مجله‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار
دوره‌ی ۱۶، شماره‌ی ۲، پاییز و زمستان ۱۳۹۸، صص ۳۷۹–۳۹۶

J. Statist. Res. Iran **16** (2019): 379–396
DOI: 10.52547/jsri.16.2.379
DOR: 20.1001.1.25385771.2020.16.2.2.2

A Method for Analyzing Censored Survival Data with Application to Coronary Heart Disease

Azam Rastin[†], Mohammad Reza Faridrohani[†] and Davoud Khalili[‡]

[†] Shahid Beheshti University
[‡] Sahahid Beheshti University of Medical Sciences

Received: 2020/23/01 Approved: 2021/16/02

Abstract. An objective of analyzing survival data via regression is to develop a predictive model given predictors. However, due to the censoring in response variables and the high dimensionality of predictors, information needed for an appropriate model specification is often inadequate. We propose a method for an integrated study of survival time and predictors. At first, variable selection methods are employed for finding the correct subset of predictors with significantly higher probability. This is based on the Lasso approach. Then, the dimension of the predictors is further reduced using sufficient dimension reduction methods. This is based on the Sliced inverse regression for censored data (DSIRII) . In particular we use the popular Cox proportional hazards model to build a predictive model for survival data. An application to Coronary heart disease (CHD) data from the Tehran Lipid and Glucose (TGLS) study further illustrates the usefulness of the work.

Keywords. Censored data; sufficient dimension reduction; central subspace; sliced inverse regression; variable selection; corronary heart disease.

MSC 2010: 62N01.

[*] Corresponding author

1 Introduction

The rapidly increasing burden of non communicable diseases is a key determinant of global public health (Reddy and Yusuf, 1998). Coronary heart disease (CHD) is one of the most common causes of morbidity and mortality in different communities (Castelli, 1984; Keil, 2005). CHD continues to be the number one cause of death in most Northern European, North American and other industrialized Caucasian societies. By the age of 60, every fifth man and one in 17 women have some form of this disease (Castelli, 1984). In the United States of America, CHD is the leading cause of death in adults, accounting for approximately one-third of all deaths in people over the age of 35 years (Thom et al., 1998). The age-adjusted prevalence of CHD in Iran is 12.7 % (Nabipour et al., 2007).

Recently, regression analysis of high-dimensional data has prevailed in many scientific fields, while the high dimensionality often poses challenges to classical regression techniques. In many applications such as biomedical, the analysis is further complicated by possible censoring in response variables. Common methods for handling censored responses include Cox's proportional hazards (PH) model (Cox, 1972), proportional odds (PO) model (Bennett, 1983), accelerated failure time model (Cox and Oakes, 1984), among many others. Most of those methods require specification of some models. However, knowledge for selecting an appropriate model is often inadequate prior to the analysis, and model specification and diagnosis can become more elusive when the predictor dimension is high. Sufficient dimension reduction (Cook, 1998), requires no model specification, it retains full regression information, and it provides a usually small set of composite variables upon which subsequent model formulation and prediction can be based. For more about sufficient dimension reduction and their methodologies, read Yoo (2016a, 2016b). Sufficient dimension reduction in survival analysis has been addressed in Li and Lu (2008) and Nadkarni et al. (2011). Some papers consider regularized estimation, for instance, Lu and Li (2011).

In prediction models, the belief that ''the more the variables, the better the performance'' is no longer acceptable; thus, during the past decades, application of variable (feature) selection techniques has been fast gaining popularity in the field of data mining (Liu and Yu, 2005). In particular, the high-dimensional aspect of clinical data sets has attracted much attention to variable selection (Saeys et al., 2007); however, the methods developed have not yet been applied broadly in clinical prediction models

(Bagherzadeh-Khiabani et al., 2015). The commonly used stepwise selection procedures (Efroymson, 1960), which can be computationally expensive and ignore stochastic errors in the variable selection process. In this article, Lasso method (Tibshirani, 1997) are proposed to handle these kinds of problems.

The aim of this paper is to propose a different procedure to identify risk factors of Coronary heart disease to predict individuals' survival probabilities. This method found projection directions in life time space without imposing any structural assumption (such as Cox) between survival time and predictors. We can use these projection directions to predict survival probabilities.

