Journal of Multivariate Analysis 127 (2014) 56–71

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

# Sufficient dimension reduction on marginal regression for gaps of recurrent events[☆]

Xiaobing Zhao[a,*], Xian Zhou[b,*]

[a] School of Mathematics and Statistics, Zhejiang University of Finance and Economics, Hangzhou, Zhejiang Province, China
[b] Department of Applied Finance and Actuarial Studies, Macquarie University, Sydney, NSW, Australia

ARTICLE INFO

*Article history:*
Received 8 November 2012
Available online 4 February 2014

*AMS 2000 subject classifications:*
62G99
62H99
62N01

*Keywords:*
Semiparametric transformation
Gap time
Recurrent event
High-dimensional covariates
Sliced regression
Sufficient dimension reduction

ABSTRACT

A semiparametric linear transformation of gap time is proposed to model recurrent event data with high-dimensional covariates and informative censoring. It is derived from a proportional hazards model for the conditional intensity function of a renewal process. To overcome the difficulty arising from high-dimensional covariates, we develop a modified sliced regression for censored data and use a sufficient dimension reduction procedure to transform them to a lower dimensional space. Simulation studies are performed to confirm and evaluate the theoretical findings, and to compare the proposed method with existing methods in the literature. An example of application on a set of medical data is demonstrated as well. The proposed model together with the dimension reduction method offers an effective alternative for the analysis of recurrent event with high-dimensional covariates and informative censoring.

## 1. Introduction

Recurrent event data arise frequently in longitudinal medical studies. Examples include repeated hospitalizations, the occurrences of opportunistic infections in HIV-infected patients, the appearances of new tumors in patients with superficial bladder cancer, and medical costs. They are also encountered in many other important areas such as criminology (recidivism), demography (repeated migrations), consumer service (warranty claims), and actuarial studies (insurance claims), among others.

To analyze recurrent event data, various approaches have been proposed in the literature with focuses on different variables of interest, such as the number of events during the observation time, the event times, and the inter-event (gap) times. These approaches include the *conditional regression models* by Prentice et al. [40] and Andersen and Gill [2], and the *semiparametric hazards model* by Chang and Wang [5]. To overcome the lack of robustness in the conditional regression models, Wei et al. [48] proposed the *marginal hazards model*, and Pepe and Cai [39] developed an intermediate alternative between the conditional regression models and the marginal hazards model. Compared with the intensity and hazard-based models, the marginal mean or rate models are often preferred by practitioners since the mean number of event has a more direct interpretation for identifying risk factors. The literature on the marginal mean/rate models can be found in [29,32], among others. More references on the recurrent event process can be found in the book of Cook and Lawless [11] and their reviews.

[☆] This work was partially supported by the NSFC under Grant No. 11271317, Zhejiang Provincial Natural Science Foundation under Grant No. LY12A01017, Zhejiang Provincial Planning Projects of Philosophy and Social Science under Grant No. 12JCJJ17YB, and the Institute of Actuaries of Australia Research Grant.

[*] Corresponding authors.
*E-mail addresses:* maxbzhao@hotmail.com (X. Zhao), xian.zhou@mq.edu.au (X. Zhou).

http://dx.doi.org/10.1016/j.jmva.2014.01.008
0047-259X/© 2014 Elsevier Inc. All rights reserved.

The existing models for recurrent event data have largely assumed low-dimensional covariates (if any). In recent years, however, various high-dimensional covariates have been found important for statistical analysis and prediction. One example is to use the gene expression profiles to predict various clinical phenotypes including tumor recurrence, drug response and survival time—particularly patient survival time and time to cancer recurrence [41]. This methodology, which *jointly* models the survival time and the high-dimensional covariates, is a ground breaking advance in biomedical and genomic research. There are, however, a number of challenges that hinder its development, such as a huge number of potential predictors and censored survival time [31]. For such data, a dimension reduction strategy is necessary to transform the high-dimensional covariates to a low-dimensional space as the first step of data analysis.

For survival analysis model with high-dimensional covariates, the existing technique of dimensional reduction can be classified roughly into two categories. The first is the likelihood-based dimension reduction, including various penalized Cox regressions. A nice review for this technique is provided by Witten and Tibshirani [49]; see also, Engler and Li [18], Van Wieringen et al. [45], Gui and Li [15], and Cook and Forzani [10]. The second is the regression-based dimension reduction. All the techniques in this category are the applications of sufficient dimension reduction, such as sliced inverse regression (SIR) proposed by Li [30]; see also Li and Li [31], Lu and Li [37] and Xia et al. [52]. A nice review of various applications of SIR can be found in [47], in which a new method free of the linearity condition in [30] is proposed. In either category, however, the dimension reduction has not been applied in survival analysis of recurrent event data with high-dimensional covariates and censored failure time.

