Estimation adaptative à noyau du risque de base dans le modèle de Cox en grande dimension

Sarah Lemler[1], Agathe Guilloux[2] & Marie-Luce Taupin[3]

[1] *Laboratoire MICS, École CentraleSupélec, France*
e-mail : sarah.lemler@centralesupelec.fr
[2] *LSTA, Université Pierre et Marie Curie, Paris 6, France*
CMAP, École Polytechnique, CNRS UMR 7641, 91128 Palaiseau, France
e-mail : agathe.guilloux@upmc.fr
[3] *LAMME, UMR CNRS 8071- USC INRA, Université d'Évry Val d'Essonne, France*
e-mail : marie-luce.taupin@genopole.cnrs.fr

Résumé. Nous proposons un nouvel estimateur à noyau pour le risque de base dans le modèle de Cox en grande dimension, pour lequel nous obtenons une vitesse de convergence non-asymptotique. Pour construire notre estimateur, nous estimons dans un premier temps le paramètre de régression du modèle de Cox à l'aide de la procédure LASSO. Puis, nous injectons cet estimateur dans l'estimateur à noyau usuel du risque de base, obtenu en lissant l'estimateur de Breslow du risque de base cumulé. Nous proposons et étudions une procédure adaptative pour sélectionner la fenêtre, dans l'esprit de Goldenshluger et Lepski (2011). Nous établissons une inégalité oracle non-asymptotique pour l'estimateur final, qui montre que la vitesse de convergence est ralentie lorsque le nombre de covariables augmente.

Mots-clés. Fonction de risque conditionnel, Modèle semi-paramétrique, Processus de comptage, Estimation à noyau, Méthode de Goldenshluger et Lepski, Inégalité oracle non-asymptotique, Analyse de survie

Abstract. We propose a novel kernel estimator of the baseline function in a general high-dimensional Cox model, for which we derive a non-asymptotic rate of convergence. To construct our estimator, we first estimate the regression parameter in the Cox model via a LASSO procedure. We then plug this estimator into the classical kernel estimator of the baseline function, obtained by smoothing the so-called Breslow estimator of the cumulative baseline function. We propose and study an adaptive procedure for selecting the bandwidth, in the spirit of Goldenshluger and Lepski (2011). We state a non-asymptotic oracle inequality for the final estimator, which leads to a reduction in the rate of convergence when the dimension of the covariates grows.

Keywords. Conditional hazard rate function, Semi-parametric model, Counting process, Kernel estimation, Goldenshluger and Lepski method, Non-asymptotic oracle inequality, Survival analysis

1 Contexte et cadre de travail

Le modèle de Cox, introduit par Cox (1972), est un modèle de régression souvent utilisé en analyse de survie pour relier la durée de survie T_i d'un individu i à un vecteur de covariables $\boldsymbol{Z_i} = (Z_{i,1}, \dots, Z_{i,p})^T$ à l'aide de la fonction de risque λ_0 :

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \exp(\boldsymbol{\beta_0^T Z_i}),$$

où $\boldsymbol{\beta_0} = (\beta_{0_1}, \dots, \beta_{0_p})^T \in \mathbb{R}^p$ le vecteur des paramètres de régression et α_0 le risque de base sont les deux paramètres inconnus du modèle. Pour chaque individu i ($i = 1 \dots, n$), les observations peuvent être décrites à partir d'un processus de comptage N_i, d'un processus prévisible Y_i à valeurs dans $[0, 1]$, qui complète l'information sur les observations, et du vecteur de covariables $\boldsymbol{Z_i} = (Z_{i,1}, \dots, Z_{i,p})^T$. Ce cadre général des processus de comptage inclut les données censurées, les processus de Poisson marqués et les processus de Markov.

