Hindawi Publishing Corporation
Journal of Quality and Reliability Engineering
Volume 2015, Article ID 795154, 9 pages
http://dx.doi.org/10.1155/2015/795154

Research Article

Variable Selection Methods for Right-Censored Time-to-Event Data with High-Dimensional Covariates

Keivan Sadeghzadeh and Nasser Fard

Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115, USA

Correspondence should be addressed to Keivan Sadeghzadeh; k.sadeghzadeh@neu.edu

Received 1 October 2014; Revised 13 April 2015; Accepted 16 April 2015

Academic Editor: Christian Kirchsteiger

Advancement in technology has led to greater accessibility of massive and complex data in many fields such as quality and reliability. The proper management and utilization of valuable data could significantly increase knowledge and reduce cost by preventive actions, whereas erroneous and misinterpreted data could lead to poor inference and decision making. On the other side, it has become more difficult to process the streaming high-dimensional time-to-event data in traditional application approaches, specifically in the presence of censored observations. This paper presents a multipurpose analytic model and practical nonparametric methods to analyze right-censored time-to-event data with high-dimensional covariates. In order to reduce redundant information and to facilitate practical interpretation, variable inefficiency in failure time is determined for the specific field of application. To investigate the performance of the proposed methods, these methods are compared with recent relevant approaches through numerical experiments and simulations.

1. Introduction

Time-to-event data such as failure or survival times have been extensively studied in reliability engineering. By the advent of modern data collection technologies, a huge amount of this type of data includes high-dimensional covariates. This massive amount of data is increasingly accessible from various sources such as transaction-based information, information-sensing devices, remote sensing technologies, machines and logistics statistics, wireless sensor networks, and analytics in quality engineering, manufacturing, service operations, and many other segments. Unlike traditional datasets with few explanatory variables, the analysis of datasets with a high number of variables requires different approaches. In this situation, variable selection techniques could be used to determine a subset of variables that are significantly more valuable to analyze high-dimensional time-to-event datasets. If data is compiled and processed correctly, it can enable informed decision making [1–7].

In many professional areas and activities, as well as manufacturing and services, decision making is increasingly based on the type and size of data, as well as analytic methods, rather than on experience and intuition. It has been suggested that business should discover new ways to collect and use data every day and develop the ability to interpret data to increase the performance of their decision makers [8, 9]. As stated in a broad survey [10], advanced analytics are among the most popular techniques used in high-dimensional and massive data analysis and decision making process. As an analytical approach, decision making is the process of finding the best option from all feasible alternatives [11, 12].

The opportunity for manufacturing and services in the era of data is to analyze their performance to enhance the quality. Quality dimensions of products and services, defined by quality experts [13, 14] or perceived by customers [15], are summarized as performance, availability, reliability, maintainability, durability, serviceability, conformance, warranty, and aesthetics and reputation. Access to valuable data for

sophisticated analytics can substantially improve the management decision making process. In reliability analysis, failure time is determined by the variables contributing to products' failure time. Complex, high-dimensional, and censored time-to-event data provide an excellent chance for manufacturers to reduce costs, improve efficiency, and ultimately improve the quality of their products by detecting failure causes faster [16, 17].

Most of the traditional variable selection methods such as Akaike information criterion (AIC) [18] or Bayesian information criterion (BIC) [19] involve computational algorithms in a class of nondeterministic polynomial-time (NP-hard) and computational cost, making these procedures infeasible. Also, recently developed methods for identifying variable efficiency may operate faster, but the robustness is not consistent. These methods involve different estimation methods and assumptions such as the Cox proportional hazard model [20], accelerated failure time [21], Buckley-James estimator [22], random survival forests [23], additive risk models [24], weighted least squares [25], or classification and regression tree (CART) [26]. Due to the presence of censoring, analyzing time-to-event data with high-dimensional covariates and recognizing efficient covariates in terms of predictive power of survival is more challenging [24, 25]. This study is motivated by the importance of the aforementioned variable selection issue.

The objective of this study is to propose a combinational methodology for the variable reduction via determining variable inefficiency in right-censored high-dimensional time-to-event data. The aim of the proposed logical analytic model as well as methods and algorithms is also to reduce the volume of the failure time data and to identify a set of the most influential variables on failure time. Variable efficiency refers to the effect of a variable on failure or survival time in a right-censored time-to-event dataset with high-dimensional covariates. This paper presents two multipurpose nonparametric methods to analyze the aforementioned class of data.