2 Theory and Methodology

2.1 Variable Selection

Tibshirani (1997) proposed a norm penalized regression procedure called Lasso, which performs regression shrinkage and variable selection simultaneously. By the Cox's proportional hazards model, the hazard function for subject i with covariates X_i is given as

$$\lambda_i(t) = \lambda_0(t) exp(X_i^T \theta), \tag{1}$$

where λ_0(t) is an unknown baseline hazard function at time t and θ is a vector of unknown parameters. The estimation of θ is commonly based on the partial log-likelihood (Cox , 1970), namely

$$l(\theta) = \sum_{i=1}^{N} \delta_i \left\{ X_i^T \theta - \log(\sum\nolimits_{j \in R_j} \exp(X_j^T \theta)) \right\} \tag{2}$$

where R_j is the risk set at time t_j (i.e. the set of subjects who have not yet experienced an event) and δ_i is an indicator for whether or not subject i is censored. Tibshirani (1997) proposed to estimate the parameters in (2) under the L_1 constraint:

$$\hat{\theta} = arg\min\{-l(\theta) + \lambda ||\theta||_1\}, \tag{3}$$

where $||\ ||_1$ denotes the L_1 norm and λ is a positive number. The tuning parameter λ can be determined, for example, by generalized cross-validation (GCV) as explained by Tibshirani (1997). The minimization problem (3) is

appropriate tool for dimension reduction of covariates. Because of the nature of this constraint, it shrinks coefficients and produces some coefficients that are exactly zero. Tibshirani (1997) indicated the lasso outperforms the full Cox model and stepwise models by shrinking the coefficients almost all of the way to zero.

2.2 Sufficient Dimension Reduction: SIR with complete data

To reduce the dimension of regression problems, Sliced inverse regression (SIR; Li, 1991) makes it possible to determine linear combinations of explanatory variables related to a response variable Y in general regression models. Li's framework for dimension reduction in regression begins with the following formulation:

$$Y = g(\beta_1^{'}X, ..., \beta_k^{'}X, \epsilon) \tag{4}$$

where Y is associated with the p-dimensional predictors X only through the linear combinations $\beta_j^{'}X$ and ϵ is a random error term independent of X. No assumption has been made about the functional form of the unknown link function g or the distribution of ϵ. The space spanned by $\beta_1^{'}X, ..., \beta_k^{'}X$, which is a subspace spanned by all the columns in X (dimension reduction subspace (Li, 1991; Cook, 1996), is called the effective dimension reduction (e.d.r.) space and any vector in this space is referred to as an e.d.r. direction. When K is small which is often the case in real applications, the original regression problem (data) can be effectively reduced by projecting X along these effective directions.

Li (2007) showed that many moment based suffcient dimension reduction methods can be formulated as a generalized eigenvalue problem in the following form

$$\mathbf{M}_n\delta_{ni} = \lambda_{ni}\mathbf{N}_n\delta_{ni}, \quad for\ i = 1, ..., p,$$

where $\mathbf{M}_n \geq 0$ is a method-specific symmetric kernel matrix, $\mathbf{N}_n > 0$ is symmetric, often taking the form of the sample covariance matrix $\mathbf{\Sigma}_n$ of x; $\delta_{n_1}, ..., \delta_{n_p}$ are eigenvectors such that $\delta^{'}{}_{ni}\mathbf{N}_n\delta_{nj} = 1$ if $i = j$ and 0 if $i \neq j$ and $\lambda_{n_1} \geq ... \geq \lambda_{n_p}$ are the corresponding eigenvalues. We use the subscript "n" to indicate that $\mathbf{\Sigma}_n, \mathbf{M}_n, \mathbf{N}_n$ and λ_{ni} are the sample versions of the corresponding population analogs $\mathbf{\Sigma}, \mathbf{M}, \mathbf{N}$ and λ_i. Under certain conditions that are usually imposed only on the marginal distribution of x, the first d eigenvectors $\{\delta_{n_1}, ..., \delta_{n_k}\}$, which correspond to the nonzero eigenvalues

Table 1. The generalized eigenvalue formulations for principle component analysis (PCA), sliced inverse regression (SIR and SIRII) and sliced average variance estimation (SAVE) methods.

Method	M	N
PCA	$\mathbf{\Sigma}$	$\mathbf{I_p}$
SIRI	$Cov(E(Z\|Y))$	$\mathbf{\Sigma}$
SIRII	$E[Cov(Z\|Y) - E(Cov(Z\|Y))]^2$	$\mathbf{\Sigma}$
SAVE	$E[Cov(Z\|Y) - I_p]^2$	$\mathbf{\Sigma}$

$\lambda_{n_1} > ... > \lambda_{n_k}$ form a consistent estimator of a basis for the central subspace. Letting $Z = \Sigma^{-1/2}\{X - E(X)\}$. Many commonly used moment based SDR methods are listed in Tabe (1) with the population versions of $\mathbf{M}_n$ and $\mathbf{N}_n$.

The number of significant projections, denoted by k, can be determined by an asymptotic Chi-square test, which sequentially tests the hypotheses $k = m$ versus $k > m$ for $m = 0, ..., p-1$, where p is the number of covariates (Li, 1991).

2.3 Modification of SIR to Censored Survival Data

SIR is a powerful method due to its simplicity. However, SIR cannot be applied directly to censored survival data. The paper by Li et al.(1999) extended the original SIR to censored data (DSIRI). Also, Rastin and Faridrohani extended (2020) SIRII of Li (1991) to the setting which allows for censoring in the data (DSIRII). We will briefly introduce DSIRI and DSIRII in this section and discuss the related issues of applying DSIRII to CHD data.

