Journal of Multivariate Analysis 161 (2017) 103–122

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Variable selection and structure identification for varying coefficient Cox models

Toshio Honda [a,*], Ryota Yabe [b]

[a] Graduate School of Economics, Hitotsubashi University, Kunitachi, Tokyo 186-8601, Japan
[b] Department of Economics, Shinshu University, Matsumoto, Nagano 390-8621, Japan

ARTICLE INFO

Article history:
Received 18 July 2016
Available online 29 July 2017

AMS 2010 subject classifications:
62G08
62N01

Keywords:
B-spline basis
Censored survival data
Group Lasso
High-dimensional data
Semi-varying coefficient model
Structured nonparametric regression model

ABSTRACT

We consider varying coefficient Cox models with high-dimensional covariates. We apply the group Lasso to these models and propose a variable selection procedure. Our procedure can cope with simultaneous variable selection and structure identification for high-dimensional varying coefficient models to find true semi-varying coefficient models from them. We also derive an oracle inequality and closely examine restrictive eigenvalue conditions. We focus on Cox models with time-varying coefficients. The theoretical results on variable selection can be extended easily to some other important models which we only mention briefly since they can be treated in the same way. The models considered here are the most popular among structured nonparametric regression models. The results of numerical studies are also reported.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Cox's proportional hazards model is one of the most popular and useful models for censored survival data analysis. Since this model was originally proposed by Cox [9], it has been extended in many ways to deal with complicated situations or to carry out more flexible analyses. In this paper, we consider varying coefficient Cox models and additive Cox models with high-dimensional covariates. When the number of covariates is moderate, these models have already been investigated in many papers, including Huang et al. [17], Cai and Sun [8], and Cai et al. [7].

We apply the group Lasso as described, e.g., in Lounici et al. [25] and Huang et al. [16], to varying coefficient models with high-dimensional covariates to carry out variable selection and structure identification simultaneously. Although we focus on time-varying coefficient models, our method can be applied to variable selection for other types of varying coefficient models and additive models, for which we briefly mention how our procedures and the corresponding theoretical results can be adapted.

Suppose that for each $i \in \{1, \ldots, n\}$, we observe a censored survival time T_i and a high-dimensional random vector $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^\top$ of covariates. More specifically, we have i.i.d. observations such that, for each $i \in \{1, \ldots, n\}$, $T_i = \min(T_{0i}, C_i)$, $\delta_i = \mathbf{1}(T_{0i} \leq C_i)$, $\boldsymbol{X}_i(t)$ is a p-dimensional covariate on the time interval $[0, \tau]$, where T_{0i} is an uncensored survival time and C_i is a censoring time satisfying subject to the independent censoring mechanism described in Section 6.2 of Kalbfleisch and Prentice [20]. Hereafter we set $\tau = 1$ for simplicity of presentation. Note that p can be very large compared to n in this paper, e.g., $p = O(n^{c_p})$ for a very large positive constant c_p or $p = O\{\exp(n^{c_p})\}$ for a sufficiently small positive

* Corresponding author.
E-mail addresses: t.honda@r.hit-u.ac.jp (T. Honda), ryotayabe@shinshu-u.ac.jp (R. Yabe).

http://dx.doi.org/10.1016/j.jmva.2017.07.007
0047-259X/© 2017 Elsevier Inc. All rights reserved.

constant c_p. We assume that the standard setup for the Cox model holds as in Chapter 5 of [20] and that T_i or $N_i(t) = \mathbf{1}(t \geq T_i)$ has the following compensator $\Lambda_i(t)$ with respect to a suitable filtration $\{\mathcal{F}_t\}$:

$$d\Lambda_i(t) = Y_i(t)\exp\{\boldsymbol{X}_i^\top(t)\boldsymbol{g}(t)\}\lambda_0(t)dt, \tag{1}$$

where $Y_i(t) = \mathbf{1}(t \leq T_i)$, $\boldsymbol{g}(t) = (g_1(t), \ldots, g_p(t))^\top$ is a vector of unknown functions on $[0, 1]$, $\boldsymbol{a}^\top$ denotes the transpose of $\boldsymbol{a}$, and $\lambda_0(t)$ is an unknown baseline hazard function; see Subsection 5.2.1 of [20] for an example of $\{\mathcal{F}_t\}$. Then as in Chapter 5 of [20], $\boldsymbol{X}_i(t)$ is predictable and $M_i(t) = N_i(t) - \Lambda_i(t)$ is a martingale process with respect to $\{\mathcal{F}_t\}$. In the original Cox model, $\boldsymbol{g}(t)$ is a vector of unknown constants and we estimate this constant coefficient vector by maximizing the partial likelihood.