The concept of time-to-event analysis and the commonly used and relevant data mining tools and techniques are presented in Section 2. The logical model for the transformation of the explanatory variable dataset to reach the logical representation of the original covariate dataset as a sort of binary variables is defined in Section 3, where each variable is represented by a Boolean vector to verify and to prove the sustainability of the transformation. Section 4 presents hybrid nonparametric variable selection methods and algorithms through variable efficiency. The validity of proposed methods is verified by results obtained in comparison to those from well-known methods through simulation patterns and by using different collected and simulated time-to-event datasets. The performance and verification of the proposed methods are presented in Section 5. Concluding remarks, including the advantages of the proposed methods, are discussed in Section 6.

The dynamic change of variables in time-dependent explanatory data streaming is of interest in the complementary level of this research. The computer software used in this research is the MATLAB R2011b programming environment.

2. Basic Definitions

In this section, an applied introduction of time-to-event data analysis and a brief review of prominent data mining tools and techniques relevant to this study are presented.

2.1. Time-to-Event Data Analysis. Time-to-event data analysis methods consider the time until the occurrence of an event. This time can be measured in any unit such as days, weeks, or years with this analysis widely used in reliability engineering. In time-to-event data, subjects are usually followed over a specified time period. The study of time-to-event data focuses on predicting the probability of survival or failure. Examples of time-to-event data are the lifetime of mechanic devices, electronic components, or complex systems [21, 27, 28].

Regression models cannot effectively perform in the presence of the censoring of observations [27, 29]. Censored data occurs when the information about the event time is not complete or missed for any reason. Right censoring occurs when a test subject does not remain under the test for a full test period or until it fails. In this paper, we focus on time-to-event data with this type of censoring.

Available methods to analyze time-to-event data and to find a relationship between survival time and other variables can be categorized in parametric, semiparametric, and nonparametric methods. Parametric survival analysis is based on survival function distributions such as the exponential function. Semiparametric models do not assume knowledge of absolute risk. These models estimate relative risk rather than absolute risk with this assumption called the proportional hazards assumption. In this category, the Cox proportional hazards regression analysis is by far the most popular model for survival data analysis. For moderate-to high-dimensional covariates, it is difficult to apply semiparametric methods [25]. In nonparametric methods which are useful when the underlying distribution of the problem is unknown, statistical assumptions are not required. These methods are commonly used to describe survivorship of a study population or compare two or more study populations. The Kaplan-Meier product limit estimate is a commonly used nonparametric method in estimating the survival function. This estimator has clear advantages since it does not require an approximation of the follow-up time assumption [27, 30].

The probability of the failure time occurring at time t is

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}. \tag{1}$$

In time-to-event or survival analysis, the information on an event status and follow-up time are used to estimate a survival function, $S(t)$, which is defined as the probability that an object survives at least until time t:

$$\begin{aligned} S(t) &= P(\text{an object survives longer than } t) \\ &= P(T > t). \end{aligned} \tag{2}$$

From the definition of the cumulative distribution function (or failure function)

$$S(t) = 1 - P(\text{an object fails before } t) = 1 - P(T \le t) = 1 - F(t). \quad (3)$$

Accordingly, the survival function is calculated by a probability density function as

$$S(t) = \int_t^{\infty} f(u)\, du. \quad (4)$$

In most applications, the survival function is shown as a step function rather than a smooth curve. The nonparametric estimate of $S(t)$ according to the Kaplan-Meier (KM) estimator for distinct ordered event times t_1 to t_n is as follows:

$$\widehat{S}(t) = \prod_{i=1}^{t} \left(1 - \frac{d_i}{n_i}\right), \quad (5)$$

where at each event time t_j there are n_j subjects at risk and d_j is the number of subjects which experienced the event, for example, failed. Let c_i denote the number of subjects censored between t_i and t_{i+1}. Then the likelihood function takes the following form:

$$L = \prod_{i=1}^{t} \left[S(t_{i-1}) - S(t_i)\right]^{d_i} \left[S(t_i)\right]^{c_i}. \quad (6)$$

For the conditional probability of surviving, if we define $\pi_i = S(t_i)/S(t_{i-1})$, then the maximum-likelihood estimation of π_i is as follows:

$$\widehat{\pi}_i = 1 - \frac{d_i}{n_i}. \quad (7)$$

Graphically, the Kaplan-Meier estimate is a step function with discontinuities which increases at observed failure times. It has been shown [31] that the KM estimator is consistent. The completely nonparametric nature of this estimator assures little or no loss in efficiency. A quick review of commonly used data mining tools and techniques in this study is presented next.

2.2. Data Mining Tools and Techniques. To analyze time-to-event data, when the size and dimensions are large, advanced analytics are advantageous. Data reduction techniques are categorized in three main strategies: dimensionality reduction, numerical reduction, and data compression [32, 33]. Dimensionality reduction is the most efficient strategy in the field of large-scale data deals by reducing the number of random variables or attributes in the special circumstances of the problem. Dimensionality reduction methods are mainly wavelet transformations and principal components analysis (PCA) [34, 35]. The transformation and projection of the original data eliminate a subset of the original data in terms of the variables' covariance.