Denote T as the true unobservable survival time, C as the censoring time, $Y = min(T, C)$ as the observed survival time and and $\delta = I(T \leq C)$ be the censoring indicator. For identifiability, we assume conditionally independence between T and C given X, denoted by $T \perp C \mid X$. Li et al.(1999) introduced a two-step procedure. The basic idea is to estimate a weight function $w(t^{'}, t, X)$ for $t > t^{'}$. Specifically, $w(t^{'}, t, X) = S_0(t \mid X)/S_0(t^{'} \mid X)$, and $S_0(t \mid X) = P(T \geq t \mid X)$= conditional survival function for T, given X. To estimate $w(t^{'}, t, X)$, we may use the smoothing techniques suggested by Li et al.(1999). The kernel smoothing approach (Beran, 1981) was considered here.

Effective dimension reduction (e.d.r.) space of life time can be identified given this weight function. Because kernel estimation is more efficient in low-

dimension spaces, one initial dimension reduction step is required. Assume that the true unobservable survival time T and the censoring time C have dimension reduction structures given by

$$T = g(\beta_1^{'}X, ..., \beta_k^{'}X, \epsilon)$$

and

$$C = h(\eta_1^{'}X, ..., \eta_c^{'}X, \epsilon^{'})$$

respectively. It is noted, though, that the bivariate response (Y, δ) is observable and is a function of (T, C). Operationally this amounts to the slicing of the bivariate response (Y, δ), that is, partitioning Y within each subsample where $\delta = 0$ and $\delta = 1$, respectively. Then the joint e.d.r. space of the true unobservable survival time and censoring time will be obtained by taking the eigenvalue decomposition on the between-slice covariance matrix with respect to the covariance matrix of X. The remaining steps remain the same as a usual SIR (SIRII). The leading eigenvectors will serve as the projection directions. This procedure is called double slicing in Li et al.(1999). Mainly due to its simplicity, double slicing SIR ,DSIRI (DSIRII), has received wide applications; see Li and Li(2004) for analysis of microarray gene expression data with censored phenotypes, and Li et al.(2007) for analysis of Tobit models in economics data.

Predictive model can be build based on the reduced data. For example, survival models, e.g., Cox proportional hazard model, can be fitted with the projected directions as the explanatory variables first and then the fitted model can be used to predict the survival probability of individuals.

3 Simulation Results: Comparison with Other Methods

We first compare DSIRII-lasso with the DSIRII estimator of Rastin, Rohani (2020) and the DSIR estimator of Li et al. (1999). The true survival time is generated as follows:

$$T = \Phi[5\epsilon\{\exp(3\beta' X) + 1\} - 2] \tag{5}$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution, ϵ follows an exponential distribution with parameter 1, and X follows a standard normal distribution independent with ϵ. We consider

d=1, p=7 and set the true parameter value to be β =(1,0,-1, 0,1, 0,-1)$'$. We generate the censoring time C as

$$C = \Phi(2X_2 + 2X_3) + U$$

where U denotes a random variable uniformly distributed on (0,c), where c is a constant controlling the proportion of censoring.

Proposed method yields smaller difference between the estimated projection matrix $\hat{P} \equiv \hat{\beta}(\hat{\beta}'\hat{\beta})^{-1}\hat{\beta}'$ and the true projection matrix $P \equiv \beta(\beta'\beta)^{-1}\beta'$, in that both the mean and variance of the largest singular value of $\hat{P} - P$ are much smaller based on the proposed method than based on DSIRI and DSIRII. In figure (1) the box plots of the three different methods are shown to provide a quick visual inspection. From these results, we can see that the method we proposed generally performs better.

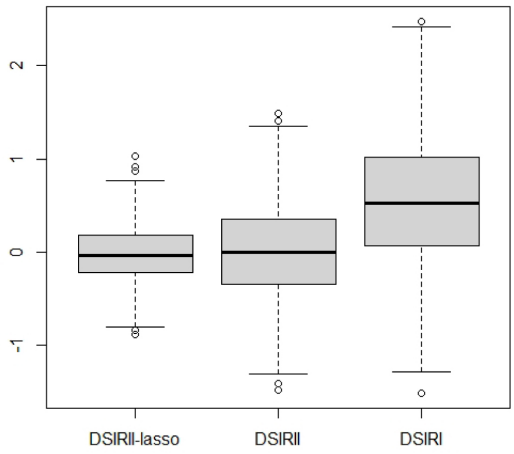

Figure 1. Boxplot of methods of our study under the 40% of censoring.

