PENALIZED METHOD BASED ON REPRESENTATIVES & NONPARAMETRIC ANALYSIS OF GAP DATA

A Thesis
Presented to
The Academic Faculty

by

Soyoun Park

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2010

PENALIZED METHOD BASED ON REPRESENTATIVES
&
NONPARAMETRIC ANALYSIS OF GAP DATA

Approved by:

Professor Jye-Chyi Lu, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Assistant Professor Nicoleta Serban
School of Industrial and Systems Engineering
Georgia Institute of Technology

Associate Professor Xiaoming Huo
School of Industrial and Systems Engineering
Georgia Institute of Technology

Associate Professor Martha Grover
Chemical & Biomolecular Engineering
Georgia Institute of Technology

Assistant Professor Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: September 7 2010

DEDICATION

To my parents, Ohksoon Yim and Kwangsoo Park;

my sisters, Eunhye and Eunkyung;

my husband, Joongsup Jay, Lee and my baby

ACKNOWLEDGEMENTS

I am heartily thankful to my advisor, Professor Jye-Chyi Lu, whose encouragement, guidance and support from the initial to the final enables me to complete this dissertation and finish the long journey to pursuit the Ph.D.

For this dissertation, I would like to thank my reading committee members: Martha Grover, Xiaoming Huo, Yajun Mei, and Nicoleta Serban for their time, interest, and helpful comments.

I would like to express my deepest gratitude to my parents, sisters, and my husband for all their love and encouragement. For my parents who are always so positive despite it all, who are so giving, who work so hard to provide the best for their children, and who have always believed in me. For my sisters Eunhye and Eunkyung who always encouraged me to continue the study. And most of all for my loving, supportive, and encouraging husband Joongsup whose faithful support and patience during all the stages of this Ph. D is so appreciated.

Lastly, I thank God for all His blessings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

When there are a large number of predictors and few observations, building a regression model to explain the behavior of a response variable such as a patient's medical condition is very challenging. This is a "$p \gg n$" variable selection problem encountered often in modern applied statistics and data mining. Chapter one of this thesis proposes a rigorous procedure which groups predictors into clusters of "highly-correlated" variables, selects a representative from each cluster, and uses a subset of the representatives for regression modeling. The proposed Penalized method based on Representatives (PR) extends the Lasso for the $p \gg n$ data and highly correlated variables, to build a sparse model practically interpretable and maintain prediction quality. Moreover, we provide the PR-Sequential Grouped Regression (PR-SGR) to make computation of the PR procedure efficient. Simulation studies show the proposed method outperforms existing methods such as the Lasso/Lars. A real-life example from a mental health diagnosis illustrates the applicability of the PR-SGR.

In the second part of the thesis, we study the analysis of time-to-event data called a gap data when missing time intervals (gaps) possibly happen prior to the first observed event time. Estimation of survival function of the first true event time is complicated by the occurrence of gaps in the gap data. If a gap occurs prior to the first observed event, then the first observed event time may or may not be the first true event time. This incomplete knowledge makes the gap data different from the well-studied regular interval censored data. We propose a Non-Parametric Estimate for the Gap data (NPEG) to estimate the survival function for the first true event time on the gap data, derive its analytic properties and demonstrate its performance in simulations. We also extend the Imputed Empirical Estimating method (IEE),

which is an existing nonparametric method for the gap data up to one gap, to handle the gap data with multiple gaps.

CHAPTER I

PENALIZED METHOD BASED ON REPRESENTATIVES (PR)

1.1 Introduction and Motivation

Consider a problem of parameter estimation for a linear regression model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon \\ &= \sum_{j=1}^{p} X_j \beta_j + \epsilon, \end{aligned} \tag{1}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$ is an n-vector of responses, $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ is an $n \times p$ matrix with linearly independent predictors $X_j = (X_{1j}, \cdots, X_{nj})^T$, $j = 1, \cdots, p$, $\beta = (\beta_1, \cdots, \beta_p)^T$ is a p-vector of parameters, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$ is an n-vector of random errors. This chapter studies a regression model in Equation (1) with a special type of data called "large-p-small-n" ($p \gg n$), where the number of predictors p is much larger than the sample size n.

Large-p-small-n data are abundant in many fields such as genomics, microarray study, computational biology, health science, manufacturing and finance to name a few. For example, expression levels of millions of genes are monitored for a few hundreds of subjects in a microarray study. Another example is the thousands of magnetic resonance images (MRI) that are collected for each patient when there are only tens of patients involved. Section 1.4 uses the high-dimensional data from the mental health diagnosis on depression to illustrate the proposed method. Note that in the study, $p = 1794$ predictors of depression symptoms are examined for each of the $n = 15$ patients.

Large-p-small-n regression has been one of the most challenging problems over the

recent years. Due to the high dimensionality of $p \gg n$ data, the parameters in the regression model (1) cannot be uniquely estimated because $(\mathbf{X}^T\mathbf{X})$ is singular ([11]). Moreover, overfitting might occur, resulting in questionable prediction quality.

The dimension reduction is an attractive method to handle the challenges caused by the high dimensionality in $p \gg n$. Principal Component Analysis (PCA) ([23]) transforms a number of possibly correlated predictors into a smaller number of uncorrelated ones called principal components. PCA uses the eigenvectors of the covariance matrix of the predictors and finds the independent axes from the linear combinations of the predictors. While the PCA only uses the highest variation across the predictors, the supervised principal components analysis ([1, 2]) performs PCA using only a subset of predictors which have strong correlations with the response rather than using all the predictors in the dataset. However, since the two PCA methods perform a coordinate rotation, the meaning of the principal components may become obscure, resulting in the loss of the physical meaning of the original predictors.

As another approach for the dimension reduction, Fan and Lv ([12]) proposed a pre-screening method called Sure Independence Screening (SIS) to filter out predictors which have a weak correlation with the response before employing a variable selection method (for example, the Lasso ([34]) or Adaptive Lasso ([40]), to further select a subset of predictors. While this correlation filtering method successfully reduces the dimensionality and keeps the original coordinates of the predictors, it does not handle the spurious high correlations among predictors caused by the high dimensionality of predictors (for example, $p = 1794 \gg n = 10$).

A series of penalized methods ([5, 22, 34, 39, 40, 41]) which impose a penalization on the L_1- or L_2-norms of the regression coefficients have emerged as highly successful techniques in handling the high dimensionality problem. In these methods,

adjustment for overfitting is directly built into the model development and estimation accuracy is improved by effectively identifying a subset of important predictors (variable selection). Hoerl and Kennard ([22]) introduced the Ridge Regression, which finds its coefficients by minimizing the sum of squared error loss subject to an L_2-norm constraint on the coefficients. That is, the solution $\hat{\beta}^{Ridge}$ can be written as follows:

$$\hat{\beta}^{Ridge}(\lambda) = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where λ is a positive regularization parameter. The Ridge Regression achieves a stable fit by shrinking each coefficient based on the variation of the predictors and its solution is simplified to $\hat{\beta}^{Ridge}(\lambda) = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$ ([20, 21]). Consequently, predictors with high positive correlations tend to yield similar coefficients. Since the Ridge Regression always keeps all the predictors in the model ([41]), the number of nonzero coefficients is larger than the sample size n when $p \gg n$.

Instead of using an L_2-norm, Tibshirani ([34]) proposed the Lasso which imposes a penalization on the L_1-norm of the coefficients. The Lasso finds its coefficients by minimizing the sum of squared error subject to an L_1norm constraint on the coefficients. Equivalently, the solution $\hat{\beta}^{Lasso}$ can be written as follows:

$$\hat{\beta}^{Lasso}(\lambda) = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1,$$

where λ is also a positive regularization parameter. Unlike the L_2-penalized methods such as the Ridge Regression, the L_1-penalized methods such as the Lasso assign zero coefficients to a subset of the predictors and hence the Lasso achieves an automatic variable selection. Nevertheless, as Zou and Hastie ([41]) discussed, the Lasso tends to select only one of the predictors when their pairwise correlations are very high, and the Lasso does not concern which one is selected in the model. Since there could be highly correlated predictors simply due to high dimensionality (for example,

see Experiments in Section 1.1.1), the arbitrary selection of only one of the highly correlated predictors can result in "incomplete" use of the selected predictors.

In the presence of high correlations between predictors, the Variance Inflation Factor (VIF) in an Ordinary Least Square (OLS) regression analysis can be used to measure how much the variance of an estimated regression coefficient increases due to the collinearity. VIF_j, VIF for a predictor X_j, is calculated as $\frac{1}{1-R_j^2}$, where R_j is the multiple correlation coefficient for an OLS regression using X_j as the response against all the other $p-1$ predictors. However, when $p \gg n$, this is not applicable, because such an OLS regression has $p-1$ parameters which is still larger than the sample size n.

