Variable selection in Cox regression models with varying coefficients

Toshio Honda[1,3] and Wolfgang Karl Härdle[2]
[1]Graduate School of Economics, Hitotsubashi University, JAPAN
[2]C.A.S.E. - Center for Applied Statistics & Economics,
Humboldt-Universität zu Berlin, GERMANY
[3]Corresponding author: Toshio Honda, e-mail: honda@econ.hit-u.ac.jp

Abstract

We deal with two kinds of Cox regression models with varying coefficients. The coefficients vary with time in one model. In the other model, there is an important random variable called an index variable and the coefficients vary with the variable. In both models, we have p-dimensional covariates and p increases moderately. However, it is the case that only a small part of the covariates are relevant in these situations. We carry out variable selection and estimation of the coefficient functions by using the group SCAD-type estimator or the adaptive group Lasso estimator. We focus on time varying coefficient models here. We examine the theoretical properties of the estimators, especially the L_2 convergence rate, the sparsity, and the oracle property.

Keywords: adaptive group Lasso, B-splines, group SCAD, high-dimensional data, oracle estimator, sparsity.

1 Introduction

The Cox regression model is one of the most popular and useful models in survival analysis. In recent years, many nonparametric and semiparametric variants of the Cox regression model have been proposed. Among them, there are varying coefficient models, partially linear models and their extensions, and additive and functional ANOVA models. In this research, we consider varying coefficient models and consider two kinds of Cox regression models with varying coefficients. The coefficients vary with time in one model and with index variable $U(t)$ in another model. We focus on the former here and omit technical details due to the space limitation. See Honda and Härdle (2012) for the details omitted here.

In recent years, a dimensional and model selection issue occurs in many applications : only a small part of the variables are relevant. Therefore statistical methods for variable selection are needed. Penalized likelihood estimators such as the Lasso or SCAD estimators have been among the standard tools in carrying out variable selection and estimation simultaneously, Tibshirani (1996) and Fan and Li (2001). Zou (2006) proposed the adaptive Lasso to correct some deficiencies of the Lasso and proved that the adaptive Lasso estimators choose the relevant variables consistently.

For Cox regression models with time varying coefficients, we deal with the cases where the number of the covariates, p, increase moderately with the sample size, for example $p = \mathcal{O}(n^{3/10})$, where n is the sample size. We conduct variable selection and estimation simultaneously by employing group SCAD-type or adaptive group Lasso estimators.

Variable selection and estimation in Cox regression models are considered in many papers. For example, Leng and Zhang (2006), Zhang and Lu (2007), Du et al. (2010), Bradic et al. (2011), Zhang et al. (2012), Lian et al. (2013), and Hu and Lian (2013). However, they have not considered varying coefficient models. Recently Yan and Huang (2012) proposed the adaptive group Lasso in a Cox regression model with time-varying coefficients. There is however still a lacuna of theoretical results that this research aims to fill. We establish the sparsity for the group SCAD-type and adaptive group Lasso estimator and the oracle property for the group SCAD-type estimator under simple and interpretable assumptions. The derivation of the theoretical results of this paper crucially depend on the methodology of Huang et al. (2000).

We state the setup of the time-varying coefficient model and define the partial likelihood estimator, the group SCAD-type estimator, and the adaptive group Lasso estimator in section 2. We consider the asymptotics and establish the sparsity and the oracle property of the estimators in section 3 . We denote the Euclidean norm and the transpose of a vector v by $|v|$ and v^{T}, respectively.

2 Assumptions and estimators

In this section, we describe the Cox regression model with time-varying coefficients, state some assumptions, and define the group SCAD-type and adaptive group Lasso estimator. In deriving the main results, we repeatedly use insights of Huang et al. (2000), of which we also borrow the notation.

Let T and C be a failure time and a censoring time. The interest is in the failure time. However, we observe only $Y = \min\{T, C\}$ on $[0, \tau]$ subject to censoring for some finite τ and $\delta = \mathbf{I}(T \leq C)$. We define

$$N(t) = \delta\, \mathbf{I}(Y \leq t) \quad \text{and} \quad Z(t) = \mathbf{I}(Y \geq t).$$

We also observe a p-dimensional time-dependent covariate $\boldsymbol{X}(t)$. Suppose that $(Y_i, \delta_i, \boldsymbol{X}_i(t))$, where $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^\top$, $i = 1, \ldots, n$, are i.i.d. observations of $(Y, \delta, \boldsymbol{X}(t))$. The hazard function of T_i w.r.t. an appropriate filtration is given by

$$\lambda(t) = \lambda_0(t) \exp\Big\{ \sum_{j=1}^{p} g_{0j}(t) X_{ij}(t) \Big\} = \lambda_0(t) \exp\{\boldsymbol{g}_0^\top(t) \boldsymbol{X}_i(t)\}, \tag{1}$$

where $\lambda_0(t)$ is an unknown hazard function and $\boldsymbol{g}_0(t) = (g_{01}(t), \ldots, g_{0p}(t))^\top$ is a vector of unknown time-varying coefficients and assumed to be twice continuously differentiable.

