Construction et estimation des capacités d'un score pronostique : intérêt de la pénalisation de LASSO et de l'estimateur bootstrap 0.632+ appliqués aux courbes ROC dépendantes du temps

Marie-Cécile Fournier [†,‡,1] & Florence Gillaizeau [†,‡,2] & Awena Le Fur [‡,3] & Jacques Dantal [‡,4] & Yohann Foucher [†,‡,5]

[†] *EA4275 Biostatistique, pharmacoépidémiologie et mesures subjectives en santé, 1 rue Gaston Veil, 44035 NANTES*

[‡] *ITUN Institut de Transplantation Urologie Néphrologie INSERM UMR1064, 1 Place Alexis Ricordeau, 44093 Nantes cedex 1*

E-mail : [1]*marie-cecile.fournier@univ-nantes.fr ;* [2]*florence.gillaizeau@univ-nantes.fr ;* [3]*awelefur@hotmail.fr ;* [4]*jacques.dantal@chu-nantes.fr ;* [5]*yohann.foucher@univ-nantes.fr*

Résumé. Les données de masse ou "big data" sont caractérisées par un petit nombre d'individus et un grand nombre de variables. Face au risque de surajustement, les méthodes de régression et de sélection classiques ne peuvent plus être utilisées. Une solution est la pénalisation de LASSO qui sélectionne les variables sur la vraisemblance pénalisée. En présence de données censurées, elle peut être appliquée au modèle de Cox pour construire un score pronostique. L'estimation des capacités prédictives du score obtenu est réalisée avec des courbes ROC dépendantes du temps et corrigée par un algorithme de rééchantillonnage : le bootstrap 0.632+. Cette méthode, publiée auparavant dans le champ des biopuces et validée par des études de simulations, est présentée ici pour une application alternative sur données cliniques réelles dans le domaine de la transplantation rénale.

Mots-clés. modèle de Cox, pénalisation de LASSO, bootstrap 0.632+, courbes ROC dépendantes du temps, surajustement

Abstract. Big data are characterized by a few number of individuals and a lot of variables. Given the risk of overfitting, classical methods of regression and selection are no longer adequate. One solution is the LASSO penalization which selects variables using the penalized likelihood. In the presence of censored data, it can be applied to the Cox model to obtain a prognostic score. The estimation of predictive capacities of the score are determined from time dependent ROC curves and corrected with a bootstrap algorithm : the bootstrap 0.632+. This method, published previously in the context of microarray data and validated by simulations, is described here with an application to real clinical data in kidney transplantation study.

Keywords. Cox model, LASSO penalization, 0.632+ bootstrap estimator, time dependent ROC curves, overfitting

1 Introduction

La prédiction du devenir d'un individu est d'un intérêt primordial pour de nombreuses disciplines, tout particulièrement dans le domaine de la santé. Dans les études de cohorte, de nombreuses variables explicatives peuvent être renseignées au moment de l'inclusion d'un individu, ce dernier étant suivi jusqu'à l'occurrence d'un certain évènement. Selon le temps de suivi, cet évènement peut ne pas être observé (censure à droite). Les courbes ROC dépendantes du temps développées par Heagerty (2000) permettent d'estimer les capacités pronostiques d'un score en tenant compte de ce phénomène de censure. L'adaptation de l'estimateur bootstrap 0.632+ aux courbes ROC dépendantes du temps présentée ici a été motivée initialement pour la construction de signatures à partir de la technique des puces à ADN. Celles-ci sont caractérisées par un nombre de variables explicatives supérieur au nombre d'individus, situation de surajustement qui se retrouve dans d'autres contextes lorsque peu d'évènements sont observés. La validation d'un score se fait habituellement sur un échantillon (dit de validation) indépendant de celui qui a servi à sa construction (échantillon d'apprentissage). Cela induit une perte de puissance et des intervalles de confiance larges, notamment en présence de surajustement. De plus, pour certains utilisateurs peu scrupuleux, ce découpage peut être répété jusqu'à l'obtention des résultats souhaités. L'intérêt d'un échantillon indépendant de validation perd alors son pouvoir de preuve. Ce travail propose l'adaptation de l'estimateur par bootstrap 0.632+ aux courbes ROC dépendantes du temps publiée par Foucher et Danger (2012). Notre objectif est de montrer l'intérêt de cette approche dans des études d'épidémiologie clinique classiques pour des chercheurs qui souhaitent proposer un score de pronostic.

