Statistics and Probability Letters 83 (2013) 61–69

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

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

Variable selection in a partially linear proportional hazards model with a diverging dimensionality

Yuao Hu, Heng Lian *

Division of Mathematical Sciences, SPMS, Nanyang Technological University Singapore, 637371, Singapore

ARTICLE INFO

Article history:
Received 25 May 2012
Received in revised form 27 August 2012
Accepted 28 August 2012
Available online 5 September 2012

Keywords:
Akaike information criterion (AIC)
Bayesian information criterion (BIC)
Cross-validation
Partial likelihood
SCAD

ABSTRACT

We consider the problem of simultaneous variable selection and estimation in partially linear proportional hazards models when the number of covariates in the linear part diverges with the sample size. We apply the smoothly clipped absolute deviation (SCAD) penalty to select the significant covariates in the linear part. Some simulations and a real data set are presented.

1. Introduction

Nowadays, more and more researchers are concerned with data analysis tasks in which a large number of predictors/features are used. This is due to the fact that, in a study where there are limited previous experiences, it is hard to identify a small number of predictors such that it is believed that only these variables contribute to the response of interest. Thus a large number of predictors suspected to be related to responses need to be collected to avoid model misspecification. On the other hand, due to the large number of predictors collected, it is desirable to select a small number of predictors that are relevant for prediction.

Variable selection is an important research topic in modern statistics. With a large number of predictors available to include into the model, many of them may not be relevant for prediction, and inclusion of these only hurts estimation performance. Recently, there has been considerable interest in investigating the variable selection problem for parametric and nonparametric models. Traditional variable selection methods such as stepwise regression and best subset selection suffer from instability, as argued in Breiman (1996), which is part of the reason why a penalization-based method (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006; Huang et al., 2008; Zhu and Zhu, 2009; Wang et al., 2012) has gained popularity in recent years. This motivates us to develop a penalization-based approach for simultaneous variable selection and estimation in partially linear Cox models.

Many studies on penalization-based variable selection for semiparametric models, including partially linear models, varying-coefficient models and additive models, can be found in the recent literature. For example, Xie and Huang (2009) studied variable selection in partially linear models, Wang et al. (2008), Wang and Xia (2009), Wei et al. (2011) and Lian (2012) investigated varying-coefficient models, Xue (2009), Meier et al. (2009) and Huang et al. (2010) investigated additive models, and Li and Liang (2008), Liu et al. (2011) and Wang et al. (2011) considered partially linear varying-coefficient models and partially linear additive models for variable selection on the linear part.

* Corresponding author.
E-mail address: henglian@ntu.edu.sg (H. Lian).

0167-7152/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2012.08.024

Parametric proportional hazards models, or Cox models, are probably the oldest and the most popular regression tools for studying the relationships between multiple covariates and censored event times in survival analysis. This class of models assume that the hazard function is related to the covariates by

$$\lambda(t|X) = \lambda_0(t)\exp\{X^T\beta_0\}, \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function and X is the covariate vector. The popularity of the proportional hazards models can be at least partially attributed to the fact that the unknown baseline function elegantly disappears from the estimating function when partial likelihood is used, making estimation and inferences much easier.

In this work, we consider proportional hazards models with a semiparametric risk that has a partially linear structure. More specifically, we assume that the hazard function is given by

$$\lambda(t|W, X) = \lambda_0(t)\exp\{\phi_0(W) + X^T\beta_0\}, \tag{2}$$

where now the model contains both the nonparametric component $\phi_0(W)$ and the parametric component $X^T\beta_0$, W is q-dimensional and X is p-dimensional. Although $q > 1$ is possible, in practice only $q = 1$ is popular to avoid the curse of dimensionality. Thus we will only consider $q = 1$ here. This model combines the flexibility of nonparametric modeling with parsimony and the easy interpretability of parametric modeling. In particular, it avoids the curse of dimensionality of a purely nonparametric model (O'Sullivan, 1993; Fan et al., 1997).