De nombreuses études se sont intéressées à l'estimation du paramètre de régression $\boldsymbol{\beta_0}$ en petite dimension (lorsque le nombre de covariables p est inférieur à la taille de l'échantillon n), mais aussi en grande dimension ($p \gg n$). Il existe un critère d'estimation introduit par Cox (1972), appelé la log-vraisemblance partielle de Cox, connu pour permettre d'estimer $\boldsymbol{\beta_0}$ sans avoir à connaître α_0. Elle est définie par :

$$l_n^*(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \ln \frac{\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}}}{S_n(t, \boldsymbol{\beta})} \mathrm{d}N_i(t), \quad \text{où } S_n(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}} Y_i(t). \tag{1}$$

Lorsque le nombre de covariables p est suffisamment petit par rapport à n, on obtient un bon estimateur en maximisant la log-vraisemblance partielle de Cox. En grande dimension, Tibshirani (1997) a proposé une procédure Lasso pour l'estimation du paramètre de régression dans le modèle de Cox : il s'agit de maximiser l'opposé de la log-vraisemblance de Cox pénalisée par la norme ℓ_1 du vecteur $\boldsymbol{\beta}$ sur lequel on minimise. Plusieurs résultats non-asymptotiques ont été démontrés récemment pour l'estimateur Lasso obtenu.

Il existe moins de résultats concernant l'estimation du risque de base α_0. Ramlau-Hansen (1983) a introduit un estimateur à noyau du risque de base, défini pour un noyau $K : \mathbb{R} \to \mathbb{R}$ tel que $\int_{\mathbb{R}} K(x)\mathrm{d}x$ et une fenêtre $h > 0$, par

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{nh} \sum_{i=1}^{n} \int_0^\tau K\Big(\frac{t-u}{h}\Big) \frac{I\{\bar{Y}(u) > 0\}}{S_n(u, \hat{\boldsymbol{\beta}})} \mathrm{d}N_i(u), \quad \text{avec } \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{2}$$

L'estimateur à noyau $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$ dépend d'un estimateur $\hat{\boldsymbol{\beta}}$ du paramètre de régression $\boldsymbol{\beta_0}$. Le choix h de la fenêtre est important et Grégoire (1993) a proposé une procédure de validation croisée pour sélectionner la fenêtre. Cependant, aucun résultat non-asymptotique n'a été établi pour l'estimateur final obtenu.

Dans ce travail, nous considérons le cas où le nombre de covariables p est possiblement supérieur à la taille de l'échantillon n. Dans ce contexte, nous proposons une

procédure d'estimation de α_0 adaptée à la grande dimension. Nous établissons une borne non-asymptotique pour l'estimateur obtenu et nous mesurons l'influence de la grande dimension sur l'estimation du risque de base.

2 Procédure d'estimation

Nous considérons une procédure d'estimation en deux étapes. La première étape consiste à estimer le paramètre de régression $\boldsymbol{\beta_0} \in \mathbb{R}^p$, à l'aide d'une procédure adaptée à la grande dimension, puis dans un deuxième temps nous estimons le risque de base α_0.

Estimation préliminaire de $\boldsymbol{\beta_0}$.

Nous estimons le paramètre de régression $\boldsymbol{\beta_0}$ à l'aide d'une procédure LASSO appliquée à la vraisemblance partielle de Cox (1). L'estimateur LASSO $\hat{\boldsymbol{\beta}}$ est alors défini par

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{B}(0,R)}{\arg\min} \{-l_n^*(\boldsymbol{\beta}) + \Gamma_n |\boldsymbol{\beta}|_1\}, \tag{3}$$

où $\Gamma_n > 0$ est un paramètre de régularisation et $\mathcal{B}(0,R) = \{b \in \mathbb{R}^p : |b|_1 \leq R\}$, $R > 0$.

Hypothèses 2.1

1. *Nous supposons que* $|\boldsymbol{\beta_0}| < +\infty$.
2. *Il existe* $B > 0$ *telle que :* $\forall (i,j) \in \{1,\dots,n\} \times \{1,\dots,p\}$, $|Z_{i,j}| \leq B$.