2 Modèle de Cox avec pénalisation de LASSO

Soient X le vecteur des p covariables potentiellement prédictives de l'évènement étudié $X = (X_1, ..., X_P)$ et $x_j = (x_{j1}, ..., x_{jP})$ les observations associées à l'individu j ($j = 1...N$). Soient T_j le temps d'évènement de l'individu j et C_j le temps de censure, on en déduit l'indicatrice de l'évènement δ_j ($\delta_j = 1$ si $T_j \leq C_j, 0$ sinon). Posons Y_j le temps de dernière information pour le sujet j tel que : $Y_j = min(T_j, C_j)$. Alors, le modèle de Cox s'écrit :

$$h(t|x_j) = h_0(t)exp(\beta x_j') \tag{1}$$

où $h_0(t)$ est la fonction de risque de base non spécifiée et $\beta = (\beta_1, ..., \beta_P)$ est le vecteur des coefficients de régression. L'estimation classique est basée sur le maximum de vraisemblance partielle $l(\beta)$.

L'approche LASSO (Tibshirani, 1996) permet de restreindre le nombre de paramètres dans un modèle en pénalisant la vraisemblance partielle du modèle de Cox telle que :

$$\widehat{\beta} = argmax \left\{ l(\beta) - \lambda \sum_{p=1}^{P} |\beta_p| \right\} \tag{2}$$

où λ est le paramètre de pénalisation positif ou nul.

3 Courbes ROC dépendantes du temps

Nous souhaitons à présent évaluer les capacités prédictives de notre index pronostique estimé précédemment $\hat{\eta} = \hat{\beta}x$. Les valeurs importantes du score η sont en faveur de l'évènement. Soit τ le temps de pronostic, alors le patient est défini à haut risque de subir l'évènement avant τ si $\eta > c$, c étant le seuil recherché. D'après la définition de Heagerty et al. (2000), le taux de faux négatifs (TFN) et le taux de faux positifs (TFP) pour un pronostic au temps τ sont respectivement $TFN_\tau(c) = P(\eta \leq c | T \leq \tau)$ et $TFP_\tau(c) = P(\eta > c | T > \tau)$. Les capacités pronostiques de $\hat{\eta}$ sont résumées par la courbe ROC dépendante du temps, notée ROCt : $1 - TFN_\tau(c)$ est tracé en fonction de $TFP_\tau(c)$ pour tous les seuils c. L'aire sous cette courbe évalue les capacités pronostiques : il s'agit de la probabilité qu'une personne subissant l'évènement avant τ ait une valeur du score supérieure à un individu libre de l'évènement au temps τ.

Ces taux d'erreur peuvent être estimés de manière non paramétrique :

$$\widehat{TFN}_\tau(c) = \{\hat{G}(c) - \hat{S}(-\infty, \tau) + \hat{S}(c, \tau)\}/\{1 - \hat{S}(-\infty, \tau)\} \tag{3}$$

$$\widehat{TFP}_\tau(c) = \hat{S}(c, \tau)/\hat{S}(-\infty, \tau) \tag{4}$$

où $\hat{G}()$ est une estimation de la fonction de répartition de η, donnée par $N^{-1}\sum_j \mathbb{1}(\hat{\eta}_j < c)$, et $S(c, \tau) = P(\eta > c, T > \tau)$ est la survie bivariée de η et T. Cette probabilité jointe peut être estimée par la méthode proposée par Akritas (1994) :

$$\hat{S}(c, \tau) = N^{-1} \sum_{j=1}^{N} \hat{S}(\tau | \hat{\eta} = \hat{\eta}_j) \mathbb{1}(\hat{\eta}_j > c) \tag{5}$$

où $\hat{S}(\tau | \hat{\eta} = \hat{\eta}_j)$ est un estimateur de la fonction de survie conditionnelle basée sur un noyau des plus proches voisins. Soit K le noyau des plus proches voisins, le principe étant de choisir les patients éligibles, c'est à dire les patients tels que la valeur de leur marqueur soit proche de la valeur $\hat{\eta}_j$ d'intérêt : $K_\pi(\hat{\eta}_j, \hat{\eta}_l) = \mathbb{1}(-\pi < \hat{G}(\hat{\eta}_j) - \hat{G}(\hat{\eta}_l) < \pi)$. 2π représente la proportion de voisins inclus. A partir de cette définition, l'estimation de la survie conditionnelle s'écrit comme un estimateur de Kaplan Meier pondéré :

$$\hat{S}(\tau | \hat{\eta} = \hat{\eta}_j) = \prod_{s \leq \tau} \left\{ 1 - \frac{\sum_l K_\pi(\hat{\eta}_j, \hat{\eta}_l) \mathbb{1}(y_l = s) \delta_l}{\sum_l K_\pi(\hat{\eta}_j, \hat{\eta}_l) \mathbb{1}(y_l \geq s)} \right\} \tag{6}$$

Nous obtenons alors les taux d'erreur de prédiction pour notre score en incluant le phénomène de censure. Mais, en estimant ces taux sur les individus ayant servis au développement du score, nous surestimons ses capacités prédictives. Pour une estimation correcte, nous proposons l'utilisation de la méthode de rééchantillonnage par bootstrap 0.632+, amélioration des méthodes par bootstrap cross-validation (BCV) ou par bootstrap 0.632.