For parametric proportional hazards models, penalized variable selection has been considered in Tibshirani (1997), Fan and Li (2002) and Zhang and Lu (2007). More recently, Bradic et al. (2011) has extended it to the case where the number of predictors can be much larger than the sample size, a "large p small n" situation which has attracted much attention in recent years. Leng and Zhang (2007) considered nonparametric model selection in Cox models in reproducing kernel Hilbert spaces. On the other hand, Du et al. (2010) has considered the additive partially linear models in Cox regression with penalized variable selection to identify significant covariates in the linear part. This work however only considered the case where the dimension of the covariates is fixed. This result thus does not apply when the number of covariates diverges. It appears that no systematic studies exist for semiparametric proportional hazards models with a diverging number of covariates.

In this work, we consider penalized variable selection for semiparametric Cox models with a diverging number of parameters. However, although the number of covariates collected for statistical analysis is large, only a subset of covariates are important in predicting the event times. Such an assumption is often reasonable with high dimensional data. More technically, we will assume that $p = o(n^{1/2})$ or $p = o(n^{1/3})$. We use the SCAD method to achieve both goals of variable selection and estimation simultaneously. For the nonparametric component, the component function is approximated and estimated using polynomial splines with the computationally favorable B-spline basis, which allows reasonable approximation of smooth functions with just a small number of basis functions.

There are some studies in the literature of variable selection that concern models with ultra-high dimensionality (Fan and Lv, 2011; Huang et al., 2010; Lian, 2012). In these studies, it is allowed that p diverges at an exponential rate with sample size. However, these studies are concerned with either parametric models or uncensored observations. We will leave the ultra-high dimensional case for future study.

Our study is conceptually most related to Xie and Huang (2009) and we intend to extend their results for partially linear models to censored data based on proportional hazards models. The rest of the article is organized as follows. In the next section, we define the SCAD-penalized estimator in the partially linear proportional hazards model. Computational aspects are also discussed and several criteria for tuning parameter selection are proposed. We investigate its asymptotic theoretical properties in Section 3, including consistency in both estimation and variable selection, as well as the asymptotic normality of the linear coefficients. The finite sample behavior of the SCAD-penalized estimator is illustrated in Section 4. Finally, the Supplementary Material online contains all the technical proofs.

2. The SCAD-penalized estimator

Let T^e and T^c be the event time and the censoring time respectively, where the hazard function of T^e is given by (2). Assume that T^e and T^c are independent, given the covariates. The true nonparametric functions and parameters will be denoted using a subscript 0. The observable random variables are (T, Δ, W, X) where $T = \min\{T^e, T^c\}$ and $\Delta = I\{T^e \le T^c\}$ ($I\{\cdot\}$ is the indicator function), and $W \in R$ and $X = (X_1, \ldots, X_p)^T \in R^p$ are the covariates in the nonparametric part and the parametric part respectively. Note that ϕ_0 is identifiable only up to a constant and thus we assume $E\Delta\phi_0(W) = 0$. We make n i.i.d. observations $(T_i, \Delta_i, W_i, X_i)$.

We use polynomial splines to approximate the nonparametric component. Let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K'} < 1 = \tau_{K'+1}$ be a partition of $[0, 1]$ into subintervals $[\tau_k, \tau_{k+1})$, $k = 0, \ldots, K'$ with K' internal knots. A polynomial spline of order r is a function whose restriction to each subinterval is a polynomial of degree $r - 1$ and globally $r - 2$ times continuously differentiable on $[0, 1]$. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\{\tilde{B}_1(x), \ldots, \tilde{B}_{\tilde{K}}(x)\}$ with $\tilde{K} = K' + r$. Because of the centering constraint $E\Delta\phi_0(W) = 0$, we instead focus on the subspace of spline functions $S^0 := \{s : s = \sum_{k=1}^{\tilde{K}} a_k\tilde{B}_k(x), \sum_{i=1}^n \Delta_i s(W_i) = 0\}$ with basis $\{B_k(x) = \sqrt{K}(\tilde{B}_k(x) - \sum_{i=1}^n \Delta_i\tilde{B}_k(W_i)/n)$, $k = 1, \ldots, K = \tilde{K} - 1\}$ (the subspace is $K = \tilde{K} - 1$-dimensional due to the empirical version of the constraint).