We use an affine invariant criterion proposed by Li (1991):

$$R^2(\hat{\beta}) = \max_{\beta \in \mathrm{B}} \frac{(\beta'\beta)^2}{\hat{\beta}'\hat{\beta}.\beta'\beta}$$

where B is the true dimension-reduction space. The performance is then compared with DSIRI and DSIRII. The results of $R^2(\hat{\beta})$ for measuring the performance in estimation are summarized in Table (2) as the value of Cen-

Table 2. Mean and standard deviation of $R^2(\hat{\beta})$ for model (5) in 100 replicates

censored rate	sample size	DSIRII-lasso	DSIRII	DSIRI
20%	100	0.973(0.08)	0.966(0.01)	0.255(0.25)
	200	0.987(0.07)	0.989(0.04)	0.283(0.24)
	300	0.990(0.01)	0.988(0.02)	0.271(0.23)
	500	0.991(0.01)	0.990(0.05)	0.323(0.25)
40%	100	0.952(0.06)	0.712(0.03)	0.222(0.20)
	200	0.963(0.04)	0.764(0.04)	0.247(0.21)
	300	0.989(0.01)	0.856(0.07)	0.293(0.27)
	500	0.990(0.01)	0.860(0.06)	0.312(0.25)
70%	100	0.939(0.07)	0.385(0.3)	0.211(0.20)
	200	0.941(0.07)	0.410(0.27)	0.221(0.29)
	300	0.948(0.06)	0.415(0.31)	0.252(0.19)
	500	0.989(0.01)	0.848(0.07)	0.278(0.28)

sored rate varies.

4 Application

4.1 Description of the Data Set

The Tehran Lipid and Glucose Study is a longitudinal study, the first phase of which was a cross-sectional or baseline examination survey from 1999 to 2001. It was designed to estimate the prevalence of CVD risk factors in a representative sample of an Iranian urban population selected by random cluster sampling from district no. 13 of Tehran. Details of the rationale and design have been published elsewhere Azizi et al.(2000). The data were collected by means of interviews, with completion of a questionnaire for demographic data and CHD history, physical examination for blood pressure, pulse rate, ECG and anthropometrical measures and laboratory measurements for lipid and glucose profiles. The data collected were stored directly in a computer database Christiansen et al. (1990). The details of data collection have already been described Azizi et al.(2002).

There were 5417 individuals aged $\geqslant$ 30 years who completed the baseline

Table 3. Clinical and biochemical characteristics of the Tehran Lipid and Glucose Study population $\geq$ 30 years.

Risk Factors	Men	Women	P-value
	Percent%	**Percent%**	
Smoking	20.00	3.6	0.000
HTNdrug	0.04	0.08	0.000
LPDdrug	0.015	0.03	0.000
	Mean(SD)	**Mean(SD)**	
Age (years)	46.86(12.71)	45.16(11.22)	0.000
SBP(mm Hg)	120.19(17.98)	119.58(18.85)	0.026
DBP (mm Hg)	78.30(10.99)	78.75(10.41)	0.004
Waist (cm)	89.32(10.54)	89.403(11.78)	0.000
BMI (kg/m^2)	26.01(3.84)	28.29(4.65)	0.000
FBS(mg/dL)	91.46(9.56)	90.29(9.82)	0.399
Chol(mg/dL)	208.39(41.60)	217.61(45.49)	0.000
TGs (mg/dL)	184.48(11.94)	164.42(96.44)	0.000
HDL(mg/dL)	38.41(9.42)	44.88(10.96)	0.000
BS2hrs (mg/dL)	104.91(31.57)	113.36(28.22)	0.000

survey and whose data were used for the current study. 332 out of the 5417 individuals had CHD before the study ended, which yields 332 (216 male and 116 female) real survival times and a heavy censoring rate of 95%. The variables in the data set include case number, the number of days between registration and the earlier of death or study analysis time, censoring indicator (1=CHD, 2=censored), age in years, Sex (0=male, 1=female), blood pressure lowering drugs (HTNdrug) (0=no, 1=yes), LPDdrug (0=no, 1=yes), Smoking (0=non smoker, 1=past smoker, 2=current smoker), cholesterol (Chol), Body mass index (BMI), Systolic blood pressure (SBP), Diastolic blood pressure (DBP), triglycerides (TGs), HDL cholesterol (HDL), Waist, blood sugar two hours after you eat 75 grams of glucose (Bs2hrs) and Fasting blood sugar (FBS).

Characteristics of the TLGS have been summarized in Table (3). Statistical analyses were done using the SPSS statistical software package, and data were presented as mean and standard deviation. All analyses were performed separately for males and females in all age strata. The degree of significance of differences between means was calculated using Student's t-test. P-values less than 0.05 were considered to be statistically significant.