We estimate $\boldsymbol{g}_0(t)$ by choosing a basis $\{B_1(t), \ldots, B_{K_n}(t)\}$ on $[0, \tau]$ and maximizing the partial likelihood with/without a penalty term. We allow p to increase moderately (e.g. $p = \mathcal{O}(n^{3/10})$) and consider variable selection.

More precisely for the basis $\{B_1, \ldots, B_{K_n}\}$, we write

$$\boldsymbol{B}(t) = (B_1(t), \ldots, B_{K_n}(t))^\top \quad \text{or} \quad \boldsymbol{B} = (B_1, \ldots, B_{K_n})^\top.$$

Then the covariate vector for the partial likelihood is $\boldsymbol{X}_i(t) \otimes \boldsymbol{B}(t)$. The approximation error ρ_n of the basis $\{B_1, \ldots, B_{K_n}\}$ is defined by

$$\rho_n = \sup_{\boldsymbol{g}_0} \sum_{j=1}^{p} \inf_{\beta_j \in \mathbb{R}^{K_n}} \sup_{0 \le t \le \tau} |\beta_j^\top \boldsymbol{B}(t) - g_{0j}(t)|, \tag{2}$$

where $\boldsymbol{g}_0 = (g_{01}, \ldots, g_{0p})^\top$ is over the set of functions satisfying Assumption G in Honda and Härdle (2012). Then we have $\rho_n = \mathcal{O}(K_n^{-2})$ by the standard theory. An example of the basis is an equi-spaced B-spline basis of order $m (m \ge 2)$.

We define the estimation space $\mathbf{G}_0$ by

$$\mathbf{G}_0 = \{(\beta_1^\top \boldsymbol{B}(t), \ldots, \beta_p^\top \boldsymbol{B}(t))^\top \,|\, \beta_j \in \mathbb{R}^{K_n},\ j = 1, \ldots, p\}. \tag{3}$$

For $\boldsymbol{g} = (g_1, \ldots, g_p)^\top \in \mathbf{G}_0$, we define $\|\boldsymbol{g}\|_{L_2}$ by

$$\|\boldsymbol{g}\|_{L_2}^2 = \sum_{j=1}^{p} \|g_j\|_{L_2}^2 = \sum_{j=1}^{p} \frac{1}{\tau} \int_0^\tau g_j^2(t) dt. \tag{4}$$

Then the partial likelihood $l_p(\boldsymbol{g})$ is defined by

$$\begin{aligned} l_p(\boldsymbol{g}) \;=\; & \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \boldsymbol{g}^\top(t) \boldsymbol{X}_i(t) dN_i(t) \\ & - \int_0^\tau \log \left[n^{-1} \sum_{i=1}^{n} Z_i(t) \exp \left\{ \boldsymbol{g}^\top(t) \boldsymbol{X}_i(t) \right\} \right] d\bar{N}(t), \end{aligned} \tag{5}$$

where $\bar{N}(t) = n^{-1} \sum_{i=1}^{n} N_i(t)$.

Finally in this section, we define three estimators of $\boldsymbol{g}_0$. The first one is the partial likelihood estimator and defined by

$$\tilde{\boldsymbol{g}}_n = \operatorname*{argmax}_{\boldsymbol{g} \in \mathbf{G}_0} l_p(\boldsymbol{g}) \tag{6}$$

It will be shown in Theorem 1 below that the L_2 convergence rate of $\tilde{\boldsymbol{g}}_n$ is :

$$r_{pn} = \max\{(pK_n/n)^{1/2}, \rho_n\}. \tag{7}$$

We introduce the sparsity assumption.

Assumption S: For some s, $g_{0j} = 0,\ s+1 \le j \le p$.

To deal with this sparsity, we present two penalized partial likelihoods $Q_p(\boldsymbol{g})$ and $\bar{Q}_p(\boldsymbol{g})$ for $\boldsymbol{g} = (g_1, \ldots, g_p)^\top \in \mathbf{G}_0$.

$$Q_p(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \sum_{j=1}^{p} p_{\lambda_n}(\|g_j\|_{L_2}), \tag{8}$$

where λ_n is a tuning parameter and $p_\lambda(\cdot)$ is a SCAD-type penalty function to be specified in Assumption P below.