All dimensionality reduction techniques are also classified as feature extraction and feature selection approaches. Feature extraction is defined as transforming the original data into a new lower dimensional space through some functional mapping such as PCA and SVD [36, 37]. Most unsupervised dimensionality reduction techniques are closely related to PCA which is one of the oldest and most well-known multivariate analysis techniques, but this technique is not applicable to large complex datasets [38]. Feature selection is denoted by selecting a subset of the original data (features) without a transformation in order to filter out irrelevant or redundant features, such as filter methods, wrapper methods, and embedded methods [39, 40]. The next section presents a proposed analytic logical model for the transformation of the explanatory variable dataset of a time-to-event data.

3. Proposed Analytic Model

The direction for developing an analytic model and analyzing datasets depends on the type and size of the data. A multipurpose, flexible, and innovative model for a type of right-censored time-to-event data with a large number of variables when the correlation between variables is complicated or unknown provides the motivation to find an applicable solution for this type of data. For such data, we propose a model to simplify the original covariate dataset into a logical dataset by transformation lemma. In order to select the most significant variables in terms of efficiency, variable reduction methods and clustering algorithms are proposed in Section 4. The analytic model [41] and its following methods and algorithms are potentially applicable solutions for many problems in a vast area of science and technology.

The original right-censored high-dimensional time-to-event dataset may include any type of explanatory data as binary, continuous, categorical, or ordinal data. The concept of this proposed analytic model is that many variables are even binary or interchangeable with a binary variable such as dichotomous and Bernoulli variables. Also, the interpretation of a binary variable is simple, understandable, and comprehensible. In addition, the model is appropriate for fast and low-cost calculation which makes the time-dependent analysis with data streaming possible.

The random variables Y and C represent the time-to-event and censoring time, respectively. Time-to-event data is represented by $(\mathbf{T}, \mathbf{\Delta}, \mathbf{U})$ where $\mathbf{T} = \min(Y, C)$ and the censoring indicator $\mathbf{\Delta} = 1$ if the event occurred, for instance failure is observed, otherwise 0. The observed covariate $\mathbf{U}$ represents a set of variables. Let denote any observations by (t_i, δ_i, u_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, p$. It is assumed that the hazard at time t only depends on the survivals at time t which assures the independent right censoring assumption [21].

In order to simplify the original complex high-dimensional time-to-event dataset, we propose a transformed logical model. Based on the concept of this model, for any n-by-p dataset matrix $\mathbf{U}$, there are n independent observations and p variables. Each array of p variables vectors will take only two possible values, canonically 0 and 1. Therefore, it is required to define a Bernoulli criterion to split all arrays as a set of binary outcomes and reach a logical dataset. Transforming numerical attributes to binary variables has been well studied [42]. In order to construct

the abovementioned transformed logical time-to-event dataset as a simplified and applied representation of the original one, we define w_j as an initial Bernoulli criterion:

$$w_j = \frac{\max\{u_{ij}\} + \min\{u_{ij}\}}{2} \quad i = 1, \ldots, n, \ j = 1, \ldots, p. \tag{8}$$

For any array u_{ij} in the n-by-p dataset matrix $\mathbf{U} = [u_{ij}]$, assign a substituting array v_{ij} as

$$v_{ij} = \begin{cases} 0, & u_{ij} < w_j \\ 1, & u_{ij} \geq w_j. \end{cases} \tag{9}$$

Note that for each of the variable vectors $\mathbf{v}_{i.}$ the criterion w_j could be defined by an expert using experimental or historical data as well (8). The proposed model assumes any array with a value of 1 as desired for an expert and 0 otherwise. In other words, $v_{ij} = 0$ represent the lack of the jth variable in the ith observation. In this fashion, only desired variables will be considered in each variable vector. The transformed dataset is used in the proposed methods and algorithms.

Therefore, the result of the transformation is an n-by-p dataset matrix $\mathbf{V} = [v_{np}]$ which will be used in the following methods and algorithms. Also, we define the time-to-event vector $\mathbf{T} = [t_n]$ including all observed failure times and $\mathbf{S} = [s_n]$ as the survival function. The proposed logical model validation and verification of the robustness were presented comprehensively in [41, 43]. The logical model initially could be satisfied by the proper design of the data collection process based on Boolean logic to generate binary attributes.