3.1 Bootstrap Cross Validation

Soit B ($b = 1, ..., B$) échantillons de bootstrap de taille N avec remise. Pour les B échantillons, l'estimation des coefficients $\widehat{\beta}_b$ est réalisée par maximisation de la vraisemblance pénalisée. Pour les B échantillons des patients non inclus dans les échantillons de bootstrap, les quantités $\widehat{FNR}_\tau^b(c)$ et $\widehat{FPR}_\tau^b(c)$ sont calculées. Ainsi, on obtient par Bootstrap Cross Validation (BCV) les estimations suivantes :

$$\widehat{FNR}_\tau^{BCV}(c) = B^{-1} \sum_{b=1}^{B} \widehat{FNR}_\tau^b(c) \tag{7}$$

$$\widehat{FPR}_\tau^{BCV}(c) = B^{-1} \sum_{b=1}^{B} \widehat{FPR}_\tau^b(c) \tag{8}$$

La courbe ROCt par BCV est définie par $1-\widehat{FNR}_\tau^{BCV}(c)$ en fonction de $\widehat{FPR}_\tau^{BCV}(c)$ pour toutes les valeurs de c. Cependant, Efron (1983) et Foucher et Danger (2012) ont montré que cette courbe peut-être associée à une sous-estimation des capacités pronostiques. En effet, si N est suffisamment grand ($N \geq 40$), la probabilité qu'un individu soit tiré au sort dans l'échantillon de bootstrap est : $1 - (1 - 1/N)^N \approx 0.632$. Cette proportion est composée d'individus pouvant être répliqués alors que la proportion $(1 - 1/N)^N \approx 0.368$ est composée d'individus indépendants entre eux. Cette situation pénalise les capacités pronostiques d'un score calculé à partir d'individus plus homogènes mais validé sur un échantillon toujours plus hétérogène.

3.2 Bootstrap 0.632

Afin de corriger ce biais, les mêmes auteurs ont proposé l'estimateur 0.632 :

$$\begin{aligned} \widehat{FNR}_\tau^{0.632}(c) &= 0.368\overline{FNR}_\tau(c) + 0.632\widehat{FNR}_\tau^{BCV}(c) \\ \widehat{FPR}_\tau^{0.632}(c) &= 0.368\overline{FPR}_\tau(c) + 0.632\widehat{FPR}_\tau^{BCV}(c) \end{aligned} \tag{9}$$

où $\overline{FNR}_\tau(c)$ et $\overline{FPR}_\tau(c)$ sont les taux apparents, calculés de la même manière que dans (7) et (8) mais en utilisant les individus inclus dans les échantillons de bootstrap. La courbe ROCt bootstrap 0.632 est définie par $1 - \widehat{FNR}_\tau^{0.632}(c)$ en fonction de $\widehat{FPR}_\tau^{0.632}(c)$ pour toutes les valeurs de c. Cette approche peut cependant surestimer la capacité pronostique en présence d'un fort sur-ajustement des données où les capacités apparentes de prédiction peuvent être proches de la perfection. Efron et Tibshirani (1997) ont amélioré la correction avec l'estimateur 0.632+.

3.3 Bootstrap 0.632+

Cet estimateur a été adapté aux données censurées par Foucher et Danger (2012) :

$$\widehat{FNR}_{\tau}^{0.632+}(c) = \{1 - \psi(\widehat{r}_{FNR,\tau}(c))\}\overline{FNR}_{\tau}(c) + \psi(\widehat{r}_{FNR,\tau}(c))\widehat{FNR}_{\tau}^{BCV}(c) \quad (10)$$

$$\widehat{FPR}_{\tau}^{0.632+}(c) = \{1 - \psi(\widehat{r}_{FPR,\tau}(c))\}\overline{FPR}_{\tau}(c) + \psi(\widehat{r}_{FPR,\tau}(c))\widehat{FPR}_{\tau}^{BCV}(c) \quad (11)$$

avec

- $\psi(x) = 0.632/(1 - 0.368x)$
- $\widehat{r}_{FNR,\tau}(c) = \left\{\widehat{FNR}_{\tau}^{BCV}(c) - \overline{FNR}_{\tau}(c)\right\}/\left\{\widehat{\gamma}_{FNR,\tau}(c) - \overline{FNR}_{\tau}(c)\right\}$
- $\widehat{r}_{FPR,\tau}(c) = \left\{\widehat{FPR}_{\tau}^{BCV}(c) - \overline{FPR}_{\tau}(c)\right\}/\left\{\widehat{\gamma}_{FPR,\tau}(c) - \overline{FPR}_{\tau}(c)\right\}$