Assumption P:
(i) $p_\lambda(t)$ is a monotone increasing and concave function on $[0,\infty)$ with $p_\lambda(0)=0$. Besides, there are positive constants a_0, b_0, and c_0 such that $p'_\lambda(t)=0,\ t\geq a_0\lambda$, and $p'_\lambda(t)\geq c_0\lambda,\ 0<t\leq b_0\lambda$.
(ii) $\lambda_n/r_{pn}\to\infty$ and $\min_{1\leq j\leq s}\|g_{0j}\|_{L_2}/\lambda_n\to\infty$.

Another penalized partial likelihood $\bar{Q}_p(\boldsymbol{g})$ is defined by

$$\bar{Q}_p(\boldsymbol{g})=l_p(\boldsymbol{g})-\lambda'_n\sum_{j=1}^{p}w_j\|g_j\|_{L_2}, \tag{9}$$

where λ'_n is another tuning parameter and w_j, $j=1,\ldots,p$, are weights to be constructed from a preliminary estimator. Notice that $\bar{Q}_p(\boldsymbol{g})$ is a concave function.

The group SCAD-type estimator $\hat{\boldsymbol{g}}_n$ and the adaptive group Lasso estimator $\bar{\boldsymbol{g}}_n$ are given by

$$\hat{\boldsymbol{g}}_n=\operatorname*{argmax}_{\boldsymbol{g}\in\mathbf{G}_0}Q_p(\boldsymbol{g})\quad\text{and}\quad\bar{\boldsymbol{g}}_n=\operatorname*{argmax}_{\boldsymbol{g}\in\mathbf{G}_0}\bar{Q}_p(\boldsymbol{g}). \tag{10}$$

We can also define $l_s(\boldsymbol{g})$, $Q_s(\boldsymbol{g})$, and $\bar{Q}_s(\boldsymbol{g})$ for the s in Assumption S by ignoring the last $(p-s)$ elements of the covariates or taking $\boldsymbol{X}_i(t)=(X_{i1}(t),\ldots,X_{is}(t))^\top$ and $\boldsymbol{g}=(g_1,\ldots,g_s)^\top$.

3 Main theorems

The L_2 convergence rate of the partial likelihood estimator $\tilde{\boldsymbol{g}}_n$ is derived in Theorem 1.

Theorem 1 *Suppose that the necessary technical assumptions hold. Then with probability tending to 1, there is a unique maximizer* $\tilde{\boldsymbol{g}}_n=(\tilde{g}_{n1},\ldots,\tilde{g}_{np})^\top$ *of* $l_p(\boldsymbol{g})$ *over* $\mathbf{G}_0$ *and we have*

$$\|\tilde{\boldsymbol{g}}_n-\boldsymbol{g}_0\|_{L_2}=\mathcal{O}_p(r_{pn}).$$

The existence of the group SCAD-type estimator is verified in Theorem 2 and the sparsity and oracle property is established is Theorem 3.

Theorem 2 *Suppose that all the assumptions in Theorem 1 and Assumptions P and S hold. Then for any positive* ϵ, *there is a positive constant* M *such that*

$$\lim_{n\to\infty}\mathsf{P}(\textit{ There is a local maximizer }\hat{\boldsymbol{g}}_n \textit{ of } Q_p(\boldsymbol{g}) \textit{ over } \mathbf{G}_0 \textit{ such that } \|\hat{\boldsymbol{g}}_n-\boldsymbol{g}_0\|_{L_2}\leq Mr_{pn})>1-\epsilon.$$

Before we present Theorem 3, we define two properties. If the local maximizer of $\hat{\boldsymbol{g}}_n=(\hat{g}_{n1},\ldots,\hat{g}_{np})^\top$ satisfies under Assumption S,

$$\hat{g}_{nj}=0,\ j=s+1,\ldots,p, \tag{11}$$

with probability tending to 1, we say that $\hat{\boldsymbol{g}}_n$ has the sparsity. The maximizer of $l_s(\boldsymbol{g})$ is called an oracle estimator since we use the knowledge of the true model under Assumption S. If an estimator is asymptotically equivalent to such an oracle estimator, we say that the estimator has the oracle property.

Theorem 3 *Suppose that the assumptions in Theorem 2 and some additional assumptions hold and let* $\{d_n\}$ *be a sequence of positive numbers satisfying* $d_n \to \infty$ *and* $\lambda_n/(d_n r_{pn}) \to \infty$ *.*
(i) *With probability tending to 1, any local maximizer* $\hat{\boldsymbol{g}}_n$ *of* $Q_p(\boldsymbol{g})$ *over* $\mathbf{G}_0$ *such that* $\|\hat{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} \leq d_n r_{pn}$ *satisfies (11).*
(ii) *With probability tending to 1, the local maximizer in* (i) *is the unique maximizer of* $l_s(\boldsymbol{g})$ *and satisfies*

$$\sum_{j=1}^{s} \|\hat{g}_{nj} - g_{0j}\|_{L_2}^2 = \mathcal{O}(r_{sn}^2).$$

Now we state the properties of the adaptive group Lasso estimator.