4. Proposed Methods and Heuristic Algorithms

In order to design appropriate methods and algorithms, a test is performed on the efficiency of a cluster of variables as a subdataset of the complete time-to-event covariates dataset. The test is done by comparing this subdataset with the complete transformed logical dataset. A key assumption in this approach is that the variable which is completely inefficient solely can provide a significant performance improvement when engaged with others, and two variables that are inefficient by themselves can be efficient together [40]. By expanding these assumptions over the presence of a large number of covariates with a complexity in correlation, the efficient subset of variables does not differ meaningfully from the effect of the whole body of variables on the time-to-event outcome. Therefore, comparing through nonparametric test, any subset from the complete dataset could be determined as an efficient or inefficient selection. Based on these assumptions, we design two methods and hybrid algorithms for the proposed analytic model for selecting inefficient variables in right-censored time-to-event datasets with high-dimensional covariates.

We use the Kaplan-Meier estimator in this study to estimate and graph survival probabilities as a function of time. In addition, a nonparametric method is used to test a null hypothesis of whether two samples are drawn from the same distribution as compared to a given alternative hypothesis. Among many nonparametric tests for comparing survival functions for the aforementioned propose, we use a log-rank test in our methods as best fit for the comparison of two nonparametric distributions. This test is the most commonly used one for a typical study under different models for the relationship between the groups [27, 30].

$\mathbf{V}$ is constructed by k observation vectors that correspond to each of the variables $\mathbf{D} = [d_{kp}]$ as a k-by-p matrix, a selected subset of $\mathbf{V}$. k is defined as the number of observations in any subset of $\mathbf{V}$, where $k \leq n$. We also define vector $\mathbf{R}$ as follows:

$$r_i = \begin{cases} 0, & \sum di. = 0 \\ t_i, & \sum di. \geq 0 \end{cases} \quad i = 1, \ldots, n. \tag{10}$$

Vector $\mathbf{R}$ is constructed by all nonzero arrays r. The preliminary step for the highest efficiency in proposed methods is to cluster the variables based on the correlation coefficient matrix of the original dataset $\mathbf{M} = [m_{ij}]$ and choose a representative variable from each highly correlated cluster and then eliminate the other variables from the dataset. Let m_{ij} denote the covariance of variables i and j. The correlation coefficient matrix is defined as follows:

$$m_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \left(v_{ik} - \overline{v}_i\right)\left(v_{jk} - \overline{v}_j\right) \quad i = 1, \ldots, p, \ j = 1, \ldots, p. \tag{11}$$

v_{ik} and $\overline{v}_i$ represent the values of variable i in observation k and the mean of variable i and the second parenthesis defined similarly for variable j.

Applying this lemma, for instance, to any given dataset, determines three highly correlated variables from $\mathbf{M}$. Only one of them is selected randomly and the other two are eliminated from the dataset. The outcome of this process ensures that the remaining variables for applying methods and heuristic algorithms are not highly correlated.

4.1. Method I: Nonparametric Test Score Variable Clustering. The log-rank test score variable selection method is applied for selecting a subset of the best and the worst variables in terms of efficiency. The nonparametric test score (NTS) method is a variable clustering technique which selects a set of size k variables from the transformed logical dataset $\mathbf{V}$ and calculates the score of each variable in two levels. The first level is to determine the priority of the variable efficiency via the scores. The scores are obtained from the frequency of each variable in the rejected subsets from comparison with the original time-to-even vector $\mathbf{T}$. We code this level of calculation with letter F. The second level rates the variables by the cumulative score of each variable from comparisons of selected subsets of all nonparametric test results with the original time-to-even vector $\mathbf{T}$. This level acts as a searching procedure to detect the less efficient variables. The code which this level is denoted by is the letter C. The randomization

Step 1. for $i = 1$ to q do
Step 2. Compose the dataset $\mathbf{D}_i$ for variable set i in $\mathbf{\Psi}$ including variables ψ_{ij} where $j = 1$ to k
Step 3. Calculate $\mathbf{R}_i$ over the dataset $\mathbf{D}_i$
Step 4. Compare $\mathbf{T}$ and $\mathbf{R}_i$ with log-rank test
Step 5. Save the test score for variables in subset i as $\psi_{i(k+1)}$
Step 6. end for
Step 7. Eliminate rows of $\mathbf{\Psi}$ where the array $k + 1$ of the row has a value of more than 0.05, and call this new set $\mathbf{\Psi}^*$
Step 8. Count the contribution (presence) of each variable h based on its identification number in the $\mathbf{\Psi}^*$ first k columns and save as γ_h
Step 9. Return $\Gamma = [\gamma_p]$ as the variable efficiency vector

Algorithm 1: NTS (F) algorithm.