4.2 Coronary Heart Disease Data Analysis based on DSIRII-lasso

Following the steps described in the Method section, we first choose the correct subset of predictors with significantly higher probability by variable selection. We applied methods of selecting covariates to go into the model. This was based on the full Cox model and stepwise model and Lasso approach. Estimated coefficients and their standard errors are shown in Table (4). For the stepwise procedure, we use the function stepAIC in R with the default P-values to enter and remove of 0.05 and 0.10 respectively. As we can see, the full Cox model gives the most conservative solution, choosing five variables, whereas stepwise chooses three more variables, "SBP," "BMI" and "HTNdrug." Lasso, on the other hand, chooses the variables "DBP" and "FBS" but not the variable "HTNdrug," suggesting that "DBP" and "FBS" are somewhat more important than "HTNdrug." Then, Variables of Age, Chol, Waist, HDL ,FBS, SBP, DBP, BMI and Smoking are the most important predictive factors of CHD in men using Lasso procedure. We again compare three variable selection algorithms in women. In Table (5) we list and compare solutions found by different techniques, including the full Cox model and stepwise models, and Lasso. The stepwise procedure gave a model with six variables, all having large Z-scores. The resulting model from the lasso looks similar to the stepwise model, with most of the effects shrunken towards zero. While the stepwise procedure often infates the Z scores of chosen variables relative to the full model fit, the lasso seems to shrink them towards zero. Variables of Age, Chol, Waist, BS2hrs, SBP, TGs and HTNdrug are the most important predictive factors of CHD in women using Lasso procedure.

The next step is to perform DSIRII on the selected variables. Rastin et al. (2019) considered an analysis of CHD data via DSIRI. The χ^2 test shows that the first two projection direction is significant ($p = 0.0000$ and $p = 0.0023$, respectively). Therefore, we have reduced the survival time space to two dimension. The basis estimates (β_1 and β_2) of the corresponding covariates are given in Table (6). Fitting a Cox proportional hazards regression model, we also list the estimates (β_0) in Table (6). We can see that basis estimates from both approaches have higher coefficients for age, smoking, sbp, dbp and bmi but sbp contributes less to the linear combination using our proposed procedure. Now we redo the analysis in women. We perform a similar procedure to the one described above. The dimension test indicate that the

Table 4. Result of variable selection for coronary heart disease in the Tehran Lipid and Glucose Study population ≥ 30 years (men).

Variables			**Full**			**Stepwise**			**Lasso**
	SE	Z-score	**Cofficient**	SE	Z-score	**Cofficient**	SE	Z-score	**Cofficient**
Age	0.006	7.378	0.049	0.006	**7.48**	0.047	0.29	**2.12**	0.049
SBP	0.005	1.697	0.008	0.004	**2.86**	0.011	0.06	**0.98**	0.009
DBP	0.008	0.792	0.006	-	-	-	0.08	**0.51**	0.005
Waist	0.015	2.217	0.034	0.015	**2.16**	0.032	0.12	**1.09**	0.026
BMI	0.041	-1.511	-0.062	0.041	**-1.48**	-0.061	0.20	**-2.01**	-0.043
FBS	0.007	-1.061	0.049	-0.008	-	-	0.08	**-0.22**	-0.007
Chol	0.001	2.998	0.005	0.002	**3.75**	0.006	0.001	**0.14**	0.005
TGs	0.000	0.653	0.000	-	-	-	0.000	0.000	0.000
HDL	0.008	-2.512	-0.022	0.008	**-2.97**	-0.023	0.11	**-1.56**	-0.021
BS2hrs	0.002	-0.272	0.000	-	-	-	0.000	0.000	0.000
Smoking	0.081	3.700	0.301	0.080	**3.61**	0.290	0.32	**3.83**	0.229
HTNdrug	0.245	1.737	0.427	0.243	**1.61**	0.392	0.000	0.000	0.000
LPDdrug	0.468	-0.634	-0.296	-	-	-	0.000	0.000	0.000

Table 5. Result of variable selection for coronary heart disease in the Tehran Lipid and Glucose Study population ≥ 30 years (women).

Variables			**Full**			**Stepwise**			**Lasso**
	SE	Z-score	**Cofficient**	SE	Z-score	**Cofficient**	SE	Z-score	**Cofficient**
Age	0.011	6.049	0.066	0.009	**6.97**	0.065	0.801	**2.619**	0.047
SBP	0.006	0.309	0.002	-	-	-	0.083	**1.021**	0.003
DBP	0.012	0.567	0.006	-	-	-	0.000	0.000	0.000
Waist	0.014	1.117	0.015	0.008	**2.55**	0.0 22	0.010	**1.180**	0.002
BMI	0.034	0.464	0.015	-	**-8**	-	0.000	0.000	0.000
FBS	0.010	-1.334	0.013	-	-	-	0.000	0.000	0.000
Chol	0.002	1.738	0.003	0.002	**2.59**	0.005	0.002	**0.971**	0.001
TGs	0.001	0.832	0.001	-	-	-	0.000	0.000	0.000
HDL	0.010	-1.059	-0.011	0.009	**-1.49**	-0.013	0.000	0.000	0.000
BS2hrs	0.003	1.746	0.006	0.003	**1.54**	0.005	0.122	**0.998**	0.004
Smoking	0.211	1.083	0.228	-	-	-	0.000	0.000	0.000
HTNdrug	0.227	3.183	0.724	0.216	**3.36**	0.72512	1.96	**3.291**	0.553
LPDdrug	0.382	-0.6787	-0.301	-	-	-	0.000	0.000	0.000

Table 6. Estimates of dimension reduction subspace directions for coronary heart disease in the Tehran Lipid and Glucose Study population ≥ 30 years (men) along with estimates using Cox regression.

