Journal of Multivariate Analysis 148 (2016) 141–159

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates

Agathe Guilloux [a,b], Sarah Lemler [c,*], Marie-Luce Taupin [d,e]

[a] Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie - Paris 6, France
[b] Centre de Mathématiques Appliquées, Ecole Polytechnique, CNRS UMR 7641, Palaiseau, France
[c] Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes, École CentraleSupélec, France
[d] Laboratoire de Mathématiques et Modélisation d'Évry, UMR CNRS 8071- USC INRA, Université d'Évry Val d'Essonne, France
[e] Unité MaIAGE, INRA Jouy-En-Josas, France

ARTICLE INFO

Article history:
Received 6 July 2015
Available online 21 March 2016

AMS subject classifications:
62G05
62G08
62N01
62N02
62P10
62H12

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

ABSTRACT

We propose a novel kernel estimator of the baseline function in a general high-dimensional Cox model, for which we derive non-asymptotic rates 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 non-asymptotic oracle inequalities for the final estimator, which leads to a reduction in the rate of convergence when the dimension of the covariates grows.

1. Introduction

The Cox model, introduced by Cox [9], is a regression model often considered in survival analysis, which relates the distribution of a time T to the values of covariates. The hazard function of T is then defined by

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) \exp(\boldsymbol{\beta_0}^\top \boldsymbol{Z}), \tag{1}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ is a p-dimensional vector of covariates, $\boldsymbol{\beta_0} = (\beta_{0_1}, \ldots, \beta_{0_p})^\top$ the vector of regression coefficients and α_0 the baseline hazard function.

* Corresponding author.
E-mail addresses: agathe.guilloux@upmc.fr (A. Guilloux), sarah.lemler@centralesupelec.fr (S. Lemler), marie-luce.taupin@genopole.cnrs.fr (M.-L. Taupin).

http://dx.doi.org/10.1016/j.jmva.2016.03.002
0047-259X/

The regression parameter $\boldsymbol{\beta_0}$ and the baseline function α_0 are the two unknown parameters in this model. In previous works, more attention has been paid to the estimation of the regression parameter than to the estimation of the baseline function.

There are good reasons for this. First, the Cox partial log-likelihood, introduced by Cox [9], allows us to estimate $\boldsymbol{\beta_0}$ without knowledge of α_0. Second, the regression parameter is directly related to the covariates. Therefore, in order to select the relevant covariates that best explain the survival time, we need to estimate the regression parameter. Many papers deal with the problem of the estimation of $\boldsymbol{\beta_0}$, the number of covariates p being large (or not) compared with the number of individuals n. When p is smaller than n, the usual estimator of $\boldsymbol{\beta_0}$ is obtained by maximizing the Cox partial log-likelihood (see Andersen et al. [2] as a good reference). When the number of covariates grows, the LASSO procedure is often considered. This consists of a minimization of the negative ℓ_1-penalized Cox partial log-likelihood. Asymptotic results are stated in Bradic et al. [4], Kong and Nan [18] and Bradic and Song [5]. Lastly, the non-asymptotic rate of convergence of the LASSO is now known to be of order $\sqrt{\ln p/n}$, see Huang et al. [17].

The estimation of the baseline function α_0 has been less studied. One known estimator of the baseline function is a kernel estimator, introduced by Ramlau-Hansen [23,24]. We present here its form in the special case of right-censoring. Let us consider, for the moment, that we observe for $i = 1, \ldots, n$, $(X_i, \delta_i, \boldsymbol{Z_i})$, where $X_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}_{\{T_i \le C_i\}}$, T_i is the time of interest, and C_i the censoring time. The usual kernel estimator is then obtained from an estimator of the cumulative baseline function A_0 defined by $A_0(t) = \int_0^t \alpha_0(s)\mathrm{d}s$. This estimator is called the Breslow estimator and is defined, for $t > 0$, by

$$\hat{A}_0(t, \hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \frac{\delta_i}{S_n(X_i, \hat{\boldsymbol{\beta}})}, \quad \text{with } S_n(t, \hat{\boldsymbol{\beta}}) = \sum_{i:T_i \ge t} \exp(\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z_i}), \tag{2}$$

see Ramlau-Hansen [24] and Andersen et al. [2] for details. From $\hat{A}_0(\cdot, \hat{\boldsymbol{\beta}})$, the kernel function estimator for α_0 is derived by smoothing the increments of the Breslow estimator. It is defined by