Zou and Hastie ([41]) introduced the Elastic-Net which uses a convex combination of the L_1- and L_2-norms. The solution $\hat{\beta}^{Elastic-Net}$ can be written as follows:

$$\hat{\beta}^{Elastic-Net}(\lambda_1, \lambda_2) = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2,$$

where λ_1 and λ_2 are positive regularization parameters. Since the Elastic-Net compromises between the Ridge and Lasso, it selects a subset of the predictors like the Lasso, and assigns similar coefficients to highly correlated predictors like the Ridge Regression so that it yields more than n nonzero coefficients. The Elastic-Net is useful in the situation when highly correlated predictors selected simultaneously in the model are meaningful. For example, consider a gene expression study where several genes share a common biological pathway so that those genes express together. Naturally, their correlations can be high ([33]), and including all those genes simultaneously has better interpretability than including only one gene. However, if there are highly correlated genes that are detected but do not have any special meaning in the gene expression study, there is a great deal of confusion and redundancy in selecting the highly correlated predictors, where the ones without the "meaningful" information are included in the model. Moreover, the Elastic-Net method may not reveal the information of which predictors are actually highly correlated based on

their estimated coefficients.

As another method to select highly correlated predictors simultaneously in the model like the Elastic-Net, Bondell and Reich ([5]) proposed the OSCAR which uses a combination of the L_1- and the pairwise $L_\infty-$norms. The solution $\hat{\beta}^{OSCAR}$ can be written as follows:

$$\hat{\beta}^{OSCAR}(\lambda, c) = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 + c\lambda \sum_{j<k} \max\{\mid \beta_j \mid, \mid \beta_k \mid\},$$

where c and λ are positive regularization parameters. The L_1-norm selects a subset of the predictors like the Lasso, while the pairwise $L_\infty-$norm yields the same coefficients for highly correlated predictors. Therefore, the OSCAR provides the additional information on which predictors are highly correlated to each other based on their estimated coefficients. That is, if two predictors have the same estimated coefficients, it means that they are highly correlated. However, as the real data study in ([5]) showed, two predictors with different estimated coefficients can have higher correlation than other two predictors with the same estimated coefficients. Therefore, the OSCAR also suffers somewhat from the inconsistent information of highly correlated predictors.

In some problems, the predictors belong to predefined groups (factors). For example, in regression problems with categorical predictors, a set of dummy predictors or two/three way interaction dummy predictors can be used to build a group of derived input predictors. When such groups are available, it may be desirable to select all the predictors from individual group. In this case, the variable selection problem becomes the selection problem of groups. Given predefined groups, Yuan and Lin ([39]) proposed a general version of the Lasso called Grouped-Lasso. Suppose that the p predictors are divided into the J factors with factor size p_j, $j = 1, \cdots, J$. Let $\mathbf{X}_j$ be an $n \times p_j$ matrix corresponding to the jth factor, and β_j be a coefficient vector of size p_j, $j = 1, \cdots, J$. Now consider a general regression problem with J factors

$\mathbf{y} = \mathbf{X}\beta + \epsilon = \sum_{j=1}^{J} \mathbf{X}_j\beta_j + \epsilon$, where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_J)$ and $\beta = (\beta_1^T, \cdots, \beta_J^T)^T$. Then, the Grouped-Lasso estimate can be written as $\hat{\beta}^{Grouped-Lasso}$:

$$\hat{\beta}^{Grouped-Lasso}(\lambda) \quad = \quad \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j}\ \|\beta_j\|_2,$$

where $\|\cdot\|_2$ is the Euclidean norm and λ is a positive constant. If $p_1 = \cdots = p_J = 1$, $\hat{\beta}^{Grouped-Lasso}(\lambda)$ is equivalent to $\hat{\beta}^{Lasso}(\lambda)$. As Yuan and Lin ([39]) discussed, the Grouped-Lasso shares the properties of both L_1- and L_2-penalized methods. Therefore, the Grouped-Lasso selects only a subset of the predefined factors by assigning nonzero coefficients, resulting in automatic variable selection. However, their procedure does not consider the problem that groups may consist of highly correlated predictors.

When $p \gg n$, one faces two challenges with the application of the penalized methods. First, the application of the penalized methods can be computationally costly. Second, there is also a higher chance of having highly correlated predictors as the data dimensionality increases. Note that the high dimensionality itself can induce spuriously highly correlated predictors as Hall *et al.* discussed ([19]). With the presence of such spurious high correlations, it is not meaningful to contain a subset of the highly correlated predictors in the model like the Elastic-Net or OSCAR. Moreover, there are no predefined groups of highly correlated predictors for the methods like Grouped-Lasso.

In this Chapter, we propose a new method called the "Penalized method based on Representatives (PR)", which can achieve dimension reduction and handle high correlations among the predictors for $p \gg n$ data. The PR procedure creates clusters of predictors based on their pairwise correlations. Moreover, the proposed method selects "representatives" from the clusters and builds a regression model based on these representatives. All aforementioned steps in the PR procedure are rigorously defined and formulated. To make computation of the PR procedure efficient, each

group of highly correlated predictors are built via the PR-Sequential Grouped Regression (PR-SGR) Algorithm, which is an adaption of fast algorithm for the PR to make its implication realistic by the Lars ([9]) algorithm in Section 1.2.4. Without the predefined groups of highly correlated predictors, the selected representatives can characterize the corresponding group of highly correlated predictors and provide a physical interpretation of the impact of the predictors to the response. We show that the PR can be interpreted as a variant of Lasso with a penalty of the L_1-norm of coefficients applied to the predictor-clusters' representatives. Its L_1-penalty achieves an automatic variable selection and builds a fitted model with less than n representatives. Moreover, selecting a representative in each group using a simple rule introduced in Section 1.2.1.2 reduces the computational complexity and achieves better interpretability of the final model.

Section 1.1.1 uses two experiments to further motivate the study of this chapter. Section 1.2 proposes the PR method and will discuss its rationale. The proposed method will be illustrated with simulated examples in Section 1.3 and real dataset in Section 1.4. Section 1.5 concludes this chapter with discussions and summaries.

1.1.1 Experiment 1 and 2

Experiments 1 and 2 are conducted to illustrate the existence of spurious high correlations between the predictors due to the high dimensionality of $p \gg n$ data and to support the motivation of this thesis study.

Experiment 1: Suppose predictors $X_1, \cdots, X_p$ are independent and follow the standard normal distribution. Based on the generated data matrix $\mathbf{X} = (X_1, \cdots, X_p)$, the maximum absolute sample correlation coefficient among predictors was calculated. We simulated 500 datasets with $n = 20, 40$ and 80 for $p = 2000$, and $n = 10, 20$ and 40 for $p = 1000$, respectively. Figure 1 shows that the distribution of the maximum absolute correlation coefficients shifts to the right as the ratio p/n increases. Although

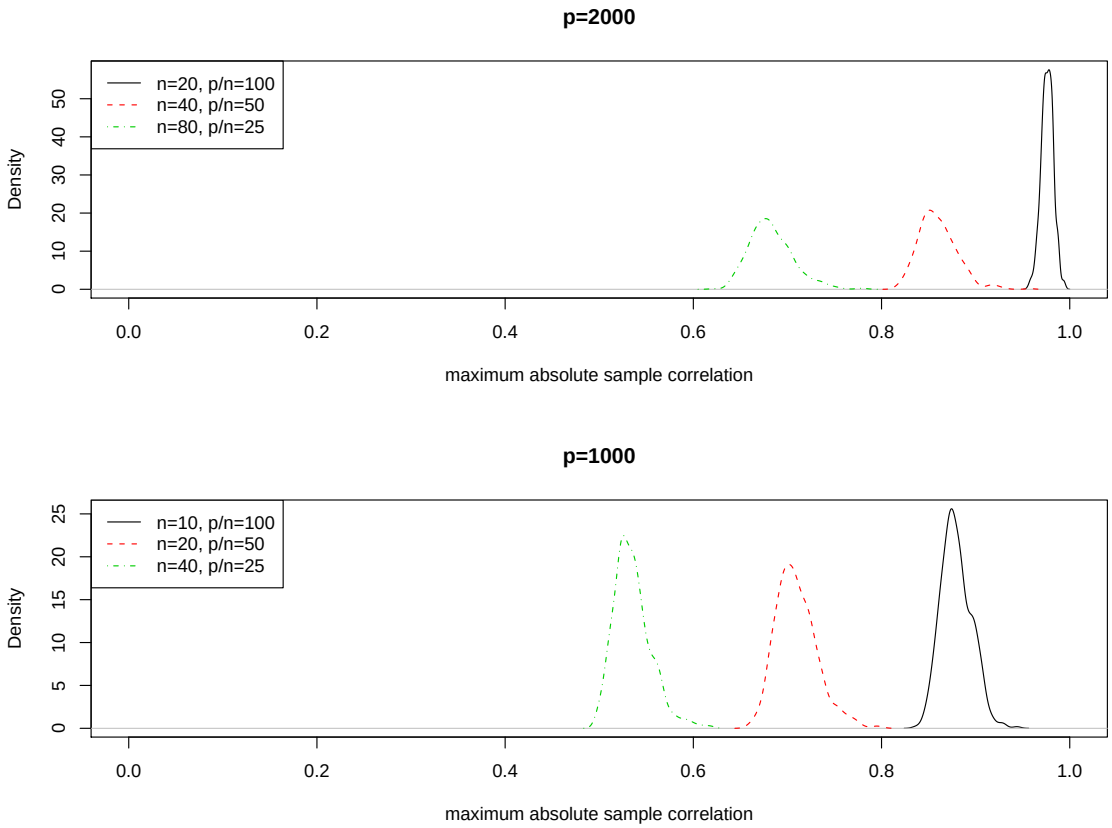


Figure 1: Distributions of the maximum absolute sample correlations for $p/n = 100$(solid black line), 50(dashed red line) and 25(dotted and dashed green line) with $n = 1000$ and 2000.