Risk Factors	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Age	0.149	**-0.152**	**0.148**
SBP	0.009	-0.026	**-0.060**
Waist	0.033	-0.029	-0.035
Chol	0.006	-0.005	-0.001
BMI	-0.061	**0.086**	**-0.076**
Smoking	0.291	**-0.283**	**0.249**
DBP	0.006	0.010	**0.101**
HDL	-0.023	0.018	0.008
FBS	-0.008	0.006	0.007

Table 7. Estimates of dimension reduction subspace directions for coronary heart disease in the Tehran Lipid and Glucose Study population ≥ 30 years (women) along with estimates using Cox regression.

Risk Factors	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Age	3.460	**-1.963**	**2.318**
Waist	1.904	-0.781	**-5.011**
BS2hrs	0.250	-0.100	0.633
Chol	0.750	-0.557	-0.509
HTNdrug	0.651	**-1.904**	**-2.301**
SBP	0.659	**-0.890**	0.153
TGs	0.307	-0.08	**1.098**

dimension reduction subspace dimension is two yet again. From Table (7), we find that some covariates such as age, waist, sbp, tg and HTNdrug have high coefficients but tg and sbp contribute less to the linear combination using our proposed procedure.

In Figure 2, $\beta_1^T X$ and $\beta_2^T X$ are plotted against the martingale residuals from the "intercept only" proportional hazards model to explore their relationship with the survival time. From Figure 2, the second dimension reduction subspace direction is approximately linearly related to the residuals of the "intercept only" proportional hazards model while the first dimension reduction subspace subspace direction is apparently not. It thus indicates that the function forms of two directions are different and the effects of covariates on the hazard function through two directions exhibit different

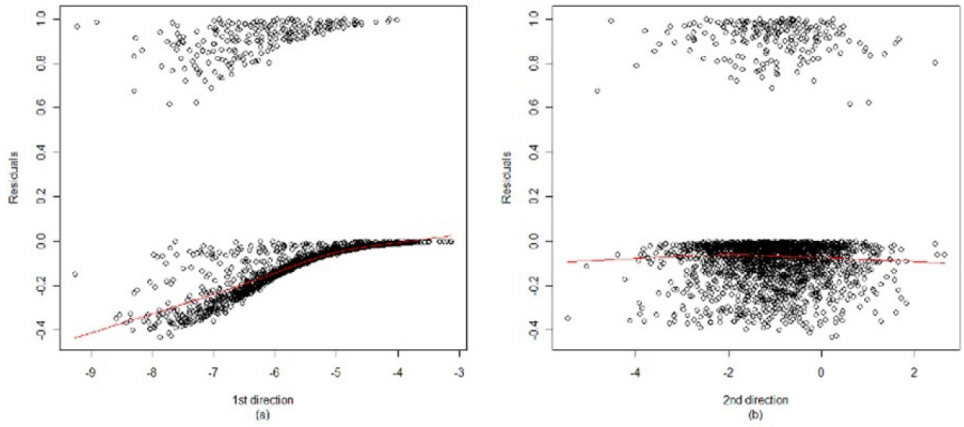

Figure 2. Residual plots for two central subspace directions for coronary heart disease in the Tehran Lipid and Glucose Study population≥ 30 years in men.

patterns. The correlations of $\beta_1^T X$ and $\beta_2^T X$ with $\beta_0^T X$ are - 0.970 and 0.043, respectively. Since the linear component can be well estimated by the proportional hazards model, it again validates that the first and second dimension reduction subspace directions correspond to nonlinear and linear components in the residuals the"intercept only" proportional hazards model respectively.

In women, We again see that the correlations of $\beta_1^T X$ and $\beta_2^T X$ with $\beta_0^T X$ are - 0.915 and 0.248, respectively. That is, there is clear nonlinear relationship between the covariates and the response which cannot be detected directly by a Cox hazards proportional model.