$\widehat{r}$ sont les taux de surajustement et $\widehat{\gamma}$ est le taux de non-information associé aux taux de faux négatifs et de faux positifs. Il peut être estimé en utilisant toutes les données et en considérant l'indépendance entre η et T : $\widehat{\gamma}_{FNR,\tau}(c) = 1 - \widehat{\gamma}_{FPR,\tau}(c)$

La courbe ROCt bootstrap 0.632+ est définie par $1 - \widehat{FNR}_{\tau}^{0.632+}(c)$ en fonction de $\widehat{FPR}_{\tau}^{0.632+}(c)$ pour toutes les valeurs de c.

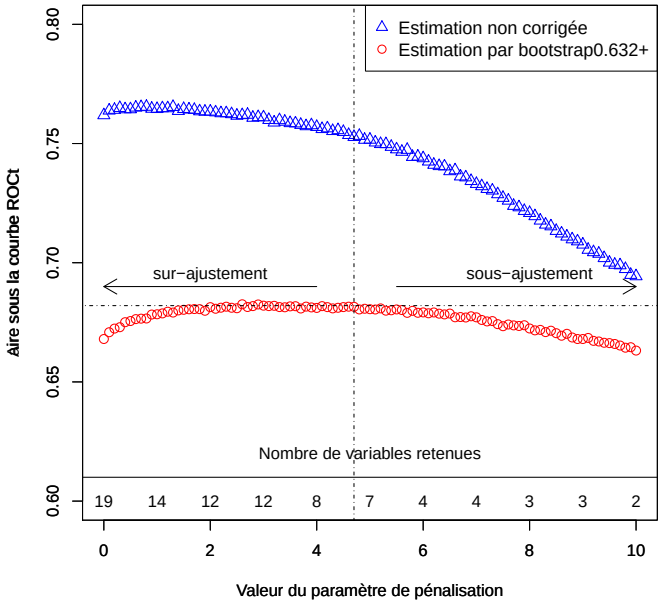
3.4 Estimation du paramètre de pénalisation

Le paramètre de pénalisation λ est classiquement estimé par validation croisée, ce qui doit être réalisé pour chaque échantillon bootstrap ($\hat{\lambda}_b$), tout le modèle devant être théoriquement réestimé à chaque itération. Cependant, la convergence de la validation croisée est discutable et cette réestimation est couteuse en temps de calcul. Foucher et Danger (2012) ont par ailleurs montré que l'estimation a priori de λ sur tout l'échantillon n'avait pas d'impact sur un éventuel surajustement. C'est aussi le choix fait par Schumacher et al. (2007). Nous appliquons ici la même simplification et proposons de fixer ce paramètre dès la première étape.

3.5 Application

Notre application a été réalisée à partir de la cohorte prospective DIVAT (Données Informatisées et VAlidée en Transplantation www.divat.fr) de Nantes incluant des patients transplantés rénaux entre 2000 et 2010. L'objectif est de proposer un score pronostique de la survenue d'un diabète post transplantation. On souhaite déterminer ce score au moment de la greffe afin de pouvoir envisager des modifications de la prise en charge des patients (modalités de suivi, prescriptions médicamenteuses, etc.). L'échantillon disponible contient 444 patients mais seulement 58 évènements sont observés. La variable d'étude est le délai entre la greffe et la survenue du diabète. Nous censurons les patients à la date du

retour en dialyse (perte du greffon), du décès du patient avec son greffon fonctionnel ou à la date des dernières nouvelles. Pour déterminer le score pronostique, nous utilisons le package ROC632 que nous avons implémenté en R. Nous obtenons alors un score comprenant 7 facteurs pronostiques. L'apport de la correction de l'estimation des capacités prédictives est illustrée dans le graphique suivant.



Bibliographie

[1] Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring. The annals of statistics. 1994 Sep;22(3):1299-1327.
[2] Efron B. Estimating the error rate of a prediction rule : improvement on cross validation. Journal of the american statistical association. 1983 Jun;78(382)n:316-331.
[3] Efron B, Tibshirani R. Improvements on cross validation : the 632+ bootstrap method. Journal of the american statistical association. 1997;92(438):548-560.
[4] Foucher Y, Danger R. Time dependant ROC curves for the estimation of true prognostic capacity of microarray. Statistical applications in genetics and molecular biology. 2012;11(6).
[5] Heagerty PJ, Lumley T, Pepe MS. Time dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000 Jun;56(2):337-344.
[6] Schumacher M, Binder, Gerds T. Assessment of survival prediction models based on microarray data. Bioinformatics, 23:1768-1774, 2007.
[7] Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the royal statistical society series B. 1996.