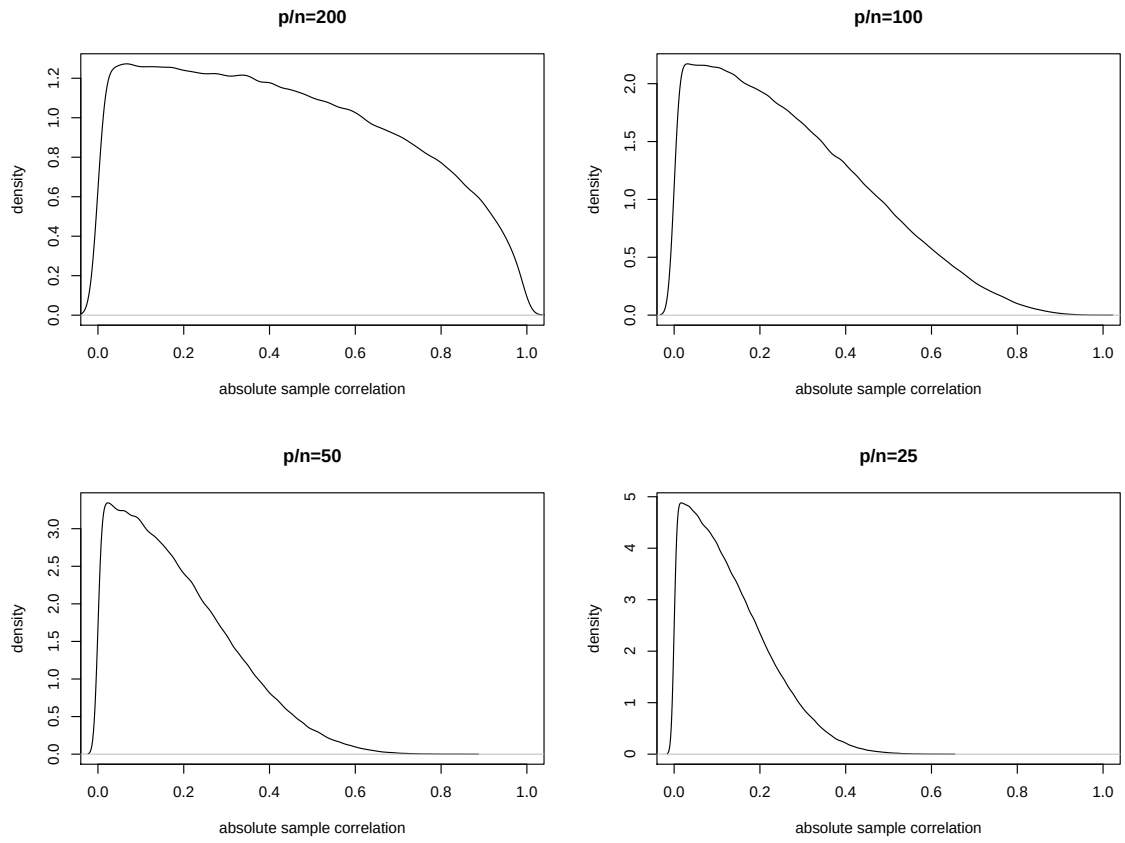


Figure 2: Distributions of absolute sample pairwise correlations with $p/n = 200$(left-top), 100(right-top), 50 (left-bottom) and 25(right-bottom) with $p = 1000$.

the predictors are generated randomly, the maximum absolute sample correlation coefficient among predictors can be very large. Figure 2 shows the density of absolute values of all the sample pairwise correlation coefficients among p predictors with $p = 1000$ and $n = 5, 10, 20$ and 40. It shows that more predictors are spuriously correlated as p/n increases. Experiment 1 showed that the chance of having spuriously highly correlated predictors increases with the dimensionality.

Experiment 2: Suppose we have two random vectors X and Y following the standard normal distribution, with correlation $\rho_0 = Corr(X, Y) = 0.9$ and the sample size $n = 50$. We also generated 10000 random vectors X' following the standard normal distribution with the same sample size, and calculated the correlation $Corr(X, X')$.

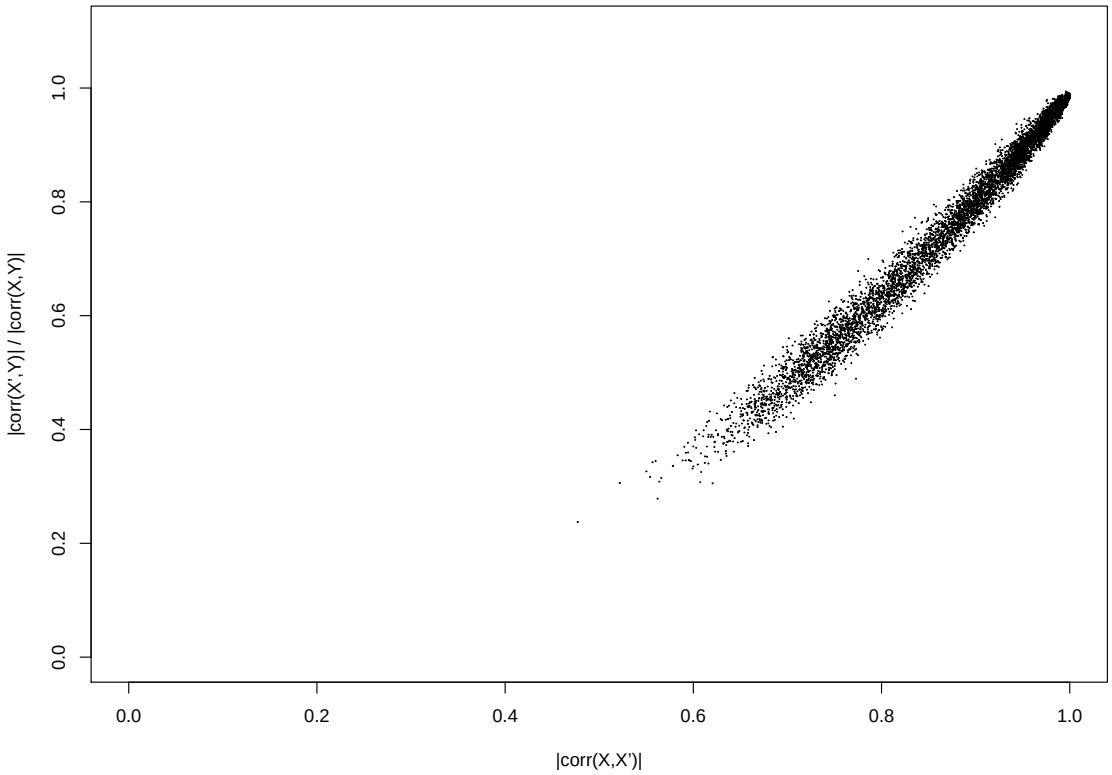


Figure 3: Ratios of $\mid Corr(X', Y) \mid / \mid Corr(X, Y) \mid$ vs. absolute values of correlation between X and X'.

Figure 3 shows that the ratio of $\mid Corr(X', Y) \mid / \mid Corr(X, Y) \mid$ increases as the value of $\mid Corr(X, X') \mid$ increases. We had the same conclusion for different values of n and correlation coefficient ρ_0. This experiment shows that for highly correlated predictors, their correlations to the response, respectively, have similar values. This means that their regression coefficients would have similar values.

1.2 Methodology

In this section, we propose a new approach called the Penalized method based on Representatives (PR) for $p \gg n$ data and correlated predictors. The proposed method formulates clusters/groups of predictors, which have high pairwise correlations, and selects one of the predictors from each cluster as a representative. Then, apply the

Lasso regularization to select regression predictors from these representatives. The PR minimizes summation of the least squared error and the total representative selection error according to the L_1-penalty. The L_1-norm achieves an automatic variable selection and selects less than n representatives in the final model. The total representative selection error depends on different representative selection rules. See Section 1.2.1.2 for details. The following formulates this method rigorously.

