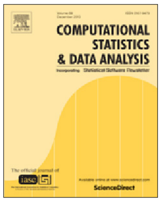
Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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



Identification of local sparsity and variable selection for varying coefficient additive hazards models

Lianqiang Qu [a], Xinyuan Song [b], Liuquan Sun [c]

[a] *School of Mathematics and Statistics, Central China Normal University, Wuhan, Hubei, 430079, PR China*
[b] *Department of Statistics, The Chinese University of Hong Kong, Hong Kong*
[c] *Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, PR China*

HIGHLIGHTS

- We propose a method to conduct variable selection and identify local sparse effects.
- We provide a smooth estimate of the varying coefficient in the non-null subregions.
- We propose an efficient computational algorithm to solve the minimization problem.

ARTICLE INFO

Article history:
Received 29 August 2017
Received in revised form 22 December 2017
Accepted 9 April 2018
Available online xxxx

Keywords:
Additive hazards models
Group penalty
Kernel smoothing
Local sparsity
Oracle property
Varying coefficients

ABSTRACT

Varying coefficient models have numerous applications in a wide scope of scientific areas. Existing methods in varying coefficient models have mainly focused on estimation and variable selection. Besides selecting relevant predictors and estimating their effects, identifying the subregions in which varying coefficients are zero is important to deeply understand the local sparse feature of the functional effects of significant predictors. In this article, we propose a novel method to simultaneously conduct variable selection and identify the local sparsity of significant predictors in the context of varying coefficient additive hazards models. This method combines kernel estimation procedure and the idea of group penalty. The asymptotic properties of the resulting estimators are established. Simulation studies demonstrate that the proposed method can effectively select important predictors and simultaneously identify the null regions of varying coefficients. An application to a nursing home data set is presented.

1. Introduction

Varying coefficient models have become popular statistical tools in many applications (Hastie and Tibshirani, 1993; Hoover et al., 1998; Fan and Zhang, 1999; Huang et al., 2002; Sun and Wu, 2005). For example, in the cross-sectional data analysis, some covariate effects may have nonlinear interactions with an exposure variable. In the longitudinal setting, the effects of covariates on the outcome of interest may change dynamically over time. Varying coefficient models provide a nice graphical summary of temporal dynamics of covariate effects, and also can reveal deep insights into functional and complex interactive effects of covariates, thereby greatly enhancing model capability and flexibility.

In the literature of survival analysis, extension from conventional survival models to their varying coefficient variants is also common (Zucker and Karr, 1990; Murphy and Sen, 1991; Marzec and Marzec, 1997; Martinussen et al., 2002; Tian et al., 2005; Chen and Tong, 2010; Chen et al., 2012). For example, Fan et al. (2006) considered a varying coefficient Cox

E-mail addresses: qulianq@amss.ac.cn (L. Qu), xysong@sta.cuhk.edu.hk (X. Song), slq@amt.ac.cn (L. Sun).

https://doi.org/10.1016/j.csda.2018.04.003
0167-9473/© 2018 Elsevier B.V. All rights reserved.

model, wherein the covariate effects vary with an exposure variable. As an important alternative of Cox-type models, additive hazards models and their varying coefficient variants have also attracted significant attention in the past years. For example, Lin and Ying (1994) introduced a semiparametric additive hazards model with constant covariate effects. McKeague and Sasieni (1994) proposed a partly Aalen's model, which allows certain covariate effects to vary with time and the rest to be time-invariant. Yin et al. (2008) extended the additive hazards model to a partially linear varying coefficient model, in which some covariate effects vary with an exposure variable and the others are constant. However, the existing works on varying coefficient survival models have mainly focused on estimation.

