilisée comme fonction d'activation pour la couche cachée et la couche de sortie. Un terme de régularisation *ridge* est ajouté à la fonction de *cross-entropy*.

3 Résultats pour le cancer du rein à cellules claires

Comme décrit dans la Section 1, deux stratégies différentes ont été proposées pour les réseaux de neurones en analyse de survie. La première est un réseau de neurones basé sur la log-vraisemblance partielle de Cox. Faraggi et Simon l'ont développé en premier en 1995 pour un nombre de covariables faible en analyse de survie et Ching *et al.* a étendu ce travail au cadre de grande dimension. La seconde stratégie est d'utiliser un modèle à temps discret, ce qui correspond au modèle développé par Biganozli *et al.* en 1998 en

petite dimension. Nous l'avons étendu à la grande dimension, notamment en travaillant sur la régularisation du réseau de neurones. On comparera donc ce travail à celui de Ching *et al.*, ainsi qu'au modèle de Cox pénalisé, qui reste la référence en analyse de survie.

Pour illustrer cette comparaison, nous utilisons la base de données TCGA (The Cancer Genome Atlas) (https://tcga-data.nci.nih.gov/tcga/). Cette base de données est constituée de différents types de cancers. Nous évaluerons les résultats des réseaux de neurones sur celui du cancer du rein à cellule claire (dont le code d'accès dans la base de données est KIRC). Nous avons récupéré ces données à l'aide du package R RTCGA, qui nous permet d'obtenir facilement les données cliniques et d'expression de gènes. Pour ce jeu de données, nous avons accès à 17 741 gènes pour 533 patients. Parmi ces individus, 67% sont censurés.

Pour évaluer les modèles présentés, nous avons utilisé comme métrique l'indice de concordance (C-index) (Gerds, 2013). Cette métrique mesure si la prédiction du modèle correspond bien aux rangs des données de survie. En effet, si le temps d'événement du patient i est plus petit que celui du patient j, un bon modèle prédira une plus grande probabilité de survie pour le sujet j. Le C-index est donc calculé de la façon suivante :

$$C(t) = E_{ij}\left[1\left\{S(t, X_i) < S(t, X_j)\right\} | T_i < T_j\right],$$

où $S(t, X_i)$ est la probabilité que l'individu i survive au temps t. Celle-ci est calculée pour des données de survie censurées et va prendre une valeur entre 0 et 1. Si cette métrique est égale à 0.5, cela signifie que le modèle est équivalent à un processus aléatoire.

	NNsurv	Cox-nnet	Cox
C-index (train)	0.92	0.85	0.74
C-index (test)	0.64	0.61	0.63

Table 1: Résultats des différentes méthodes sur le jeu données KIRC.

Le réseau de neurones développé par Biganozli *et al.* (NNsurv dans la table 1) et auquel nous avons appliqué des données de grande dimension donne des résultats très encourageants. En effet, il a de meilleurs résultats en grande dimension que Cox-nnet et Cox. On obtient sur le jeu de donnés test, correspondant à 20% des données, un C-index égal à 0.64 contre respectivement 0.61 et 0.63 pour Cox-nnet et Cox. De plus, il a l'avantage d'estimer directement la fonction de risque contrairement aux deux autres méthodes estimant seulement les coefficients de régression et nécessitant donc une deuxième procédure pour obtenir la fonction de risque. Enfin, nous avons vu que la procédure d'estimation pénalisée du modèle de Cox n'est pas stable en sélection, ce qui peut ajouter un avantage à l'utilisation de réseaux de neurones. Ce réseau semble donc intéressant pour étudier la prédiction de survie des patients en grande dimension.

Bibliographie

Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17(10), 1169–1186.

Brown, S. F., Branford, A. J., and Moran, W. (1997). On the use of artificial neural networks for the analysis of survival data. *IEEE Transactions on Neural Networks*, 8(5), 1071–1077.

Ching, T., Zhu, X., and Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4), e1006076.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2), 269–276.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6), 3567–3604.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. *Institute of Mathematical Statistics*.

Faraggi, D. and Simon, R. (1995). A neural network model for survival data. *Statistics in Medicine*, 14(1), 73–82.

Gensheimer, M. F. and Narasimhan, B. (2018). A Scalable Discrete-Time Survival Model for Neural Networks. *arXiv:1805.00917 [cs, stat]*.

Gerds, T. A., Kattan, M. W., Schumacher, M., and Yu, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13), 2173–2184.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 24.

Lemler, S. (2016). Oracle inequalities for the Lasso in the high-dimensional Aalen multiplicative intensity model. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(2), 981–1008.

Rodrigo, H. and Tsokos, C. P. (2017). Artificial Neural Network Model for Predicting Lung Cancer Survival. *Journal of Data Analysis and Information Processing*, 05, 33.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, 16(4), 385–395.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2), 301–320.