We applied the censorSIR estimation to the CHD data and We next obtained the two censorSIR variates, $\nu_1 = \beta_1^T X$ and $\nu_2 = \beta_2^T X$ where β_1 and β_2 are the censorSIR estimators. To complete the analysis, we next fit the Cox PH model with full quadratic effects, that is $\lambda(t) = \lambda_0(t)exp(\theta_1\nu_1 + \theta_2\nu_2 + \theta_3\nu_1^2 + \theta_4\nu_2^2 + \theta_5\nu_1 \times \nu_1)$, and find the main effect term ν_1 and the quadratic effect ν_1^2 are significant. We then refit the model with these two terms, yielding $\lambda(t) = \lambda_0(t)exp(-3.68\nu_1 - 0.23\nu_1^2)$, with the corresponding P-values for both coefficients less than 1×10^{-4}. Figure 3 shows the Kaplan-Meier estimate of the survival curves for the two risk groups determined by the estimated risk scores. A good separation of the two risk curves is observed, suggesting that the model fits the data reasonably well. The log-rank test shows that the survival rates of these two groups are significantly

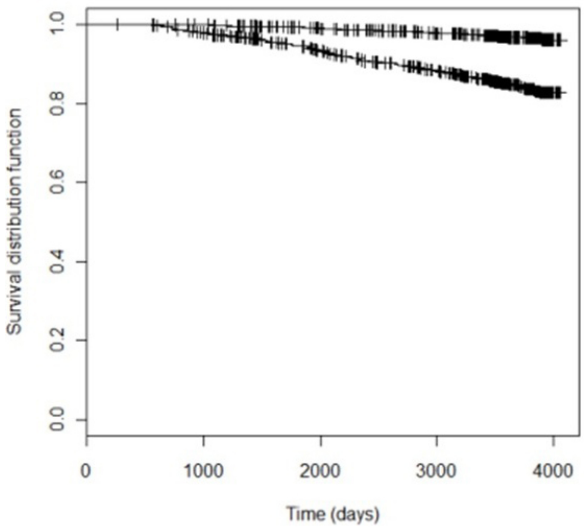

Figure 3. The Kaplan-Meier estimates of survival rates. Survival rates are estimated for the two risk groups in men. The log-rank test comparing the two curves gives a P-value of $p = 2e^{-03}$.

different ($p = 2e^{-03}$, Figure 3).

5 Conclusion

In this paper, we have introduced a method for the analysis of survival time and CHD risk factors. By an appropriate combination of variable selection and dimension reduction, we find a method of identifying risk factors which is effective for survival prediction. The proposed method was applied to the problem of risk factors of CHD separately for men and women based on data from the Tehran Lipid and Glucose Study. Two single projection direction (linear combination of CHD risk factors) was identified that is significantly related to the survival time. Compared with most available methods, our method has two major advantages. Firstly, the subset of predictors which gives the best fit in modeling the time to CHD are finded by employing variable selection methods; Secondly, the dimension of the predictors is further reduced using sufficient dimension reduction methods for censored data. Specifically, no model assumption is required on the relationship between risk

factors and survival time. A clear nonlinear relationship between the covariates and the response also was identified which cannot be detected directly by a Cox hazards proportional model.

It is also possible to use other alternative methods instead of SIR such as SAVE or SIR_α for example. These approaches are based on some properties of the conditional variance of X given Y, see for instance Li (1991) or Shao et al. (2009). Another possible extension is to investigate the case of a multivariate response variable Y : the idea is then to use multivariate SIR approach instead of univariate SIR methods, see for instance Barreda et al. (2007), Saracco (2005) or Lue (2009).

References

Kundu, D. and Dey, A.K. (2009). Estimating the Parameters of the Marshall-Olkin Bivariate Weibull Distribution by EM Algorithm. *Computational Statistics and Data Analysis*, **53**, 956-965.

Azizi, F., Rahmani, M., Emami, H. and Madjid, M. (2000). Tehran Lipid and Glucose Study: Rationale and Design. *CVD prevention* **3**, 242-247.

Azizi, F., Rahmani, M., Emami, H., Mirmiran, P., Hajipour, R. and Madjid, M. (2002). Cardiovascular Risk Factors in an Iranian Urban Population: Tehran Lipid and Glucose Study (Phase1). *Sozial-und pr-ventivmedizin*, **476**, 408-426.

Barreda, L., Gannoun, A. and Saracco, J. (2007). Some Extensions of Multivariate SIR. *Journal of Statistical Computation and Simulation*, **77**, 1-17.

Bennett, S. (1983). Analysis of Survival Data by the Proportional Odds Model. *Statistics in Medicine*, **2**, 273-277.

Beran, R.Z. (1981). Nonparametric Regression with Randomly Censored Survival Data. *Technical Report, Univ. California, Berkeley.*

Castelli W.P. (1984). Epidemiology of Coronary Heart Disease: the Framingham Study. *American journal of medicine*, **76**, 4-12.

Christiansen D.H., Hosking, J.D., Dannenberg, A.L., and Williams, O.D. (1990). Computer-Assisted Data Collection in Multicenter Epidemiologic Epidemiologic Research: the Atherosclerosis Risk in Communities (ARIC) Study. *Controlled Clinical Trials*, **11**, 101-115.

Cook, R.D. (1998). *Regression Graphics*. Wiley, New York.

Cook, R.D. (1996). *Graphics for Regressions with a Binary Response*. Wiley, New York.

Cox, D.R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society*, **Series B 34**, 187-220.

Cox, D.R., and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, New York.

