ARTICLE IN PRESS
Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates

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

[a] Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie - Paris 6, France
[b] Laboratoire de Mathématiques et de Modélisation d'Évry, UMR CNRS 8071- USC INRA, Université d'Évry Val d'Essonne, France
[c] Unité MaIAGE, INRA Jouy-En-Josas, France

ARTICLE INFO

Article history:
Received 26 February 2015
Received in revised form 4 September 2015
Accepted 26 November 2015
Available online xxxx

Keywords:
Survival analysis
Conditional hazard rate function
Cox's proportional hazards model
Right-censored data
Semi-parametric model
Nonparametric model
High-dimensional covariates
Model selection
Non-asymptotic oracle inequalities
Concentration inequalities

ABSTRACT

The purpose of this article is to provide an adaptive estimator of the baseline function in the Cox model with high-dimensional covariates. We consider a two-step procedure : first, we estimate the regression parameter of the Cox model via a Lasso procedure based on the partial log-likelihood, secondly, we plug this Lasso estimator into a least-squares type criterion and then perform a model selection procedure to obtain an adaptive penalized contrast estimator of the baseline function.

Using non-asymptotic estimation results stated for the Lasso estimator of the regression parameter, we establish a non-asymptotic oracle inequality for this penalized contrast estimator of the baseline function, which highlights the discrepancy of the rate of convergence when the dimension of the covariates increases.

1. Introduction

Consider the following Cox model, introduced by Cox (1972) and defined, for a vector of covariates $\boldsymbol{Z} = (Z_1, \dots, Z_p)^T$, by

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

where λ_0 denotes the hazard rate, $\boldsymbol{\beta_0} = (\beta_{0_1}, \dots, \beta_{0_p})^T \in \mathbb{R}^p$ is the regression parameter and α_0 is the baseline hazard function. The Cox partial log-likelihood, introduced by Cox (1972), allows to estimate $\boldsymbol{\beta_0}$ without the knowledge of α_0, considered as a functional nuisance parameter. For the estimation of α_0, one common way is to use a two step procedure, starting with the estimation of $\boldsymbol{\beta_0}$ alone and then to plug this estimator into a non parametric type estimator α_0, usually a kernel type estimator.

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

http://dx.doi.org/10.1016/j.jspi.2015.11.005

Please cite this article in press as: Guilloux, A., et al., Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates. J. Statist. Plann. Inference (2015), http://dx.doi.org/10.1016/j.jspi.2015.11.005

Let us be more specific.

When p is small compared to n, $\boldsymbol{\beta_0}$ is usually estimated by minimization of the opposite of the Cox partial log-likelihood. We refer to Andersen et al. (1993), as a reference book, for the proofs of the consistency and the asymptotic normality of $\hat{\boldsymbol{\beta}}$ when p is small compared to n. Those strategies only apply when $p < n$ and even more, they only apply when p is small compared to n. When p grows up, becoming of the same order as n and possibly larger than n, various well known problems appear. Among them, the minimization of the opposite of the Cox partial log-likelihood becomes difficult and even impossible if $p > n$.

In high-dimension, when p is large compared to n, the Lasso procedure is one of the classical considered strategies. The Lasso (Least Absolute Shrinkage and Selection Operator) has been first introduced by Tibshirani (1996) in the linear regression model. It has been largely considered in additive regression model (see for instance Knight and Fu, 2000, Efron et al., 2004, Donoho et al., 2006, Meinshausen and Bühlmann, 2006, Zhao and Yu, 2006, Zhang and Huang, 2008, Meinshausen and Yu, 2009 and also Juditsky and Nemirovski, 2000, Nemirovski, 2000, Bunea et al., 2006, 2007c,a, Greenshtein and Ritov, 2004 or Bickel et al., 2009), and in density estimation (see Bunea et al., 2007b and Bertin et al., 2011). In the particular case of the semi-parametric Cox model, Tibshirani (1997) has proposed a Lasso procedure for the regression parameter. The Lasso estimator of the regression parameter $\hat{\boldsymbol{\beta}}$ is defined as the minimizer of the opposite of the Cox partial log-likelihood under an ℓ_1 type constraint, that is, suitably penalized with an ℓ_1-penalty function. Recent results exist on the estimation of $\boldsymbol{\beta_0}$ in high-dimension setting. Among them one can mention (Bradic et al., 2012) who have proved asymptotic results for Lasso estimator. More recently, Bradic and Song (2012), Kong and Nan (2012) and Huang et al. (2013) establish the first non-asymptotic oracle inequalities (estimation and prediction bounds) for the Lasso estimator.