1.2.1 Penalized Method Based on Representatives (PR)

1.2.1.1 Definition and Notation

Suppose that there are n data points, $\{(\mathbf{X}, \mathbf{y}) : \mathbf{X} \in \mathcal{X} \subset \mathbb{R}^{n \times p},\ \mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^n\}$, sampled from an unknown distribution. Here, $\mathbf{y}$ is the response and $\mathbf{X}$ is the matrix of $p-$predictors. Let $X_j = (X_{1j}, \cdots, X_{nj})^T$, $j = 1, \cdots, p$ be vectors representing the predictors where $\mathbf{X} = (X_1, X_2, \cdots, X_p)$, and $\mathbf{y} = (y_1, \cdots, y_n)^T$. Assume that the predictors have been standardized and the response has been centered so that

$$\sum_{i=1}^{n} X_{ij} = 0,\ \sum_{i=1}^{n} X_{ij}^2 = 1,\ \text{and}\ \sum_{i=1}^{n} y_i = 0,\ \text{for}\ j = 1, 2, \cdots, p.$$

Suppose that the p predictors are divided into the m clusters $\mathcal{C}_1, \cdots, \mathcal{C}_m$ with cluster size $p_j,\ j = 1, \cdots, m$. Let $\mathbf{X}_j$ be an $n \times p_j$ matrix corresponding to the cluster $\mathcal{C}_j$, β_j be a coefficient vector of size p_j, $j = 1, \cdots, m$, and $\beta = (\beta_1^T, \cdots, \beta_m^T)^T$.

Let $\mathcal{R}$ be the set of indices of representatives $X^1, \cdots, X^m \in \{X_1, \cdots, X_p\}$ where $X^j = X_k$ for some $k \in \mathcal{C}_j,\ j = 1, \cdots, m$ (i.e., the representative X^j of the cluster $\mathcal{C}_j$ is X_k for some $k \in \mathcal{C}_j$). For example, where a cluster $\mathcal{C}_1 = \{1, 2, 3\}$ is given (i.e., the cluster consists of $X_1,\ X_2$ and X_3) and X_2 is selected as a representative X^1 of $\mathcal{C}_1$, then $\{2\} \in \mathcal{R}$. Let $\mathbf{X}_\mathcal{R} = (X^1, \cdots, X^m)$ be the submatrix of $\mathbf{X}$ corresponding to the representatives, and $\beta_\mathcal{R}$ be the m-dimensional vector of regression coefficient corresponding to $\mathbf{X}_\mathcal{R}$. The representative set $\mathbf{X}_\mathcal{R}$ is determined by one of the representative selection rules described in Section 1.2.1.2. After selecting the representatives, the original problem of estimating the p-vector β from the linear regression model (1) is

reduced to the problem of estimating a $m-$dimensional vector $\beta_{\mathcal{R}} = (\beta_{(1)}, \cdots, \beta_{(m)})^T$ based on a smaller sub-model,

$$\mathbf{y} = \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}} + \epsilon.$$

The PR method estimates $\beta_{\mathcal{R}}$ by minimizing a loss function of errors. This loss function quantifies both the quality of model fitting and the cost of selecting variables among representatives.

The loss function for the PR with a given representative set $\mathbf{X}_{\mathcal{R}}$ is defined as

$$L((\mathbf{y}, \mathbf{X}_{\mathcal{R}}), \beta_{\mathcal{R}}) = \|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}}\|_2^2 + wG(\beta_{\mathcal{R}}, (\mathbf{y}, \mathbf{X}_{\mathcal{R}})),\ w > 0, \tag{2}$$

where

$$G(\beta_{\mathcal{R}}, (\mathbf{y}, \mathbf{X}_{\mathcal{R}})) = \sum_{j=1}^{m} c_j I(\beta_{(j)} \neq 0)$$

is the total representative selection error, and

$$c_j \ = \ \frac{1}{p_j} \sum_{k \in \mathcal{C}_j} \mid R^2_{\mathbf{y}|X^j} - R^2_{\mathbf{y}|X_k} \mid,\ 0 \leq c_j \leq 1$$

is the representative selection error on the cluster $\mathcal{C}_j$. $R^2_{\mathbf{y}|X}$ denotes the coefficient of determination, R^2, which is the portion of the variability in the response $\mathbf{y}$ accounted for by the fitted simple regression model $\mathbf{y} = \hat{\alpha} + \hat{\beta}X$. If $p_1 = \cdots = p_m = 1$, $c_j = 0$ and thus $G(\beta_{\mathcal{R}}, (\mathbf{y}, \mathbf{X}_{\mathcal{R}})) = 0$. Consider the example with $\mathcal{C}_1 = \{1, 2, 3\}$ and then calculate $R^2_{\mathbf{y}|X_k},\ j = 1, 2, 3$. Since the presentative X^1 of $\mathcal{C}_1$ is X_2, the representative selection error c_1 on the cluster $\mathcal{C}_1$ is

$$\begin{aligned} c_1 \ &= \ \frac{1}{3}\left\{ \mid R^2_{\mathbf{y}|X^1} - R^2_{\mathbf{y}|X_1} \mid + \mid R^2_{\mathbf{y}|X^1} - R^2_{\mathbf{y}|X_2} \mid + \mid R^2_{\mathbf{y}|X^1} - R^2_{\mathbf{y}|X_3} \mid \right\} \\ &= \ \frac{1}{3}\left\{ \mid R^2_{\mathbf{y}|X_2} - R^2_{\mathbf{y}|X_1} \mid + \mid R^2_{\mathbf{y}|X_2} - R^2_{\mathbf{y}|X_3} \mid \right\}. \end{aligned}$$

The clusters are formulated in Step 3 of the PR-Sequential Grouped Regression (PR-SGR) Algorithm, based on the given minimum pairwise correlations ρ among the

predictors (see Section 1.2.4 for details). With this clustering method, any two predictors in a cluster have at least ρ pairwise correlation, while correlations between the predictors from different clusters are less than ρ. Please see Section 1.2.4.2 for the decisions of ρ value. While the value of ρ in Example 11 is selected by a Cross-Validation, the values of ρ in Examples 12-20 are 0.9, 0.8, $\cdots$, 0.1, respectively to investigate the sensitivity of ρ selection. As discussed in Section 1.3, the performance of the PR-SGR for a fixed representative selection rule does not change much for large enough ρ values ($\rho \geq 0.3$), and the PR-SGR with the MAX representative selection rule is the least sensitive to the changes in ρ among all the representative selection rules.

For a predictor X_k in $\mathcal{C}_j$ and the representative X^j of the same cluster $\mathcal{C}_j$, the difference $\mid R^2_{\mathbf{y}|X^j} - R^2_{\mathbf{y}|X_k} \mid$ is the extra amount of the variability explained by the fitted simple regression model of the response on the representative X^j but not explained by the fitted model on the predictor X_k. Thus, c_j, the representative selection error of the cluster $\mathcal{C}_j$ is calculated as the average amount of such extra variability explained by the fitted model on the cluster representative X^j which cannot be explained by the fitted model on another predictor in the cluster. The value of $G(\beta_{\mathcal{R}}, (\mathbf{y}, \mathbf{X}_{\mathcal{R}}))$ is defined as the summation of the representative selection errors from all clusters. We can control the effect of the representative selection procedure by changing the value of w.

The PR method is proposed to find the coefficients by minimizing the loss function $L((\mathbf{y}, \mathbf{X}_{\mathcal{R}}), \beta_{\mathcal{R}})$ in Equation (2) subject to an L_1-norm constraint on the coefficients of the representatives. Equivalently, the solution $\hat{\beta}^{PR}_{\mathcal{R}}$ is determined by minimizing:

$$
\begin{aligned}
& L((\mathbf{y}, \mathbf{X}_{\mathcal{R}}), \beta_{\mathcal{R}}) \text{ subject to } \|\beta_{\mathcal{R}}\|_1 \leq t \qquad (3) \\
\Leftrightarrow \quad & \|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}}\|_2^2 + wG(\beta_{\mathcal{R}}, (\mathbf{y}, \mathbf{X}_{\mathcal{R}})) \text{ subject to } \|\beta_{\mathcal{R}}\|_1 \leq t \\
\Leftrightarrow \quad & \left\| \mathbf{y} - \sum_{j=1}^{m} X^j \beta_{(j)} \right\|_2^2 + w \sum_{j=1}^{m} c_j I(\beta_{(j)} \neq 0) \text{ subject to } \sum_{j=1}^{m} \mid \beta_{(j)} \mid \leq t,
\end{aligned}
$$

where $t \geq 0$ is a tuning parameter. The parameter t controls the amount of shrinkage that is applied to the estimates. Please see Section 1.2.3 for the decisions of the tuning parameters w and λ (see Equation (4) below).

This constrained loss function can be written as a penalized function, and thus the PR solution $\hat{\beta}_{\mathcal{R}}^{PR}$ can be written as:

$$\hat{\beta}_{\mathcal{R}}^{PR}(w, \lambda) = \arg\min_{\beta_{\mathcal{R}}} \|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}}\|_2^2 + wG(\beta_{\mathcal{R}}, (\mathbf{y}, \mathbf{X}_{\mathcal{R}})) + \lambda\|\beta_{\mathcal{R}}\|_1, \tag{4}$$

where λ and w are positive constants. The $\hat{\beta}_{\mathcal{R}}^{PR}$ is equivalent to $\hat{\beta}^{Lasso}$ for $p_1 = \cdots = p_m = 1$, because $c_j = 0$ and $G(\beta_{\mathcal{R}}, (\mathbf{y}, \mathbf{X}_{\mathcal{R}})) = 0$. By imposing the L_1-penalty, the PR assigns zero coefficients to a subset of the representatives and thus none of the predictors in some clusters may be selected. Consequently, the PR achieves an automatic variable selection. As the simulation experiments in Section 1.3 and real-life case study in Section 1.4 show, the proposed method performs superior to the Lasso/Lars, Elastic-Net, grouped-Lasso and grouped-Lars and that it tends to select less variables than those comparable methods.