Efroymson, M. (1960). *Multiple Regression Analysis.* In Mathematical Methods for Digital Computers (eds A. Ralston and H. Wilf). Wiley, New York.

Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E.W., and Khalili, D. (2015). A Tutorial on Variable Selection for Clinical Prediction Models: Feature Selection Methods in Data Mining Could Improve the Results. *Journal of Clinical Epidemiology*, **71**, 76-85.

Liu, H., and Yu, L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Trans Knowl Data Eng*, **17**, 491-502.

Saeys, Y., Inza, I., and Larra naga, P.A. (2007). Toward Integrating Feature Selection Algorithms for Classification and Clustering. *Bioinformatics*, **23**, 2507-2517.

Keil, U. (2005). Das Weltweite WHO-MONICAProjekt: Ergebnisse und Ausblic [The Worldwide WHO MONICA Project: Results and Perspectives]. *Gesundheitswesen*, **67**, S38-45.

Li, K.C. (1991). Sliced Inverse Regression for Dimension Reduction (with Discussion). *Journal of the American Statistical Association*, **86**, 316-327.

Li, L. (2007). Sparse Sufficient Dimension Reduction. *Biometrika*, **94**, 603-613.

Li, K.C., Wang, J.L., and Chen, C.H. (1999). Dimension Reduction for Censored Regression Data. *The Annals of Statistics*, **27**, 1-23.

Li, L., and Li, H. (2004). Dimension Reduction Methods for Microarrays with Application to Censored Survival Data. *Bioinformatics*, **20**, 3406-3412.

Li, L., Simonoff, J.S., and Tsai, C.L. (2007). Tobit Model Estimation and Sliced Inverse Regression. *Statistical Modelling*, **7**, 107-123.

Lue, H.H. (2009). Sliced Inverse Regression for Multivariate Response Regression. *J. Statist. Plann. Inference*, **139**, 2656-2664.

Lu, W., and Li, L. (2011). Sufficient Dimension Reduction for Censored Regressions. *Journal of the International Biometric society*, **67**, 513-523.

Li, L., and Lu, W. (2008). Sufficient Dimension Reduction with Missing Predictors. *Journal of the American Statistical Association*, **03(482)**, 822-831.

Nabipour, I., Amiri, M., Imami, S.R., Jahfari, S.M., Shafeiae, E., Nosrati, A., Iranpour, D., and Soltanian, A.R. (2007). The Metabolic Syndrome and Nonfatal Ischemic Heart Disease; a Population-Based Study. *International Journal of Cardiology*, **118**, 48-53.

Nadkarni, N.V., Zhao, Y., and Kosorok, M. (2011). Inverse Regression Estimation for Censored Data. *Journal of the American Statistical Association*, **106**, 178-190.

Reddy K.S., and Yusuf, S. (1998). Emerging Epidemic of Cardiovascular Disease in Developing Countries. *Circulation*, **97**, 596-601.

Rastin, A., and Faridrohani, M. (2020). Modification of Sliced Inverse Regression to Censored Survival Data. *J. of Stat. Sci.*, **13**, 427-440.

Rastin, A., Faridrohani, M., Momenan, A., Eskandari, F., and Khalili, D. (2019). Analysis of Censored Survival Data with Dimension Reduction Methods: Tehran Lipid and Glucose Study. *Andishe*, **23**, 17-25.

Saracco, J. (2005). Asymptotics for Pooled Marginal Slicing Estimator Based on SIR_α. *Journal of Multivariate Analysis*, **96**, 117-135.

Shao, Y., Cook, R.D., and Weisberg, S. (2009). Partial Central Subspace and Sliced Average Variance Estimation. *J. Statist. Plann. Inference*, **139**, 952-961.

Thom, T.J. et al. (1998). *Incidence, Prevalence and Mortality of Cardiovascular Disease in the United States.* Hurst's the heart, 9th ed. McGraw–Hill, New York.

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Stat. Med*, **16**, 385-395.

World Health Organization, Eastern Mediterranean Regional Office (1995). Prevention and control of cardiovascular diseases. *Alexandria: WHO-EMRO:* **24**.

Yoo, J.K. (2016a). Tutorial: Dimension Reduction in Regression with a Notion of Sufficiency. *Communications for Statistical Applications and Methods*, **23**, 93-103.

Yoo, J.K. (2016b). Tutorial: Methodologies for sufficient dimension reduction in regression. *Communications for Statistical Applications and Methods* **23**, 105-117.

Azam Rastin
Department of Statistics,
Shahid Beheshti University,
Tehran, Iran.
email: *rastinstat@gmail.com*

Mohammad Reza Faridrohani
Department of Statistics,
Shahid Beheshti University,
Tehran, Iran.
email: *m_ faridrohani@sbu.ac.ir*

Davoud Khalili
Research Institute for Endocrine Sciences,
Sahahid Beheshti University of Medical Sciences,
Tehran, Iran.
email: *dkhalili@endocrine.ac.ir*