Proposition 2.2 *Soient* $k > 0$, $c > 0$ *et* $s = \mathrm{Card}\{j : (\beta_0)_j \neq 0\}$ *l'indice de sparsité de* $\boldsymbol{\beta_0}$. *Sous les Hypothèses 2.1 et sous une condition de compatibilité sur la matrice Hessienne de la log-vraisemblance partielle de Cox (cf. Guilloux et al. (2016)), nous avons avec probabilité supérieure à* $1 - cn^{-k}$, *pour une constante* $C(s) > 0$ *qui dépend de* s,

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq C(s)\sqrt{\frac{\ln(pn^k)}{n}}.$$

Estimation de α_0.

Pour l'estimation de α_0, nous avons considéré l'estimateur à noyau (2) proposé par Ramlau-Hansen (1983), dans lequel nous avons injecté l'estimateur LASSO (3). Nous avons besoin des hypothèses classiques suivantes sur la function de risque.

Hypothèses 2.3

1. *Pour tout* $i \in \{1,\dots,n\}$, *le processus aléatoire* Y_i *est à valeurs dans* $\{0,1\}$.
2. *Pour* $S(t,\boldsymbol{\beta_0}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0}^T \boldsymbol{Z_i}} Y_i(t)]$, *il existe* $c_S > 0$ *telle que* $S(t,\boldsymbol{\beta_0}) \geq c_S$, $\forall t \in [0,\tau]$.

3. $||\alpha_0||_{\infty,\tau} := \sup_{t\in[0,\tau]} \alpha_0(t) < \infty$.

Afin de choisir une fenêtre h pertinente, nous considérons dans un premier temps une grille de fenêtres $h > 0$, notée $\mathcal{H}_n$. Nous obtenons alors sur cette grille une collection d'estimateurs à noyau $\mathcal{F}(\mathcal{H}_n) = \{\hat{\alpha}_h^{\hat{\beta}}, h \in \mathcal{H}_n\}$. Nous faisons les hypothèses suivantes sur les fenêtres :

Hypothèses 2.4

1. $\text{Card}(\mathcal{H}_n) \leq n$, $nh \geq 1$ *et* $0 < h < 1$.

2. Pour une constante $a \geq 0$, $\sum_{h\in\mathcal{H}_n} \frac{1}{nh} = \mathcal{O}(\ln^a(n))$.

3. Pour tout $b > 0$, $\sum_{h\in\mathcal{H}_n} \exp(-b/h) < +\infty$.

Nous voulons sélectionner la fenêtre la plus pertinente dans la grille $\mathcal{H}_n$ afin de pouvoir ensuite sélectionner un estimateur dans la collection $\mathcal{F}(\mathcal{H}_n)$.

La fenêtre optimale, appelée oracle, est celle qui minimise l'excès de risque $\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2$. Elle est inaccessible car elle dépend de la fonction inconnue α_0. L'idée de la méthode est alors d'estimer une majoration de l'excès de risque, en introduisant un pseudo-estimateur $\bar{\alpha}_h$:

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2] \leq C\Big\{||\alpha_0 - K_h * \alpha_0||_2^2 + \mathbb{E}[||\bar{\alpha}_h - K_h * \alpha_0||]_2^2 + \mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||]_2^2\Big\},$$

où $C > 0$, $K_h(.) = (1/h)K(./h)$, et le pseudo-estimateur $\bar{\alpha}_h$, qui vérifie $\mathbb{E}[\bar{\alpha}_h] = K_h * \alpha_0$, est défini par

$$\bar{\alpha}_h(t) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \frac{K_h(t-u)}{S(u, \boldsymbol{\beta_0})} \mathrm{d}N_i(u). \tag{4}$$

En fait, $\bar{\alpha}_h$ correspond à l'estimateur à noyau de α_0 lorsque $S(u, \boldsymbol{\beta_0}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0}^T \boldsymbol{Z_i}} Y_i(u)]$ est connu. Dans un deuxième temps, nous bornons $\mathbb{E}||\bar{\alpha}_h - \hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}||^2$ par une constante $C(\hat{\beta}, \beta_0)$ qui ne dépend pas de h (cf. Guilloux et al. (2016)). Nous obtenons donc