1.2.1.2 Representative Selection Rules

For each cluster $\mathcal{C}_j$, $j = 1, \cdots, m$, the following representative selection rules can be used to determine a representative X^j of the cluster $\mathcal{C}_j$ where $X^j = X_k,\ k \in \mathcal{C}_j$.

1. (**MAX** rule) Select a predictor X_k whose correlation with the response is maximized:

$$k = \arg\max_{i \in \mathcal{C}_j} Corr(X_i, \mathbf{y}) = \arg\max_{i \in \mathcal{C}_j} \mid X_i^T \mathbf{y} \mid$$

2. (**MIN** rule) Select a predictor X_k whose correlation with the response is minimized:

$$k = \arg\min_{i \in \mathcal{C}_j} Corr(X_i, \mathbf{y}) = \arg\min_{i \in \mathcal{C}_j} \mid X_i^T \mathbf{y} \mid$$

3. (**MED** rule) Select a predictor X_k whose correlation with the response is a median value among correlations between the response and the predictors in $\mathcal{C}_j$:

$$k = \arg \text{median}_{i \in \mathcal{C}_j} Corr(X_i, \mathbf{y}) = \arg \text{median}_{i \in \mathcal{C}_j} \mid X_i^T \mathbf{y} \mid$$

4. (**RAN** rule) Select a predictor $X_k,\ k \in \mathcal{C}_j$ at random.

5. (**CRT** rule) Select a predictor X_k which minimizes the representative selection error on $\mathcal{C}_j$:

$$\begin{aligned} k &= \arg \min_{\substack{i \in \mathcal{C}_j \\ X^j = X_i}} c_j \\ &= \arg \min_{i \in \mathcal{C}_j} \frac{1}{p_j} \sum_{l \in \mathcal{C}_j} \mid R^2_{\mathbf{y}|X_i} - R^2_{\mathbf{y}|X_l} \mid . \end{aligned}$$

The simulation experiments and real-life case study show the effects of each representative selection rule and suggest the MED rule as the best.

1.2.2 Solvable Problem PR

This section shows that the PR is a solvable problem by solving a set of subproblems which are of standard forms (i.e., "minimize $f(x)$ subject to $g_i(x) \leq 0,\ i = 1, \cdots, k$", where f is a convex function and $g_i,\ 1 \leq i \leq k$, are convex functions), although the objective function of the PR is not convex due to G. The number of the subproblems of the PR could be large for $p \gg n$ data, because the number of the subproblems increases exponentially as the number of clusters m increases. We show a solution to the PR that can be obtained by solving smaller number of subproblems (i.e., the number of subproblems increases only linearly).

Table 1 shows that the minimization problem (3) has 2^m subproblems in Table 2, because G has total 2^m possible values. w in the minimization problem (3) is a positive constant and $wG(\beta_{\mathcal{R}}, (\mathbf{y}, \mathbf{X}_{\mathcal{R}})) = \sum_{j=1}^{m} wc_j I(\beta_{(j)} \neq 0) = \sum_{j=1}^{m} c'_j I(\beta_{(j)} \neq 0)$ by

Table 1: Possible cases.

Case #	Value of $\beta_{\mathcal{R}} = (\beta_{(1)}, \cdots, \beta_{(m)})'$
1	None of $\beta_{(j)}$ is zero
2	Only $\beta_{(1)}$ is zero and all the others are nonzero.
$\vdots$	$\vdots$
$m+1$	Only $\beta_{(m)}$ is zero and all the others are nonzero.
$\vdots$	$\vdots$
$2^m - 1$	Only $\beta_{(m)}$ is nonzero and all the others are zero.
2^m	$\beta_{\mathcal{R}}$ is zero.

letting $c'_j = wc_j$. For simplicity, we use c_j instead c'_j in this section. The constraint regions $\mathbf{R}_j,\ j = 1, \cdots, 2^m - 1$ in Table 2 are not convex and hence the subproblems except for the case number 2^m are not of standard forms. However, by taking the closure of their constraint regions, the subproblems become solvable problems in Table 3 with the corresponding constraint region $\overline{\mathbf{R}}_j = \{\beta_{\mathcal{R},j} : g_j(\beta_{\mathcal{R},j}) \leq t\}$. That is, each subproblem in Table 3 is of standard form with $\overline{\mathbf{R}}_j$:

$$\text{(Subproblem } j) \qquad \text{minimize } f_j(\beta_{\mathcal{R},j}) \text{ subject to } g_j(\beta_{\mathcal{R},j}) \leq t,$$

The following example illustrates that the PR with $m = 2$ is of standard form:

$$\min \sum_{i=1}^{n} \left(y_i - x_i^1 \beta_{(1)} - x_i^2 \beta_{(2)}\right)^2 + c_1 I(\beta_{(1)} \neq 0) + c_2 I(\beta_{(2)} \neq 0)$$
$$\text{subject to} \quad \mid \beta_{(1)} \mid + \mid \beta_{(2)} \mid \leq t.$$

A constraint region $\mathbf{R}_1 = \{\beta_{\mathcal{R}} = (\beta_{(1)}, \beta_{(2)})^T :\mid \beta_{(1)} \mid + \mid \beta_{(2)} \mid \leq t, \beta_{(1)} \neq 0, \beta_{(2)} \neq 0\}$ contains the border and inside area of the rhombus excluding two lines on axes between $-t$ and t as in the figure on the left panel in Figure 4. Therefore, the constraint region $\mathbf{R}_1$ is not convex due to such open points on the horizontal and vertical axes. On the other hand, the closure of $\mathbf{R}_1$, $\overline{\mathbf{R}}_1$, on the right panel is convex.

Table 2: Subproblems of the minimization problem (3)

Case #	Subproblem with a constraint region $\mathbf{R}$
1	$\min \|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}}\|_2^2 + \sum_{j=1}^{m} c_j$ subject to $\begin{cases} \|\beta_{\mathcal{R}}\|_1 \le t \\ \beta_{(j)} \neq 0 \ , \forall j \end{cases}$ with $\mathbf{R}_1 = \{\beta_{\mathcal{R}} : \|\beta_{\mathcal{R}}\|_1 \le t,\ \beta_{(j)} \neq 0 \ , \forall j\}$
2	$\min \|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}}\|_2^2 + \sum_{j=2}^{m} c_j$ subject to $\begin{cases} \|\beta_{\mathcal{R}}\|_1 \le t \\ \beta_{(1)} = 0 \\ \beta_{(j)} \neq 0 \ , \forall j = 2, \cdots, m \end{cases}$ with $\mathbf{R}_2 = \{\beta_{\mathcal{R}} : \|\beta_{\mathcal{R}}\|_1 \le t,\ \beta_{(1)} = 0, \beta_{(j)} \neq 0 \ , \forall j = 2, \cdots, m\}$
$\vdots$	$\vdots$
$m+1$	$\min \|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}}\|_2^2 + \sum_{j=1}^{m-1} c_j$ subject to $\begin{cases} \|\beta_{\mathcal{R}}\|_1 \le t \\ \beta_{(m)} = 0 \\ \beta_{(j)} \neq 0 \ , \forall j = 1, \cdots, m-1 \end{cases}$ with $\mathbf{R}_{m+1} = \{\beta_{\mathcal{R}} : \|\beta_{\mathcal{R}}\|_1 \le t,\ \beta_{(m)} = 0, \beta_{(j)} \neq 0 \ , \forall j = 1, \cdots, m-1\}$
$\vdots$	$\vdots$
$2^m - 1$	$\min \|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}}\|_2^2 + c_m$ subject to $\begin{cases} \|\beta_{\mathcal{R}}\|_1 \le t \\ \beta_{(j)} = 0, \forall j = 1, \cdots, m-1 \\ \beta_{(m)} \neq 0 \end{cases}$ with $\mathbf{R}_{2^m-1} = \{\beta_{\mathcal{R}} : \|\beta_{\mathcal{R}}\|_1 \le t, \beta_{(j)} = 0 \ , \forall j = 1, \cdots, m-1,\ \beta_{(m)} \neq 0\}$
2^m	$\min \|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R}}\|_2^2$ subject to $\begin{cases} \|\beta_{\mathcal{R}}\|_1 \le t \\ \beta_{\mathcal{R}} = 0 \end{cases}$ with $\mathbf{R}_{2^m} = \{\mathbf{0}\}$

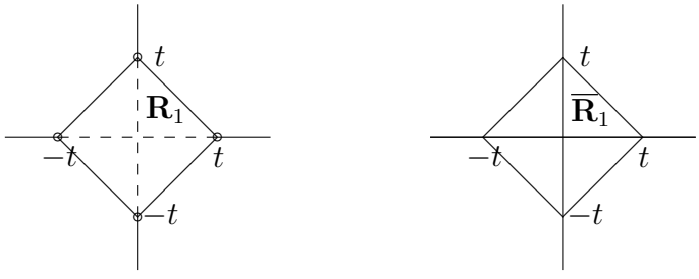


Figure 4: Constraint region $\mathbf{R}_1$ (left) and its closure $\overline{\mathbf{R}}_1$ (right) of the PR with $m = 2$.