Step 1. for $i = 1$ to l do
Step 2. Compose the dataset $\mathbf{D}_i$ for variable set i in $\mathbf{\Psi}$ including variables ψ_{ij} where $j = 1$ to k
Step 3. Calculate $\mathbf{R}_i$ over the dataset $\mathbf{D}_i$
Step 4. Compare $\mathbf{T}$ and $\mathbf{R}_i$ with log-rank test
Step 5. Save the test score for variables in subset i as $\psi_{i(k+1)}$
Step 6. end for
Step 7. Assume $\mathbf{\Omega} = [\omega_p]$ as the reverse variable efficiency vector where initially each array as the cumulative contribution score corresponding to a variable is zero
Step 8. for $i = 1$ to l do
Step 9. for $j = 1$ to k do
Step 10. Add the value of $\psi_{i(k+1)}$ to the cumulative contribution score ω_p of the variable i based on its identification number $= \psi_{ij}$
end for
Step 11. end for
Step 12. Return $\mathbf{\Omega} = [\omega_p]$ as the variable inefficiency vector

Algorithm 2: NTS (C) algorithm.

(RN) algorithm randomly chooses a defined l subset of k from the $\mathbf{V}$ transformed logical dataset of p variable. We define a randomization dataset matrix $\mathbf{\Psi} = [\psi_{lk}]$ where each row is formed by k variable identification numbers in any selected subsets for overall l subsets. The heuristic algorithm of NTS method level F is as in Algorithm 1.

The heuristic algorithm of NTS method level C is presented is Algorithm 2. The experiment results for these algorithms are followed in Section 5.

4.2. Method II: Splitting Semigreedy Clustering Algorithm. The splitting semigreedy (SSG) method to select an inefficient variable subset is proposed. And a clustering procedure through a randomly splitting approach to select the best local subset according to a defined criterion is incorporated. The concept of this method is inspired by the semigreedy heuristic [44, 45] and tabu search [46]. The criterion of this search is similar to *Method I* which is to collect the most inefficient variable subset via a log-rank test score. At each of l trials, all p variables from the transformed logical dataset $\mathbf{V}$ are randomly clustered into subsets of size k variables, where one cluster possibly contains less than k variables and the number of clusters is equal to $\lceil p/k \rceil$. To calculate the score summation for each variable over all trials, we define a randomization dataset matrix $\mathbf{\Xi} = [\xi_{lk}]$ where each row is formed by k variable identification numbers in any selected subsets for all l selected subsets. The heuristic algorithm of SSG is presented in Algorithm 3. Comprehensive experimental results for the validation of the proposed methods by comparison with similar methods are presented next.

5. Experimental Results and Analysis

To evaluate the performance of the proposed methods, the designed algorithms under different types of datasets including collected and simulated data are investigated. In order to obtain an estimation of desired number of variables in any selected subset of NTS (F and C levels) and SSG methods for calculations in hybrid algorithms (Algorithms 1, 2, and 3), we use principal component analysis (PCA) scree plot criterion [47]. The criterion for this evaluation is based on the eigenvalue of components in a dataset. The cut-off in the scree plot is interpreted as a set of significant eigenvalues among all components which leads to determining k, number of efficient variables in a given dataset (see Figure 1). We apply this technique on the original dataset to determine k.

Step 1. for $i = 1$ to l do
Step 2. Split the data into equally sized subsets
Step 3. Compose the dataset **D** for each subset
Step 4. Calculate **R** over the **D** for each subset
Step 5. Compare **T** and **R** with log-rank test and save the test score for each subset one by one
Step 6. Select a subset with the highest test score
Step 7. Save the test score for variables in the selected subset as $\xi_{i(k+1)}$
Step 8. end for
Step 9. Assume $\boldsymbol{\Theta} = [\theta_p]$ as the reverse variable efficiency vector where initially each array as the cumulative contribution score corresponding to a variable is zero
Step 10. for $i = 1$ to l do
Step 11. for $j = 1$ to k do
Step 12. Add the value of $\xi_{i(k+1)}$ to the cumulative contribution score θ_p of the variable i based on its identification number = ξ_{ij}
end for
Step 13. end for
Step 14. Return $\boldsymbol{\Theta} = [\theta_p]$ as the variable inefficiency vector

Algorithm 3: SSG algorithm.

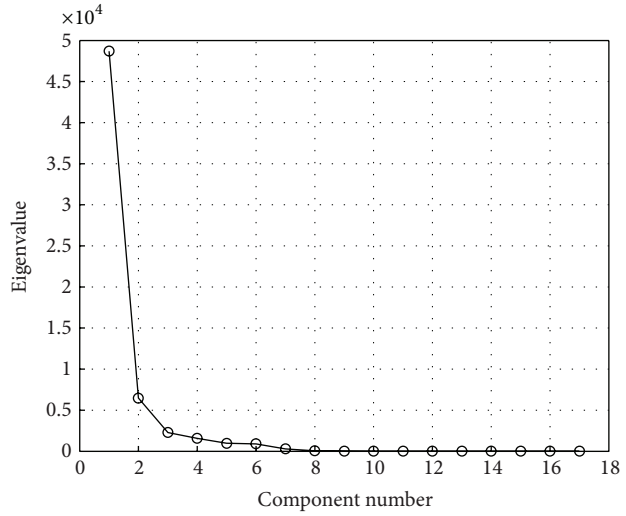

Figure 1: Scree plot of the original PBC dataset including 17 variables and 276 observations.