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{h} \int_0^\tau K\Big(\frac{t-u}{h}\Big) \mathrm{d}\hat{A}_0(u, \hat{\boldsymbol{\beta}}), \quad \tau \ge 0, \tag{3}$$

with $K : \mathbb{R} \mapsto \mathbb{R}$ a kernel with integral 1, and h a positive parameter called the bandwidth. This estimator was introduced and studied by Ramlau-Hansen [23,24] within the framework of the multiplicative intensity model for counting processes, thereby extending its use to censored survival data. Consistency and asymptotic normality are proven in Ramlau-Hansen [24] with fixed bandwidth.

The choice of the bandwidth in kernel estimation is crucial, in particular when one is interested in establishing non-asymptotic adaptive inequalities. State-of-the-art methods are based on cross-validation. Ramlau-Hansen [22] has suggested the cross-validation method to select the bandwidth but without any theoretical guarantees. For randomly censored survival data, Marron and Padgett [21] have shown that the cross-validation method gives the optimal bandwidth for estimating the density: the ratio between the integrated squared error for the cross-validation bandwidth and the infimum of the integrated squared error for any bandwidth almost surely converges to 1. Grégoire [15] has considered the cross-validated method suggested by Ramlau-Hansen [22] for adaptive estimation of the intensity of a counting process and has proved some consistency and asymptotic normality results for the cross-validated kernel estimator.

However, all the results for the adaptive kernel estimator with a cross-validated bandwidth are asymptotic. No non-asymptotic oracle inequalities have to date been stated for the kernel estimator of the baseline function. In addition, to our knowledge, the construction of $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$ has not yet been considered for high-dimensional covariates. The goal of the present paper is thus twofold: whatever the dimension, we aim to propose an estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}$ of the baseline function, for which we can establish a non-asymptotic oracle inequality to measure its performance. The loss of prediction quality of $|\hat{\alpha}^{\hat{\boldsymbol{\beta}}} - \alpha_0|$ when p increases will be quantified.

To fulfill these purposes, the idea is to first estimate the regression parameter $\boldsymbol{\beta_0}$ via a LASSO procedure applied to the Cox partial log-likelihood, then to plug this estimator in the usual kernel estimator (3) of the baseline hazard function; then, lastly, to select a data-driven bandwidth, following a procedure adapted from Goldenshluger and Lepski [14]. In the latter, the problem of bandwidth selection in kernel density estimation is addressed and an adaptive estimator is derived, which satisfies non-asymptotic minimax bounds. This method was then considered by Doumic et al. [11] for estimating the division rate of a size-structured population in a non-parametric setting, by Bouaziz et al. [3] to estimate the intensity function of a recurrent event process, and by Chagny [8] for the estimation of a real function via a warped kernel strategy. In the present paper, we consider this method in order to obtain an adaptive kernel estimator of the baseline function with a data-driven bandwidth. We establish the first adaptive and non-asymptotic oracle inequality, which guarantees the theoretical performance of this kernel estimator. The oracle inequality depends on non-asymptotic control of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ deduced from an estimation inequality in Huang et al. [17] and extended to the case of unbounded counting processes (see Guilloux et al. [16] for details).

The paper is organized as follows. In Section 3, we describe the two-step procedure to estimate the baseline function: first, we describe the estimation of $\boldsymbol{\beta_0}$ as a preliminary step and give the bound for $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$, and then we focus on the