In recurrent event data with multiple (repeated) survival times, the intra-subject dependence resulted from informative censoring cannot be ignored in statistical inference. In this paper, we study a gap time model of recurrent event data with high-dimensional covariates and censored observations. Motivated by a dimension reduction technique free of the linearity condition proposed by Wang and Xia [47], we develop a modified sliced regression (MSR) method that can accommodate censored failure times and transform the high-dimensional covariates into a low-dimensional space. Local regression is employed to estimate an unknown increasing function of gap times. By using a semiparametric approach combined with local regression, our model and estimation procedure provide more flexibility to fit the data and resemble the practical situation. The proposed model and method differ from that of Xia et al. [52] in that it is based on the sliced regression. In addition, the model and method in this paper do not need the application of the inverse regression considered in [37]; hence the linearity condition can be relaxed.

In Section 2, we review existing models and then specify our models. Sections 3 and 4 develop the sufficient dimension reduction method, including the motivation, the algorithm, and a cross-validation (CV) criterion for the central subspace dimension determination. Section 5 establishes the asymptotic properties of the estimators. In Section 6, we report some simulation results in five examples to evaluate the performance of our proposed method and compare it with a number of existing methods in the literature. Section 7 demonstrates an example of application, and concluding remarks are discussed in Section 8.

## 2. Model specification

Consider a longitudinal study that involves $n$ independent subjects, each may experience recurrences of an event of interest. Throughout this paper, $N_i(t) = \int_0^t dN_i(s)ds$ represents the number of events in time interval $[0, t]$ for subject $i$, $i = 1, 2, \ldots, n$, where $dN_i(s)$ denotes the number of events in $[s, s + ds]$. Subject $i$ is observed to experience $m_i$ events at times $T_{i,1}, \ldots, T_{i,m_i}$ over the interval $[0, C_i]$, where $C_i$ represents the censoring time of subject $i$. We assume that $C_i$ is independent of the recurrent process $\{N_i(t); t \geq 0\}$ (cf. [4]). For subject $i$ with a $p \times 1$ vector $Z_i$ of covariates, the observation time of the $k$th event is given by $T_{i,k} \wedge C_i$, where $a \wedge b = \min(a, b)$, and its event indicator is $\delta_{i,k} = I(T_{i,k} \leq C_i)$. The inter-event (gap) times are denoted by $Y_{i,k} = T_{i,j} - T_{i,j-1}$ with $T_{i,0} = 0$.

We are interested in the gap times between recurrent events and the effects of covariates on these gap times. Let $\mathcal{N}_i = \{Y_{i,j} : j = 1, 2, \ldots, \}$ and $\{\mathcal{N}_i, Z_i, C_i\}$, $i = 1, \ldots, n$, be $n$ independent and identically distributed (i.i.d.) replicates of $\{\mathcal{N}, Z, C\}$ under the assumption that the recurrent event process is a renewal process [4]. For this model, Huang and Chen [27] and Schaubel and Cai [42] discussed regression analysis of recurrent gap times under the proportional hazards model, and Sun et al. [43] studied a recurrent gap time model with additive hazards by using estimating equation similar to that of Huang and Chen [27].

We here propose a semiparametric linear transformation model as follows, which arises from a general form of the proportional hazards model introduced by Cox [14] and is an extension of the model to fit the recurrent gap times by Lu [36]:

$$H(Y_{i,k}) = -\phi(Z_i^\top \beta_1, \ldots, Z_i^\top \beta_d) + \varepsilon_{i,k}, \quad k = 1, 2, \ldots, m_i;\ i = 1, 2, \ldots, n, \tag{2.1}$$

where $\beta_j = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{pj})^\top \in R^p$ is a $p$-dimensional vector of unknown regression parameters (covariate coefficients), $j = 1, 2, \ldots, d$, $\phi(\cdot, \ldots, \cdot)$ is an unknown function on $R^p$, $H(\cdot)$ is an unknown increasing function, and $\varepsilon_{i,k}$ is the error term with a continuous distribution that is independent of censoring variable $C_i$ and covariate vector $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})^\top \in R^p$. In addition, $\{(\varepsilon_{i,1}, \varepsilon_{i,2}, \ldots, ), i = 1, 2, \ldots, n\}$ are i.i.d. random vectors. For each $i$ and any $k \neq j$, the error terms $\varepsilon_{i,k}$ and $\varepsilon_{i,j}$ are possibly correlated, but assumed to be exchangeable with a common specified marginal distribution.

For simplicity, we denote $B = (\beta_1, \beta_2, \ldots, \beta_d) \in R^{p\times d}$ and $\phi(u_i) = \phi(u_{i1}, u_{i2}, \ldots, u_{id})$ with $u_{ij} = Z_i^\top \beta_j$. In this paper we consider the case of orthonormal $\beta_1, \beta_2, \ldots, \beta_d$ such that $B^\top B = I_d$ with $d \geq 1$. Then model (2.1) can be rewritten as

$$H(Y_{i,k}) = -\phi(Z_i^\top B) + \varepsilon_{i,k}, \quad k = 1, 2, \ldots, m_i;\ i = 1, 2, \ldots, n. \tag{2.2}$$