5.1. PBC Data Numerical Experiment. First, the well-known primary biliary cirrhosis (PBC) dataset (Fleming and Harrington 1991) is considered as the sample collected dataset. These data are from a double-blinded randomized trial including 312 observations. This right-censored dataset contains 17 variables in addition to censoring information, identification number, and event times for each observation. Experimental results for the 276 observations with complete records are presented. For the original PBC dataset, approximate value of k is 3. This value is the subset size in the algorithms applied on **V** transformed logical dataset in the n-by-p matrix, shown in Figure 1.

To verify the performance of the proposed methods, the result of methods and algorithms for the transformed logical PBC dataset is compared with the results of random survival forests (RSF) method [23, 48], additive risk model (ADD) [24], and weighted least square (LS) [25] for variable selection in the PBC dataset given in Table 1. RSF is a method for the analysis of right-censored survival data using splitting rules for growing survival trees, and it is based on the random forests (RF) approach that ensembles learning which can be improved further by injecting randomization into the base learning process. Therefore, the RSF methodology is an extension of the RF method in which randomization is used in two forms; by randomly drawing a bootstrap sample of the data and by randomly selecting a subset of variables [23]. The ADD proposes a principal component regression to give unique and numerically stable estimators. This model assumes that the hazard function is the summation of the baseline hazard function and the regression function of covariates [24]. The LS estimator uses KM weights to account for censoring in the least squares criterion and makes the adaptation of this approach to multiple covariate settings computationally feasible [25]. A comprehensive comparison of RSF, ADD, and LS performance with other relevant methods in high-dimensional time-to-event data analysis such as Cox's proportional hazard model, LASSO, and PCR has been presented in [23–25].

Each number in Tables 1 and 2 represents a specific variable in experiment dataset. For example, in Table 1, variable 15 is among the selected less efficient variables by NTS (F), SSG, RSF, and LS methods.

From the results shown on Table 1, the NTS (F) method has the closest performance to the SSG. This is an advantage of these methods that more than 80% of inefficient variables which have been detected by other methods (RSF, LS, and ADD) are collected by proposed algorithms at significantly short periods of calculation. The robustness of this class of methods has been examined for several time-to-event data samples collected with high-dimensional covariates in this study.

Table 1: Selected less efficient variables in proposed methods as compared to those from RSF, ADD, and LS methods.

Method	Selected less efficient variables
NTS (F)	1, 3, 5, 6, 10, 15, 17
NTS (C)	2, 5, 10, 11, 13, 17
SSG	1, 3, 5, 10, 15, 17
RSF	1, 3, 5, 12, 13, 14, 15, 17
ADD	1, 2, 5, 12, 14
LS	1, 2, 3, 14, 15, 17

Table 2: Selected less efficient variables in all proposed methods and comparison to simulation defined pattern.

Method	Selected less efficient variables
NTS (F)	5, 10, 25
NTS (C)	5, 10, 25
SSG	5, 10, 25
Definition	5, 10, 15, 20, 25

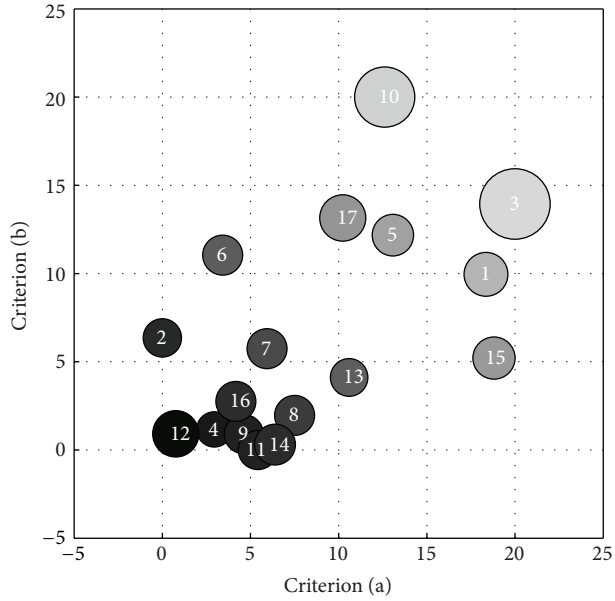

Figure 2: Hybrid scatter plot (upper right variables have less efficiency).