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2] \leq C(\hat{\beta}, \beta_0) + C\Big\{||\alpha_0 - K_h * \alpha_0||_2^2 + \mathbb{E}||\bar{\alpha}_h - K_h * \alpha_0||_2^2\Big\}.$$

Soit donc h^* tel que

$$h^* = \underset{h\in\mathcal{H}_n}{\arg\min}\Big\{||\alpha_0 - K_h * \alpha_0||_2^2 + \mathbb{E} \parallel \bar{\alpha}_h - K_h * \alpha_0 \parallel_2^2 \Big\}.$$

L'idée est alors d'estimer h^* et plus précisément le terme inconnu $B(h) = ||\alpha_0 - K_h * \alpha_0||_2^2$, qui peut être vu comme un terme de biais. Sur le principe de la méthode de Goldenshluger et Lepski (2011), nous estimons le terme de biais comme suit :

$$A^{\hat{\boldsymbol{\beta}}}(h) = \sup_{h'\in\mathcal{H}_n} \Big\{||\hat{\alpha}_{h,h'}^{\hat{\boldsymbol{\beta}}} - \hat{\alpha}_{h'}^{\hat{\boldsymbol{\beta}}}||_2^2 - V(h')\Big\}_+,$$

où pour tout $t \geq 0$ et h, h' des réels positifs, $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}(t) = K_{h'} * \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h}(t)$, et où $V(h)$ est une borne supérieure du terme de variance $\mathbb{E}||\bar{\alpha}_h - K_h * \alpha_0||_2^2$. Plus précisément, $V(h)$ est obtenu à partir du contrôle de $\mathbb{E}[A^{\hat{\boldsymbol{\beta}}}(h)]$ en appliquant l'inégalité de Talagrand et nous obtenons un terme de l'ordre suivant $V(h) = \mathcal{O}(1/nh)$. À partir de ces définitions, nous déduisons le choix suivant de fenêtre :

$$\hat{h}^{\hat{\boldsymbol{\beta}}} = \underset{h \in \mathcal{H}_n}{\arg\min}\{A^{\hat{\boldsymbol{\beta}}}(h) + V(h)\}.$$

L'estimateur final est alors $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$.

3 Inégalité oracle non-asymptotique

Nous présentons à présent l'inégalité oracle non-asymptotique pour notre estimateur final. Ce résultat garantit les performances théoriques de notre estimateur.

Théorème 3.1 *Sous les Hypothèses 2.1, 2.3, 2.4 et sous une condition de compatibilité, il existe une constante κ telle que pour n suffisamment grand et $k \geq 12$,*

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||_2^2] \leq C \inf_{h \in \mathcal{H}_n} \Big\{ ||\alpha_h - \alpha_0||_2^2 + V(h) \Big\} + C'(s) \frac{\ln^a(n) \ln(pn^k)}{n}, \tag{5}$$

avec

$$V(h) = \kappa \frac{||\alpha_0||_{\infty,\tau} \tau}{c_S^2} \Big(||\alpha_0||_{\infty,\tau} \mathbb{E}[\mathrm{e}^{2\boldsymbol{\beta_0}^T \boldsymbol{Z_1}}] \tau + \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0}^T \boldsymbol{Z_1}}] \Big) \frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{nh},$$