In this paper, we are interested in estimating $\boldsymbol{g}(t)$ in (1). Recently, with the fast development of data collection technologies, (ultra) high-dimensional data are becoming more frequent. In such high-dimensional data, usually only a small fraction of covariates is relevant. However, we cannot directly apply standard or traditional estimating procedures to such high-dimensional data. Many methods for variable selection have thus been developed, e.g., the SCAD and the Lasso. See Bühlmann and van de Geer [6] and Hastie et al. [14] for excellent reviews of these procedures for variable selection; see also Bickel et al. [2] and Zou [41] for the Lasso and the adaptive Lasso, respectively.

Cox models with constant coefficients have already been studied in high-dimensional contexts by Bradic et al. [3], who studied the SCAD, and by Huang et al. [18], Kong and Nan [22], and Lemler [23], who considered the Lasso. Zhang and Luo [36] proposed an adaptive Lasso estimator for the Cox model. The authors of [18] developed new ingenious techniques to derive oracle inequalities. We will fully use their techniques to derive our theoretical results such as an oracle inequality. Sun et al. [28] modified the Lasso penalty to incorporate side information. Wang et al. [32] proposed a hierarchical group penalty. Some variable screening procedures have also been proposed in Zhao and Li [39] and Yang et al. [34], to name just a few. Estimation of the baseline hazard function is considered in Guilloux et al. [13] in a high-dimensional setup. A model-free screening procedure for censored data with high-dimensional covariates is proposed in Song et al. [27].

In this paper, we propose a group Lasso procedure to select relevant covariates and identify the covariates with constant coefficients among the relevant covariates, namely the true semi-varying coefficient model from a much larger varying coefficient model. We can achieve this goal by a suitable two-stage procedure consisting of the proposed group Lasso either with an adaptively weighted Lasso procedure as in [15,33], or with the SCAD. In [33], the authors proposed an adaptive Lasso procedure for structure identification but did not provide any theoretical support. Our procedure can be applied to the varying coefficient model with an index variable $Z_i(t)$, viz.

$$d\Lambda_i(t) = Y_i(t)\exp[g_0\{Z_i(t)\} + \boldsymbol{X}_i^\top(t)\boldsymbol{g}\{Z_i(t)\}]\lambda_0(t)dt$$

and to the additive model

$$d\Lambda_i(t) = Y_i(t)\exp\Big[\sum_{j=1}^{p} g_j\{X_{ij}(t)\}\Big]\lambda_0(t)dt.$$

We will return to the latter models in Section 4.

Some authors considered the same problem by using the SCAD penalty; see, e.g., Lian et al. [24] and Zhang et al. [37]. They proved the existence of a local optimizer satisfying the same convergence rate as ours. In contrast, we prove the existence of the global solution with desirable properties. In Bradic and Song [4], the authors applied penalties similar to ours to additive models and obtained theoretical results with possible model misspecifications. In our context, the convergence rates are different from theirs; see Remark 1 in Section 3 for more details. We also carefully examined the restrictive eigenvalue (RE) conditions. While some authors considered the L_2 norm of the estimated second derivatives for additive models, we adopt the orthogonal decomposition approach to structure identification. This is explained around (2) and (20) and in Appendix A. We give some details on why we have adopted the orthogonal decomposition approach in Appendix C.

This paper is organized as follows. In Section 2, we describe our group Lasso procedure for time-varying coefficient models. Then we present our theoretical results in Section 3. We mention the two other models in Section 4. The results of numerical studies are reported in Section 5. The proofs of our theoretical results are postponed to Section 6 and Section 7 concludes this paper. We collected useful properties of our basis functions and the proofs of technical lemmas in Appendices A–D.

Before proceeding, we define some notation and symbols here. In this paper, $C, C_1, C_2, \ldots$ are positive generic constants and their values change from line to line. For a vector $\boldsymbol{a}$, $|\boldsymbol{a}|$, $|\boldsymbol{a}|_1$, and $|\boldsymbol{a}|_\infty$ mean the L_2 norm, the L_1 norm, and the sup norm, respectively. For a function g on $[0, 1]$, $\|g\|$, $\|g\|_1$, and $\|g\|_\infty$ stand for the L_2 norm, the L_1 norm, and the sup norm, respectively. For a symmetric matrix A, we denote the minimum and maximum eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. Besides, $\mathrm{sign}(a)$ is the sign of a real number a and $a_n \sim b_n$ means there are positive constants C_1 and C_2 such that $C_1 < a_n/b_n < C_2$. We write $\overline{\mathcal{S}}$ for the complement of a set $\mathcal{S}$. For a function g and a constant c, $g \equiv c$ and $g \not\equiv c$ means that a function g is c and it is not a constant c, respectively.

2. Group Lasso procedure

For each $j \in \{1, \ldots, p\}$, we first decompose $g_j(t)$ into its constant and non-constant part, viz.

$$g_j(t) = g_{cj} + g_{nj}(t), \tag{2}$$

Download English Version:

https://daneshyari.com/en/article/5129317

Download Persian Version:

https://daneshyari.com/article/5129317

Daneshyari.com