5.2. Simulation Numerical Experiment. Continuing the validation of the proposed methods, we set $n = 400$ observations and $p = 25$ variables and simulated event times from a pseudorandom algorithm. In addition to constant and periodic binary numbers, normal and exponential distributed pseudorandom numbers are generated as independent values of explanatory variables. The censored observations were also generated independent of the events. Additionally, some variables are set as a linear function of event time intentionally. We present the results of methods and algorithms applying the simulated data in Table 2. These results are compared with the simulation defined pattern and the comparison verifies the performance of all proposed methods and algorithms.

Inefficiency analysis results for the simulation experiment show that variables with identification numbers 5, 10, and 25 are detected as less efficient variables among all 25 simulated variables by NTS (F), NTS (C), and SSG methods. In this example, one could eliminate these less efficient identified variables if the objective is to reduce the number of variables in the dataset for further analysis.

In addition, a set of numerical experiments are presented to evaluate the performance of the proposed methods as shown in Table 3. Six different simulation models are defined. These models are set as explained above for $n = 400$ observations and $p = 25$ variables following same data generating algorithm. The number of significant inefficient variables in each simulation model is set $m = \{4, 5, 6\}$ as presented in *Definition* in Table 3. The simulations are repeated 100 times independently. For each method, m represents integer average number of determined and selected inefficient variables and p denotes the performance of the method, for 100 trails, where p is as follows:

$$p = \frac{m_{\text{method}}}{m_{\text{definition}}}, \tag{12}$$

and $\overline{p}$ is the average of performance of each method. The adjusted censoring rates for all simulation numerical experiments are $50 \pm 5\%$.

5.3. Hybrid Scattering for Variable Inefficiency. Figure 2 displays a graphical interpretation of the results from the proposed methods, hybrid scattering for variable inefficiency for three criteria as (a) NTS (F) score, (b) NTS (C) score, and (c) SSG score for PBC data. Each variable is exposed by a circle icon where two axes represent the standardized criteria (a) and (b) and the diameter of each circle demonstrates the standardized criterion (c). Therefore, it is interpreted from the hybrid scatter plot that each variable with larger diameter and more distance from the center has less efficiency and is an ideal candidate to remove from the dataset if it is desired.

6. Conclusions

This paper presents applied procedures beneficial to covariate reduction through an inefficient variable selection approach which enables one to obtain an appropriate variable subset in a right-censored high-dimensional and large-scale time-to-event data. Two hybrid methods are introduced in which the experimental results and comparisons demonstrate their performance. By using such novel methods in the field of reliability, data analysis and decision making processes will be faster, simpler, and more accurate.

We aim to apply median rank estimates in the proposed nonparametric methods and develop a weighted score and penalized nonparametric maximum likelihood and least square in the proposed logical time-to-event model by the accelerated failure time model. Also, the next challenge in this research is to face the data streaming circumstances in which

Table 3: Performance measures based on the six simulation experiments; m is the integer average number of selected inefficient variables and p is the performance of the method based on 100 replications.

Method	Simulation model												Ave.
	#1		#2		#3		#4		#5		#6		
	m	p	m	p	m	p	m	p	m	p	m	p	$\overline{p}$
NTS (F)	3	0.75	3	0.75	3	0.60	4	0.80	4	0.67	5	0.83	0.73
NTS (C)	3	0.75	3	0.75	4	0.80	4	0.80	5	0.83	5	0.83	0.79
SSG	2	0.50	3	0.75	3	0.60	4	0.80	4	0.67	4	0.67	0.66
Definition	4		4		5		5		6		6		

variables are time-dependent and real-time analysis of data is more complicated.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] F. Yao, "Functional principal component analysis for longitudinal and survival data," *Statistica Sinica*, vol. 17, no. 3, pp. 965–983, 2007.

[2] J. Hellerstein, "Parallel Programming in the Age of Big Data," Gigaom Blog, 2008.

[3] T. Segaran and J. Hammerbacher, *Beautiful Data: The Stories Behind Elegant Data Solutions*, O'Reilly Media, 2009.

[4] D. Feldman, M. Schmidt, and C. Sohler, "Turning Big data into tiny data: constant-size coresets for k-means, PCA and projective clustering," in *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '13)*, pp. 1434–1453, January 2013.

[5] J. Manyika, M. Chui, B. Brown et al., *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, 2011.

[6] J. Moran, *Is Big Data a Big Problem for Manufacturers?* Sikich, 2013.

[7] B. Brown, M. Chul, and J. Manyika, "Are you ready for the era of 'big data'?" *McKinsey Quarterly*, no. 4, pp. 24–35, 2011.

[8] S. Lohr, "The age of big data," *New York Times*, 2012.

[9] E. Brynjolfsson, *A Revolution in Decision-Making Improves Productivity*, MIT Press, 2012.

[10] P. Russom, "Big data analytics," TDWI Best Practices Report, Fourth Quarter, 2011.