For the baseline hazard function and when p is small compared to n, the common estimator is a kernel estimator, which depends on $\hat{\boldsymbol{\beta}}$ obtained by minimization of the opposite of the Cox partial log-likelihood. This kernel estimator has been introduced by Ramlau-Hansen (1983a,b) from the Breslow estimator of the cumulative baseline function (see Ramlau-Hansen, 1983b and Andersen et al., 1993 for more details). In this context, Ramlau-Hansen (1983b) and Grégoire (1993) proved asymptotic results. No non-asymptotic results and no adaptive results have to date been established for the kernel estimator of the baseline function. Finally, when p is large compared to n, to our knowledge, the construction of an estimator of the baseline function has not been yet considered.

In this paper, we consider a two-step procedure to estimate $\boldsymbol{\beta_0}$ and α_0, the two parameters in the Cox model. But our contributions focus more on the estimation of α_0. In the Cox model we consider, it is noteworthy that the high-dimension only concerns the regression parameter, whereas the baseline function is a time function. Its estimation would not require a procedure specific to high-dimension, besides the first step concerning the estimation of $\boldsymbol{\beta_0}$. We propose a procedure for the construction of an estimator of the baseline hazard function α_0, p being either smaller than n or greater than n. It combines a Lasso procedure for $\boldsymbol{\beta_0}$ as a first step and a second step based on a model selection strategy for the estimation of the baseline function α_0. This model selection procedure takes its origins in the works of Akaike (1973) and Mallows (1973), more recently formalized by Birgé and Massart (1997) and Barron et al. (1999) for the estimation of densities and regression functions (see the book of Massart, 2007 as a reference work on model selection). In survival analysis, the model selection has also been documented. Letué (2000) has adapted these methods to estimate the regression function of the non-parametric Cox model, when $p < n$. More recently, Brunel and Comte (2005), Brunel et al. (2009) and Brunel et al. (2010) have obtained adaptive estimation of densities in a censoring setting. Model selection methods have also been used to estimate the intensity function of a counting process in the multiplicative Aalen intensity model (see Reynaud-Bouret, 2006 and Comte et al., 2011). However, the model selection procedure has never been considered, to our knowledge, for estimating the baseline hazard function in the Cox model.

Our contributions are at least threefold: Our procedure is the first that focus on the estimation of baseline function of the semi-parametric Cox model with high-dimensional covariates. This procedure provides an adaptive estimator of the baseline function that works as well for small p and large p compared to n (that is for possibly high-dimensional covariates). Furthermore, for this estimator, we state non-asymptotic oracle inequalities, that hold, once again, p being either smaller than n or greater than n. More precisely, we prove that the risk of this estimator achieves the best risk among estimators in a large collection. For each model, the risk of an estimator is bounded by the sum of three terms. The first term is a bias term involving to the approximation properties of the collection of models, through the distance evaluated in $\boldsymbol{\beta_0}$ between the true baseline and the orthogonal projection of α_0 on the best selected model. The second term is a penalty term of the same order than the variance on one model, that is of order the dimension of one model over n, as expected with ℓ_0-penalty. These two terms are the "usual" terms appearing in nonparametric estimation. It is noteworthy that these two terms do not involve any quantity related to the risk of the Lasso estimator of $\boldsymbol{\beta_0}$. The last term precisely comes from the properties of the Lasso estimator of $\boldsymbol{\beta_0}$. This last term is of order $\log(np)/n$, as expected for a Lasso estimator.

When p is small, the third last term is of order $\log(n)/n$ and, the rate is governed by the first two terms. In that case, the penalty term being of the same order than the variance over one model, we conclude that the model selection procedure achieves the "expected rate" of order $n^{-2\gamma/(2\gamma+1)}$ when the baseline function belongs to a Besov space with smoothness parameter γ. This continues to hold when p is of the same order than the sample size n. When p is larger than n, that is in the so-called ultra-high dimension (see Verzelen, 2012), the rate for estimating α_0 is changed, and more precisely degraded as a price to pay for being with high dimension covariates. This degradation follows accordingly to the order of p compared to n.

Please cite this article in press as: Guilloux, A., et al., Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates. J. Statist. Plann. Inference (2015), http://dx.doi.org/10.1016/j.jspi.2015.11.005