Table 3: Subproblems of the minimization problem (3) which are of standard forms.

j	$\beta_{\mathcal{R},j}$	f_j	g_j
1	$\beta_{\mathcal{R},1} = (\beta_{(1)}, \cdots, \beta_{(m)})^T$	$\|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R},1}\|_2^2 + \sum_{j=1}^{m} c_j$	$\|\beta_{\mathcal{R},1}\|_1$
2	$\beta_{\mathcal{R},2} = (0, \beta_{(2)}, \cdots, \beta_{(m)})^T$	$\|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R},1}\|_2^2 + \sum_{j=2}^{m} c_j$	$\|\beta_{\mathcal{R},2}\|_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m+1$	$\beta_{\mathcal{R},m+1} = (\beta_{(1)}, \cdots, \beta_{(m-1)}, 0)^T$	$\|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R},1}\|_2^2 + \sum_{j=1}^{m-1} c_j$	$\|\beta_{\mathcal{R},m+1}\|_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$2^m - 1$	$\beta_{\mathcal{R},2^m-1} = (0, \cdots, 0, \beta_{(m)})^T$	$\|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R},1}\|_2^2 + c_m$	$\|\beta_{\mathcal{R},2^m-1}\|_1$
2^m	$\beta_{\mathcal{R},2^m} = \mathbf{0}$	$\|\mathbf{y} - \mathbf{X}_{\mathcal{R}}\beta_{\mathcal{R},1}\|_2^2$	$\|\beta_{\mathcal{R},2^m}\|_1$

As the solutions of a convex problem always happen on the boundary of its constraint region ([4]), the solution $\hat{\beta}_{\mathcal{R},j}$ minimizes the objective function f_j when $g_j(\hat{\beta}_{\mathcal{R},j}) = t$. Moreover, the solution becomes exactly the solution of the dual problem of general penalized form (or Lagrangian form) "minimize $f_j(\beta_{\mathcal{R},j}) + \lambda_j g_j(\beta_{\mathcal{R},j})$" with a positive λ which depends on the value of t.

Now, the PR estimate $\hat{\beta}_{\mathcal{R}}^{PR}$ is defined based on $\hat{\beta}_{\mathcal{R},j}$'s as:

$$\hat{\beta}_{\mathcal{R}}^{PR} = \arg \min_{\hat{\beta}_{\mathcal{R},j}, j=1,\cdots,2^m,} f_j(\hat{\beta}_{\mathcal{R},j}).$$

Since each subproblem has the global optimum due to its convex objective function and convex constraint region, the PR estimate becomes the global optimum.

Since the number of the subproblems in Table 3 increases exponentially as m increases, we suggest the following stepwise procedure to alleviate the computational burden. If all the representatives are orthogonal, the PR estimate becomes the solution of a subproblem whose constraint region is the closest to the OLS estimate $\hat{\beta}_{\mathcal{R}}^{OLS}$. This can be explained by the geometric interpretation which is used in the Lasso ([34]). The contour of f_j becomes a sphere because all the representatives are orthogonal, and the contour is centered at $\hat{\beta}_{\mathcal{R}}^{OLS}$ as the Lasso explained. Therefore, the solution of (Subproblem j) is the first place that the contour touches its constrain

region $\overline{\mathbf{R}}_j$. That is, the solution has the shortest distance from $\hat{\beta}_{\mathcal{R}}^{OLS}$ to the contour. In this case, we need to calculate the OLS estimate and 2^m distances from $\hat{\beta}_{\mathcal{R}}^{OLS}$ to constraint regions. And then we need to find the subproblem having the shortest distance and solve the subproblem, instead of solving all 2^m subproblems.

On the other hand, for non-orthogonal representatives, the contour of f_j becomes elliptical. Similarly with the orthogonal case, the solution for the subproblem j is the first place that the contour touches its constraint region. Thus, we can find the PR estimate by first solving the subproblems with all m nonzero coefficient $\beta_{(i)}$'s, say the solution $\hat{\beta}_{\mathcal{R},1}$, searching a subproblem with $m-1$ nonzero coefficient $\beta_{(i)}$'s whose contour is the closest to $\hat{\beta}_{\mathcal{R},1}$ and solving the problem. Continue this process until a subproblem with only one nonzero coefficient is found. Therefore, by solving only m subproblems, we can find the final PR solution. Consequently, the number of subproblems solved to find the solution of the PR increases linearly instead of exponentially as the number of clusters m increases.

1.2.3 Tuning Parameters

This section addresses the selection of the tuning parameters ρ (or m), w and λ (or t). The tuning parameter ρ (or m) controls the property of clusters, w controls the effect of the representative selection to the loss function, and λ (or t) controls the amount of the shrinkage.

If a validation dataset is available, tuning parameters can be estimated directly by applying the PR procedure on the validation dataset and minimizing the estimate of the prediction error. If only a training dataset is available, $k-$fold Cross-Validation (k-CV) (for example, $k = 10$) is a popular method for estimating a prediction error and comparing different models ([20]). Since there are three tuning parameters, we need to cross-validate on a three-dimensional surface, as the Elastic-Net ([41]) cross-validates on a two-dimensional surface for its two tuning parameters. We first pick a

(relatively large) value for $\rho \in (0, 1)$, say (0.99, 0.98, $\cdots$, 0.91, 0.90, 0.85, 0.80, 0.70, 0.60). Then, for each ρ, we pick a value for w, say (0, 0.01, 0.1, 1, 5, 10). Then, for each $(\rho,\ w)$, we select the other tuning parameter λ by $k-$CV. At the end, we select $(\rho,\ w)$ that gives the smallest CV error. We used 10$-$CV in the simulation studies and 5$-$CV in the real-life study. Since we use only several values for ρ and w, there is a chance that other $(\rho,\ w)$ gives the lower prediction error than the selected tuning parameters, but the results in the simulation studies and real-life study show that the proposed method with selected tuning parameters among the values described above outperforms other comparable variable selection methods.

For each $(\rho,\ w)$, the computational cost of $k-$CV is the same as that of k-time OLS fitting with extra cost of clustering. Since the PR imposes a regularization on the L_1-norm of the regression coefficients for representatives resulting in at most n representatives in the final model, the computational cost can be manageable in the three-dimensional CV method if the number of pairs $(\rho,\ w)$ tested are moderate.

Other techniques to select tuning parameters are AIC, BIC, or C_p that estimate a prediction error. As a tuning-parameter selection method, each method is computationally more efficient than $k-$CV. Efron *et al.* ([9]) showed that the number of nonzero coefficients is an unbiased estimate of the degrees of freedom (*df*) for the Lasso. They calculated C_p as

$$C_p(\hat{\mu}) \quad = \quad \frac{\|\mathbf{y} - \hat{\mu}\|^2}{\sigma^2} - m + 2df,$$

where $\hat{\mu}$ is a Lars estimate and $\mathbf{y} \sim (\mu, \sigma^2 I)$ with known $\sigma^2 > 0$. Since the PR can be interpreted as a variant of Lasso, we can also select the number of nonzero representative coefficients as an estimate of the degrees of freedom for the PR. This is a measure of the model complexity for the proposed algorithm in the final model.

1.2.4 PR-Sequential Grouped Regression (PR-SGR)

As Madigan and Greg ([28]) discussed, the Lasso had not been popular in statistical practice due to the relative inefficiency of its original algorithm until Efron *et al.* ([9]) provided an efficient and simple algorithm, the Least Angle Regression (Lars), for the Lasso. Therefore, L_1-penalized methods such as the Elastic Net ([41]) and Adaptive Lasso ([40]) adapt the Lars Algorithm for better implementation. The PR, as an L_1-penalized method, also faces implementation difficulty and needs an efficient algorithm to achieve better applicability. This section introduces the PR-Sequential Grouped Regression (PR-SGR) Algorithm that makes the implementation of the PR realistic by adapting the Lars Algorithm.

1.2.4.1 PR-SGR Algorithm

The PR-SGR Algorithm operates in a similar way with the Lars Algorithm except that it additionally builds a cluster and selects a representative of the cluster whenever it adds a new predictor in the model. We describe the PR-SGR Algorithm using the same mathematical notations Efron *et al.* ([9]) used.