[11] K. Sadeghzadeh and M. B. Salehi, "Mathematical analysis of fuel cell strategic technologies development solutions in the automotive industry by the TOPSIS multi-criteria decision making method," *International Journal of Hydrogen Energy*, vol. 36, no. 20, pp. 13272–13280, 2011.

[12] J. Chai, J. N. K. Liu, and E. W. T. Ngai, "Application of decision-making techniques in supplier selection: a systematic review of literature," *Expert Systems with Applications*, vol. 40, no. 10, pp. 3872–3885, 2013.

[13] A. R. Tenner and I. J. Detoro, *Total Quality Management: Three Steps to Continuous Improvement*, Addison-Wesley, 1994.

[14] D. A. Garvin, "Competing on the eight dimensions of quality," *IEEE Engineering Management Review*, vol. 24, no. 1, pp. 15–23, 1996.

[15] D. H. Besterfield, C. Besterfield-Michna, G. H. Besterfield, and M. Besterfield-Sacre, *Total Quality Management*, Prentice Hall, Englewood Cliffs, NJ, USA, 1995.

[16] E. Auschitzky, M. Hammer, and A. Rajagopaul, *How Big Data Can Improve Manufacturing*, McKinsey & Company, 2014.

[17] M. Nemschoff, *How Big Data Can Improve Manufacturing Quality*, 2014.

[18] H. Akaike, "An information criterion (AIC)," *Mathematical Sciences*, 1976.

[19] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[20] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 34, pp. 187–220, 1972.

[21] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, 2011.

[22] J. Buckley and I. James, "Linear regression with censored data," *Biometrika*, vol. 66, no. 3, pp. 429–436, 1979.

[23] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[24] S. Ma, M. R. Kosorok, and J. P. Fine, "Additive risk models for survival data with high-dimensional covariates," *Biometrics*, vol. 62, no. 1, pp. 202–210, 2006.

[25] J. Huang, S. Ma, and H. Xie, "Regularized estimation in the accelerated failure time model with high-dimensional covariates," *Biometrics. Journal of the International Biometric Society*, vol. 62, no. 3, pp. 813–820, 2006.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks, 1984.

[27] E. T. Lee and J. W. Wang, *Statistical Methods for Survival Data Analysis*, John Wiley & Sons, 2003.

[28] Jr. Hosmer D. W. and S. Lemeshow, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, John Wiley & Sons, New York, NY, USA, 1999.

[29] L.-P. Kronek and A. Reddy, "Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data," *Bioinformatics*, vol. 24, no. 16, pp. i248–i253, 2008.

[30] T. R. Holford, *Multivariate Methods in Epidemiology*, Oxford University Press, Oxford, UK, 2002.

[31] A. V. Peterson Jr., "Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions," *Journal of the American Statistical Association*, vol. 72, no. 360, p. 854, 1977.

[32] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2006.

[33] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison Wesley, Pearson, 2006.

[34] I. Jolliffe, *Principal Component Analysis*, John Wiley & Sons, New York, NY, USA, 2005.

[35] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[36] H. Motoda and L. Huan, *Feature Selection, Extraction and Construction*, Communication of IICM, 2002.

[37] D. Addison, S. Wermter, and G. Z. Arevian, "A comparison of feature extraction and selection techniques," in *Proceedings of the International Conference on Artificial Neural Networks*, Istanbul, Turkey, 2003.

[38] J. Ye, *Unsupervised and Supervised Dimension Reduction: Algorithms and Connections*, Arizona State University, Tempe, Ariz, USA, 2005.

[39] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[40] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[41] K. Sadeghzadeh and N. Fard, "Nonparametric data reduction approach for large-scale survival data analysis," *IEEE*. In press.

[42] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: an enabling technique," *Data Mining and Knowledge Discovery*, vol. 6, no. 4, pp. 393–423, 2002.

[43] K. Sadeghzadeh and N. Fard, "Multidisciplinary decision-making approach to high-dimensional event history analysis through variable reduction methods," *European Journal of Economics and Management*, vol. 1, no. 2, 2014.

[44] T. A. Feo and M. G. C. Resende, "Greedy randomized adaptive search procedures," *Journal of Global Optimization*, vol. 6, no. 2, pp. 109–133, 1995.

[45] J. P. Hart and A. W. Shogan, "Semi-greedy heuristics: an empirical study," *Operations Research Letters*, vol. 6, no. 3, pp. 107–114, 1987.

[46] M. Gendreau and J. Y. Potvin, "Tabu Search," in *Search Methodologies*, Springer, Berlin, Germany, 2005.

[47] W. R. Zwick and W. F. Velicer, "Comparison of five rules for determining the number of components to retain," *Psychological Bulletin*, vol. 99, no. 3, pp. 432–442, 1986.

[48] I. Ishwaran and U. B. Kogalur, "Random survival forests for R," *Rnews*, 2007.