In substantive study, identifying potential risk factors and selecting a plausible but parsimonious model is also of scientific interest. Besides traditional variable selection methods, such as Akaike Information Criterion, Bayesian Information Criterion (BIC), and stepwise selection procedure, penalized variable selection methods have received much attention in the past decades. The commonly used penalized methods include Lasso (Tibshirani, 1996), bridge regression (Fu, 1998; Knight and Fu, 2000), SCAD (Fan and Li, 2001), adaptive Lasso (Zou, 2006), group lasso (Yuan and Lin, 2006) and MCP (Zhang, 2010). These variable selection procedures have also been extensively studied in the presence of censored data (Tibshirani, 1997; Fan and Li, 2002; Zhang and Lu, 2007; Liu and Zeng, 2013). In addition, some penalized variable selection methods have been developed for varying coefficient models (Fan et al., 2006; Wang et al., 2008; Wang and Xia, 2009; Yan and Huang, 2012; Xiao et al., 2016). For example, Fan et al. (2006) extended the SCAD to the Cox model with coefficients varying with an exposure variable. Xiao et al. (2016) proposed a kernel group nonnegative garrote method for automatic model structure selection and coefficient estimation in the time-varying coefficient Cox model.

Although the aforementioned methods are useful for selecting relevant variables, they cannot identify the local sparsity of relevant predictors. The local sparsity of a predictor means that the varying effect of the predictor is zero only over part of its domain. When local sparsity exists, detecting zero and nonzero subregions of varying effects can provide additional insights into a nice graphical summary of the predictors. Thus, it is desirable for a regression procedure to be capable of producing estimates that are exactly zero over certain regions and have varying effects over the remaining regions. For the functional linear model, James et al. (2009) proposed a so-called "FLiRTI" approach to determine zero and nonzero subregions of the coefficient function, Zhou et al. (2013) suggested a two-stage method to simultaneously identify the null region of the coefficient function and provided the estimation procedure on the non-null region, and Lin et al. (2017) developed one-stage procedure to simultaneously identify null subregions of the coefficient function and produced a smooth estimate in non-null subregions.

In some situations, both variable selection and detection of local sparsity are important. However, the existing methods were developed solely on variable selection or on the detection of local sparsity. This limitation motivates us to consider a new method that is able to select significant predictors and identify their local sparsity simultaneously. Moreover, it is very desirable to develop a locally sparse estimator that can work with the prevalent kernel smoothing methods in a very natural way. In this article, we develop an effective method to eliminate unimportant variables and identify null subregions for important variables in the context of a varying coefficient additive hazards model. The proposed method combines the kernel smoothing technique and the group adaptive penalty into a single optimization objective function. Compared with the usual variable selection procedure, the group adaptive penalty allows us not only to perform variable selection, but also to detect the null subregions of varying coefficients. Moreover, by solving the optimization problem, we can identify the null subregions and provide a smooth estimate of the varying coefficient in the non-null subregions. Thus, the proposal inherits many nice statistical properties from both the kernel smooth estimation and nonconcave penalized methods. We study the asymptotic properties of the resulting estimators such as the sparsity and oracle property. In order to implement our method in practice, we propose an efficient computational algorithm to solve the minimization problem involved in the estimation procedure. To our knowledge, this research is the first attempt to simultaneously accomplish the two important tasks. As we will demonstrate via simulation studies, the proposed method performs as well as the oracle one in terms of both variable selection and detection of local sparsity.

The rest of this article is organized as follows. Section 2 describes the proposed model and estimation procedure, and the associated asymptotic properties are established as well. Section 3 discusses technical issues about tuning parameter selection and variance estimation. The empirical performance of the proposed method is demonstrated via simulation studies in Section 4. Section 5 presents an application to a nursing home data. Section 6 concludes the article with discussion. Proofs and technical details are provided in the Appendix.

2. Methodology

2.1. Notation and model

Let T be the failure time and C be the censoring time. Define $X = \min(T, C)$, and $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Let $Z(t)$ be the $p \times 1$ vector of external time-dependent covariates (Kalbfleisch and Prentice, 2002), and V be an exposure variable allowed to be time dependent. Assume that T and C are independent given $Z(\cdot)$ and V. The observed data consist of n independent and identically distributed replicates of $(X, \Delta, Z(\cdot), V)$, denoted by $\{(X_i, \Delta_i, Z_i(\cdot), V_i), i = 1, \ldots, n\}$.

The varying coefficient additive hazards model specifies that, given $Z(t)$ and V, the hazard function of T takes the form

$$h(t|Z(t), V) = h_0(t, V) + \beta(V)^T Z(t), \tag{1}$$

Please cite this article in press as: Qu L., et al., Identification of local sparsity and variable selection for varying coefficient additive hazards models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.04.003.