- Initial Setting

  Given the data $\{(\mathbf{X}, \mathbf{y}) : \mathbf{X} \in \mathbb{R}^{n\times p},\ \mathbf{y} \in \mathbb{R}^n\}$, where n is the sample size. Denoted by the columns of $\mathbf{X}$ as $X_j = (X_{1j}, \cdots, X_{nj})^T,\ j = 1, \cdots, p$, these columns represent the predictors. Assume that each column has been standardized so that $\sum_{i=1}^{n} X_{ij} = 0$ and $\sum_{i=1}^{n} X_{ij}^2 = 1$. Also assume that the response $\mathbf{y} = (y_1, \cdots, y_n)^T$ has been centered so that $\sum_{i=1}^{n} y_i = 0$.

- PR-SGR Stage

  1. (Compute residual) The initial residual is $\mathbf{r} = \mathbf{y} - \overline{y}\mathbf{1} = \mathbf{y}$, where $\overline{y}$ denotes the mean of the response.

2. (Find the active set) Compute the absolute correlations of the predictors with the residual and select the predictor with the largest absolute-correlation:

$$k^* = \arg \max_{k \in \{1, \cdots, p\}} Corr(X_j, \mathbf{r}) = \arg \max_{k \in \{1, \cdots, p\}} \mid X_j^T \mathbf{r} \mid / \mid \mathbf{r} \mid .$$

The corresponding predictor forms the active set $\mathcal{A} = \{k^*\}$ and its complement $\mathcal{A}^c = \{1, \cdots, p\} \backslash \mathcal{A}$.

3. (Build a cluster) For the selected predictor X_{k^*}, build a cluster $\mathcal{C}_{k^*} \subseteq \{1, \cdots, p\}$ such that

$$\mathcal{C}_{k^*} = \mathcal{C}_{k^*}(\rho) = \{j \ : \ \mid X_{k^*}^T X_j \mid > \rho, \ \forall j \in \mathcal{A}^c\} \cup \{k^*\},$$

for a given ρ. Update the active set $\mathcal{A} = \mathcal{A} \cup \mathcal{C}_{k^*}$ and its complement $\mathcal{A}^c = \mathcal{A}^c \backslash \mathcal{C}_{k^*}$.

Remark : Please see Section 1.2.4.2 for the decisions of ρ value. While the value of ρ in Example 11 is selected by a Cross-Validation, the values of ρ in Examples 12-20 are 0.9, 0.8, $\cdots$, 0.1, respectively to investigate the sensitivity of ρ selection. As discussed in Section 1.3, the performance of the PR-SGR for a fixed representative selection rule does not change much for large enough ρ values ($\rho \geq 0.3$), and the PR-SGR with the MAX representative selection rule is the least sensitive to the changes in ρ among all the representative selection rules.

4. (Determine a representative variable) Select a predictor $X_i, \ i \in \mathcal{C}_{k^*}$ as a representative of the cluster $\mathcal{C}_{k^*}$ based on a selection rule described in Section 1.2.1.2, and form the representative index set $\mathcal{R} = \{i\}$.

5. Repeat the following steps while $\mid \mathcal{R} \mid \leq n$ and $\sum_{j \in \mathcal{R}} \mid X_j^T \mathbf{r} \mid > 0$.

   (a) Let $\mathbf{u}_{\mathcal{R}}$ be the unit vector toward the least squares direction of $\{X_j \ : \ j \in \mathcal{R}\}$

(b) If $\mathcal{A}^c \neq \emptyset$, then solve how much to extend $\mathbf{u}_\mathcal{R}$ so that the correlations with the representatives decrease to zero. That is, find the solution $\hat{\lambda}$ of the following equation: for $\lambda > 0$ and any $j \in \mathcal{R}$,

$$X_j^T(\mathbf{r} - \lambda \mathbf{u}_\mathcal{R}) = 0.$$

(c) For every $X_k,\ k \in \mathcal{A}^c \neq \emptyset$, solve how much to extend $\mathbf{u}_\mathcal{R}$ so that the correlation with the predictor X_k is the same as those $X_j,\ j \in \mathcal{R}$. This is to find a solution $\hat{\lambda}$ of the following equation for all $X_k,\ k \in \mathcal{A}^c$ and $\lambda > 0$:

$$X_k^T(\mathbf{r} - \lambda \mathbf{u}_\mathcal{R}) = X_j^T(\mathbf{r} - \lambda \mathbf{u}_\mathcal{R}),$$

for any $j \in \mathcal{R}$. Among the solution $\hat{\lambda}$'s, choose the smallest positive value, and denote it $\hat{\lambda}^*$. Let j^* be the index of the corresponding predictor. Now, the active set $\mathcal{A}$ extends to $\mathcal{A} = \mathcal{A} \cup \{j^*\}$.

(d) For X_{j^*}, establish a cluster $\mathcal{C}_{j^*}$ and enlarge the active set $\mathcal{A} = \mathcal{A} \cup \mathcal{C}_{j^*}$. Then, $\mathcal{A}^c = \mathcal{A}^c \backslash \mathcal{C}_{j^*}$. Select a representative $X_i,\ i \in \mathcal{C}_{j^*}$, and enlarge the representative index set $\mathcal{R} = \mathcal{R} \cup \{i\}$.

(e) Compute the residuals: $\mathbf{r} = \mathbf{r} - \hat{\lambda}^* \mathbf{u}_\mathcal{R}$.

The unit vector $\mathbf{u}_\mathcal{R}$ in Step 5 can be simplified as Efron *et al.* ([9]) did. Let $\mathbf{X}_\mathcal{R}$ denote the submatrix of $\mathbf{X}$ for the representatives corresponding to the representative index set $\mathcal{R}$. Then, $\mathbf{u}_\mathcal{R} = \mathbf{A}_\mathcal{R} \mathbf{X}_\mathcal{R} \mathbf{G}_\mathcal{R}^{-1} \mathbf{c}_\mathcal{R}$, where

$$\mathbf{c}_\mathcal{R} = \mathbf{X}_\mathcal{R}^T \mathbf{r},\ \mathbf{G}_\mathcal{R} = \mathbf{X}_\mathcal{R}^T \mathbf{X}_\mathcal{R},\ \mathbf{A}_\mathcal{A} = (\mathbf{c}_\mathcal{R}^T \mathbf{G}_\mathcal{R}^{-1} \mathbf{c}_\mathcal{R})^{-1/2}.$$

If the size of each clusters, $\mid \mathcal{C}_k \mid$, is one, then the PR-SGR Algorithm is equivalent to the Lars Algorithm and $\mathcal{A} = \mathcal{R}$.

Remark : The entire sequence of steps of the Lars Algorithm ([9]) with $p \gg n$ requires $O(n^3)$. Since the PR-SGR uses a clustering algorithm in each step and the

clustering algorithm scan at most p predictors to build a cluster, the computational complex of the PR-SGR is $O(n^3 p \log p)$.

Remark : The Lars Algorithm ([9]) works as follows: at the beginning, it always puts a predictor that is the most correlated with the response into the active set to contain the indices of the predictors selected in the model. The coefficient of each predictor in the active set increases in the direction of the sign of its correlation with the response, until a predictor, currently not in the active set, is as much correlated with the current residual as the predictor in the active set is. When two predictors are entered into the active set, their coefficients move in their joint least square direction until another predictor, not in the active set, is as much as correlated with the current residual based on the two selected predictors. A new predictor is added to the model in a subsequent step. The algorithm continues until either the size of the active set reaches the sample size n, or until all the predictors are selected in the model and the model attains the ordinary least squared fit. As the Lars algorithm proceeds, all the predictors in the active set carry the same correlation with the current residual.

1.2.4.2 Tuning Parameters

As discussed in Section 1.2.3, the parameter ρ (or m), w, and λ (or t) can be estimated directly by applying the PR-SGR Algorithm on a given validation dataset and minimizing the estimate of the prediction error. If only a training dataset is available, $k-$fold Cross-Validation ($k-$CV) can be used to select the parameter value in a three-dimensional surface way. In Section 1.3, a validation dataset is generated for each setting to choose the tuning parameters resulting in the smallest estimated prediction error. The real-life case study uses $5-$CV to select the ρ value due to its small sample size, while the simulation studies use $10-$CV.

1.3 Simulation Study

A simulation study is carried out to evaluate the performance of the PR-SGR Algorithm under various conditions. Each dataset is simulated from the regression model $\mathbf{y} = \mathbf{X}\beta + \sigma\epsilon,\ \epsilon \sim N(0, I)$, where $\sigma > 0$. For each simulation setting, 100 datasets are generated. Each dataset consists of a training dataset of size n_1, an independent validation dataset of size n_2, and an independent test dataset of size n_3. Regression models are fitted on the training datasets, and the validation datasets are used to select the tuning parameters yielding the lowest prediction error. If tuning parameters are selected on the training dataset, the selected model tends to contain n predictors to minimize the prediction error. Therefore, the tuning parameters are selected on the validation dataset to minimize overfitting. We compute test errors, by the Mean-Squared Errors (MSE), on the test datasets and the model complexities by the number of nonzero coefficients. Each simulation setting uses $n_1 = 20$, $n_2 = 20$, $n_3 = 20$, $\sigma = 5$ and $p = 40$ where the number of predictors is p. Each predictor is generated from the standard normal distribution.