Theorem 4 *Suppose that the assumptions in Theorem 1 and Assumption S hold and that* $\lambda_n' \sqrt{s} \max_{1 \leq j \leq s} w_j / r_{pn} = \mathcal{O}_p(1)$ *. Then with probability tending to 1, there is a unique maximizer* $\bar{\boldsymbol{g}}_n = (\bar{g}_{n1}, \ldots, \bar{g}_{np})^\top$ *of* $\bar{Q}_p(\boldsymbol{g})$ *over* $\mathbf{G}_0$ *and we have*

$$\|\bar{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} = \mathcal{O}_p(r_{pn}).$$

Theorem 5 *Suppose that the assumptions in Theorem 4 and some additional assumptions hold and that* $\lambda_n' \min_{s < j \leq p} w_j/(r_{pn}) \to \infty$ *in probability. Then with probability tending to 1, the unique maximizer* $\bar{\boldsymbol{g}}_n$ *has the sparsity and is equal to the unique maximizer of* $\bar{Q}_s(\boldsymbol{g})$ *. In addition we have*

$$\sum_{j=1}^{s} \|\bar{g}_{nj} - g_{0j}\|_{L_2}^2 = \mathcal{O}_p(r_{sn}^2).$$

Remark 1 Suppose that the true model is a semi-varying coefficient model. Then we can detect the semi-varying coefficient model with probability tending one by modifying the estimators in the following way. We decompose g_j of $\boldsymbol{g} = (g_1, \ldots, g_p) \in \mathbf{G}_0$, by

$$g_j(t) = \frac{1}{\tau} \int_0^\tau g_j(s) ds + \left\{ g_j(t) - \frac{1}{\tau} \int_0^\tau g_j(s) ds \right\} = g_{aj} + g_{bj}(t)$$

and $\|g_j\|_{L_2}^2 = |g_{aj}|^2 + \|g_{bj}\|_{L_2}^2$. Then we define $Q_p'(\boldsymbol{g})$ and $\bar{Q}_p'(\boldsymbol{g})$ by

$$Q_p'(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \sum_{j=1}^{p} \{ p_{\lambda_n}(|g_{aj}|) + p_{\lambda_n}(\|g_{bj}\|_{L_2}) \}$$

and

$$\bar{Q}_p'(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \lambda_n' \sum_{j=1}^{p} (w_{1j}|g_{aj}| + w_{2j}\|g_{bj}\|_{L_2}),$$

where w_{1j} and w_{2j} are weights.

References

[1] Bradic, J., Fan, J. and Jiang, J. (2011) "Regularization for Cox's proportional hazards model with NP-dimensionality," *Ann. Statist.*, 39, 3092-3120.

[2] Du, P., Ma, S. and Liang, H. (2010) "Penalized variable selection procedure for Cox models with semiparametric relative risk," *Ann. Statist.*, 38, 2092-2117.

[3] Fan, J. and Li, R. (2001) "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, 96, 1348-1360.

[4] Honda, T. and Härdle, W. K. (2012) "Variable selection in Cox regression models with varying coefficients," SFB 649 Discussion Paper 2012-061, Humboldt-Universität zu Berlin.

[5] Hu, Y. and Lian, H. (2013) "Variable selection in a partially linear proportional hazards model with a diverging dimensionality," *Statist. Probab. Letters*, 83, 61-69.

[6] Huang, J. Z., Kooperberg, C., Stone, C. J. and Truong, Y. K. (2000) "Functional ANOVA modeling for proportional hazards regression," *Ann. Statist.*, 28, 961-999.

[7] Leng, C. and Zhang, H. H. (2006) "The L_2 rate of convergence for event history regression with time-dependent covariates," *J. Nonparametric Statist.*, 18, 417-429.

[8] Lian, H., Li, J. and Hu, Y. (2013) "Shrinkage variable selection and estimation in proportional hazards models with additive structure and high dimensionality," *Comp. Statist. Data Anal.*, 63, 99-112.

[9] Tibshirani, R. J. (1996) "Regression shrinkage and selection via the Lasso," *J. R. Statist. Soc.* B, 58, 267-288.

[10] Yan, J. and Huang, J. (2012) "Model selection for Cox models with time-varying Coefficients," *Biometrics*, 68, 419-428.

[11] Zhang, H. H. and Lu, W. (2007) "Adaptive Lasso for Cox's proportional hazards model," *Boimetrika*, 94, 691-703.

[12] Zhang, S., Wang, L. and Lian, H. (2012) "Estimation by polynomial splines with variable selection in additive Cox models," forthcoming in *Statistics*, DOI:10.1080/02331888.2012.748770.

[13] Zou, H. (2006) "The adaptive Lasso and its oracle properties," *J. Amer. Statist. Assoc.*, 101, 1418-1429.