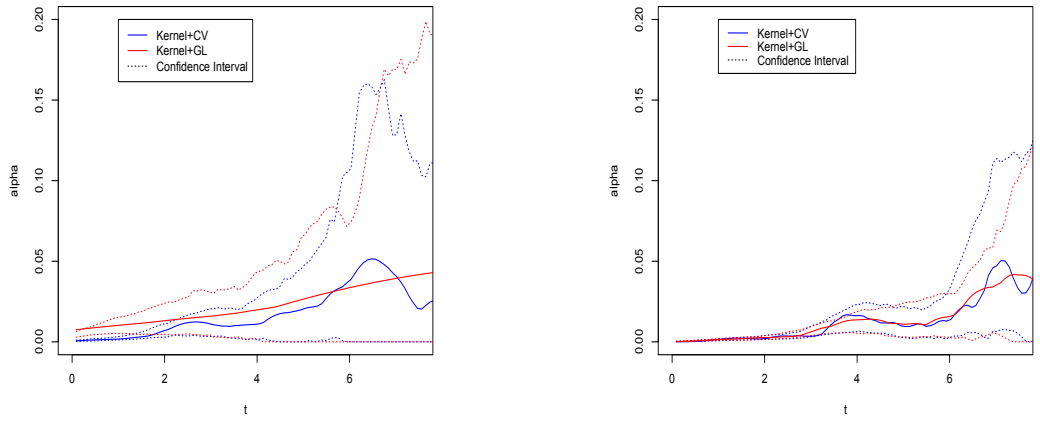
où C est une constante, $C'(s)$ une constante qui dépend de τ, κ_b la borne de l'inégalité de Bürkholder, B, $|\boldsymbol{\beta_0}|_1$, s, R, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_{\mathbb{L}^1(\mathbb{R})}$, $||K||_{\mathbb{L}^2(\mathbb{R})}$.

Cette inégalité assure que l'estimateur adaptatif à noyau $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$ réalise automatiquement le compromis biais-variance. Sans le terme en $\ln(p)/n$, nous retrouvons la vitesse de convergence dans le cas de l'estimation purement non-paramétrique, où le terme en $\ln^{a+1}(n)/n$ provient du contrôle de la différence entre l'estimateur à noyau (2) et le pseudo-estimateur (4). Le terme d'ordre $\ln(p)/n$ vient du contrôle de $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ donné par la Proposition 2.2. Ce terme est classique lorsque l'on estime le paramètre de régression en grande dimension.

4 Applications

Simulations. Nous avons généré des données simulées à partir d'un modèle de Cox avec des données censurées. Nous avons comparé les erreurs moyennes quadratiques intégrées (MISE) de notre estimateur à noyau $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$ avec l'estimateur à noyau dont la fenêtre a été sélectionnée par validation croisée. Notre estimateur fait mieux que celui obtenu avec la validation croisée. Nous avons donc, pour notre estimateur, à la fois une garantie théorique avec l'inégalité oracle, et de bons résultats en pratique.

Application à un jeu de données réelles. Nous avons appliqué les deux procédures d'estimation à noyau précédentes à un jeu de données réelles pour étudier la durée de survie sans rechute au cancer du sein en fonction d'un grand nombre de covariables (variables cliniques et niveaux d'expression de gènes, p=1000) chez deux groupes de patientes, les unes traitées au tamoxifen (n=142), les autres non traitées (n=104). Les données sont disponible sur le site www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6532. Nous avons représenté les risques de base estimés par les estimateurs à noyau avec la fenêtre sélectionnée soit par validation croisée, soit par la méthode de Goldenshluger et Lepski.



(a) Patients non-traités (p=1000). (b) Patients traités (p=1000).

Figure 1: Estimateurs à noyau avec une fenêtre choisie par validation croisée (bleu) ou avec la méthode de Goldenshluger et Lepski (rouge), et les intervalles de confiances à 90%.

Bibliographie

[1] Cox, R. (1972), Regression models and life-tables, *Journal of the Royal Statistical Society, Series B (Methodological)*, 34, 187–220.
[2] Goldenshluger A. et Lepski O. (2011), Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality, *Annals of Statistics*, 39, 1608–1632.
[3] Grégoire, G. (1993), Least squares cross-validation for counting process intensities *Scandinavian Journal of Statistics*, 20, 343–360.
[4] Guilloux, A., Lemler S. et Taupin M-L. (2016), Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates, *accepté dans JMVA*.
[5] Ramlau-Hansen, H. (1983), The choice of a kernel function in the graduation of counting process intensities, *Scandinavian Actuarial Journal*, 1983,165–182.
[6] Ramlau-Hansen, H. (1983), Smoothing counting process intensities by means of kernel functions, *Annals of Statistics*, 11, 453–466.
[7] Tibshirani, R. (1997), The Lasso method for variable selection in the Cox model, *Statistics in Medicine*, 16, 385–395.