Twenty scenarios are considered. The first four, Examples 1-4, consider the cases with one cluster of the identical predictors but with different number of nonzero β_i's. The purpose of changing the number of nonzero coefficients is to compare model complexities in terms of the number of nonzero estimated coefficients in the final model. The effect of the representative selection rules can be removed by using only one cluster of all the identical predictors. The next four, Examples 5-8, consider the cases with one cluster of highly correlated predictors. By changing the minimum correlation ρ_0 among the predictors in the cluster, the effect of the representative selection rules can be studied. The next three, Examples 9-11, consider the cases with two clusters of highly correlated predictors. In Example 9, one fixed ρ_0 value is used, so that the effect of the number of clusters can be studied. In Examples 10 and 11, the value of ρ_0 is varied. The last nine, Examples 12-20, have the same settings

with Example 11, but use different ρ values to build clusters. While the value of ρ in Example 11 is selected by 10−CV, the values of ρ in Examples 12-20 are fixed as 0.9, 0.8, $\cdots$, 0.1, respectively. Here, the sensitivity of ρ selection is investigated. The details of twenty scenarios are as follows:

- In Example 1, a cluster consisting of five identical predictors $X_1, X_2, \cdots, X_5$ is generated. The other predictors not in the cluster are assumed to be identically distributed. $\beta = (3, \underbrace{0, \cdots, 0}_{4}, 1.5, 2, 4, 5, \underbrace{0, \cdots, 0}_{31})^T$ with five nonzero β_i's.
- Example 2 uses the same setting as in Example 1 except for β in which the number of nonzero coefficients is set to be 10: $\beta = (3, \underbrace{0, \cdots, 0}_{4}, \underbrace{2, \cdots, 2}_{9}, \underbrace{0, \cdots, 0}_{26})^T$.
- In Example 3, the same setting as in Example 1 except for β is used. The number of nonzero coefficients is set to be 20: $\beta = (3, \underbrace{0, \cdots, 0}_{4}, \underbrace{2, \cdots, 2}_{19}, \underbrace{0, \cdots, 0}_{16})^T$.
- In Example 4, we use the same setting as in Example 1 except for β. The number of nonzero coefficients in β is 30: $\beta = (3, \underbrace{0, \cdots, 0}_{4}, \underbrace{2, \cdots, 2}_{29}, \underbrace{0, \cdots, 0}_{6})^T$.
- Example 5 has one cluster of three highly correlated predictors X_1, X_2 and X_3 whose correlations with other predictors in the cluster are greater than 0.9. $\beta = (3, 0, 0, 1.5, 2, 4, 5, \underbrace{0, \cdots, 0}_{33})^T$ has only 5 nonzero β_i's. The other predictors not in the cluster are assumed to be identically distributed and correlated with the predictors in the cluster with less than 0.9 correlation.
- Example 6 also has one cluster of three highly correlated predictors as in Example 5, but β is set to have 10 nonzero β_i's: $\beta = (3, 0, 0, \underbrace{2, \cdots, 2}_{9}, \underbrace{0, \cdots, 0}_{28})^T$.
- In Example 7, the same setting as in Example 5 except for β is used. The number of nonzero coefficients is 20: $\beta = (3, 0, 0, \underbrace{2, \cdots, 2}_{19}, \underbrace{0, \cdots, 0}_{18})^T$.

- In Example 8, we use the same setting as in Example 5 except for β. The number of nonzero coefficients in β is increased up to 30: $\beta = (3, 0, 0, \underbrace{2, \cdots, 2}_{29}, \underbrace{0, \cdots, 0}_{8})^T$.

- In Example 9, a cluster is generated to have three highly correlated predictors X_1, X_2, and X_3 whose correlations among themselves are set to be greater than 0.9. Another cluster is generated that contains two highly correlated predictors X_4 and X_5 whose correlation is set to be greater than 0.9. We set the correlations between the predictors from different clusters less than 0.9. All the other predictors not in the clusters are assumed to be identically distributed with 10 nonzero β_i's: $\beta = (3, 0, 0, 1.5, 0, \underbrace{2, \cdots, 2}_{8}, \underbrace{0, \cdots, 0}_{27})^T$

- Example 10 uses the same setting as in Example 9 except that the minimum correlations in the clusters are set to be 0.8.

- Example 11 also uses the same setting as in Example 9 except that the minimum correlation was 0.9 for one cluster and 0.8 for the other.

- Examples 12-20 use the same setting as in Example 11 but with different ρ values to build clusters. Selected ρ values for each example are 0.9, 0.8, 0.7, $\cdots$, 0.1 respectively.

Tables 4 - 8 and Figures 5 - 9 (box plots) summarize the prediction results in terms of MSE values. The PR-SGR with various representative selection rules has better performance in most of examples than the ones from the Lasso/Lars, Elastic-Net, grouped-Lasso or grouped-Lars. The median values of MSEs for the PR-SGR with different representative selection rules are either the best or second best in all examples. The PR-SGR with some representative selection rules behaves almost identically or similarly with the Lasso/Lars in some examples (see Example 7). Table 9 shows the reduction rates of the PR-SGR with all the representative selection rules. The

maximum reduction rate for the PR-SGR is 13.12 % and the PR-SGR outperforms 11 times out of 15 examples against the Lasso/Lars and other methods.

Tables 4 - 7 present the model complexity of each simulation setting in terms of the number of nonzero coefficients in the final model. They show that the PR-SGR tends to select smaller number of predictors than other methods do. Figure 10 shows that the model complexity tends to increase as the number of nonzero coefficients increases, and that the PR-SGR with the MAX representative selection rule tends to select the smallest number of predictors in the final model.

The results of Example 5-11 evaluate the effect of the different representative selection rules. Table 10 summarizes the average values of the median MSEs and the average ranks based on the median values for all the representative selection rules. Table 10 also summarizes the average values of the model complexities and the average ranks based on the model complexities for the PR-SGR with all the representative selection rules. To get the values in Table 10, we first calculate the ranks of the interesting values (either the median MSEs or the model complexities) for all the representative selection rules in each example, and then calculate the average ranks over all the examples for each representative selection rule. Table 10 shows that the RAN representative selection rule is the best and the MED representative selection rule is the second best, based on the median MSEs. Based on the model complexities, the MIN representative selection rule is the best and the MED representative selection rule is the second best. However, if we consider the ranks, there is no difference among the representative selection rules. In this comparison, we exclude Examples 1-4 and 12-20, because Examples 1-4 consider only one cluster of identical predictors trivially resulting in no difference among the representative selection rules, and Examples 12-20 are conducted to compare the sensitivity to the selection of ρ.

We compare the performance of the PR-SGR with other methods in Examples 13-16. The value of ρ is selected by 10-fold CV in Example 11, whereas the ρ values

Table 4: Statistics of MSE and model complexity in Examples 1 - 3 for the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules.

Example 1									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	40.20	54.13	38.20	38.26	37.32	36.78	37.11	35.59	39.17
$q_{0.25}$	53.02	78.05	47.39	52.37	45.35	47.00	45.67	41.87	46.85
$q_{0.50}$	73.45	97.40	64.78	69.46	61.83	60.18	62.23	62.58	67.62
$q_{0.75}$	90.11	150.78	83.93	94.31	84.78	79.97	82.12	82.34	85.16
$q_{0.90}$	147.96	178.64	103.48	140.88	110.94	96.35	105.15	101.99	109.59
s.e	11.41	12.01	9.15	11.72	9.23	8.42	9.10	9.27	9.64
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	8.24	7.47	7.23	7.29	6.15	7.10	6.99	7.02	7.29
s.e	2.76	2.01	2.41	2.84	1.97	2.01	1.94	2.00	2.12

Example 2									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	49.22	43.30	42.34	42.24	43.88	46.11	49.55	41.34	47.83
$q_{0.25}$	57.39	56.72	56.32	56.35	54.67	56.69	54.98	54.81	54.84
$q_{0.50}$	71.54	69.97	73.10	72.99	68.52	71.67	74.15	70.03	72.06
$q_{0.75}$	102.91	97.28	93.96	101.88	88.33	92.26	93.38	90.28	94.03
$q_{0.90}$	131.79	125.29	120.54	130.12	109.91	115.18	112.86	112.93	110.67
s.e	10.94	9.46	9.69	11.21	8.81	8.97	9.00	9.85	9.00
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	8.40	7.24	8.24	7.21	7.59	7.01	7.54	7.18	7.66
s.e	2.29	2.04	2.03	1.99	1.90	1.41	1.22	1.56	1.88

Example 3									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	70.49	60.36	62.92	71.37	68.22	69.95	66.94	70.86	67.09
$q_{0.25}$	87.72	84.86	87.32	89.25	85.52	84.66	87.93	83.83	81.41
$q_{0.50}$	109.38	105.22	107.30	113.77	101.58	107.63	107.38	107.97	104.93
$q_{0.75}$	144.95	123.23	130.87	144.72	125.57	133.48	129.94	138.22	127.06
$q_{0.90}$	216.67	145.72	174.73	195.50	168.27	154.54	160.64	179.58	150.01
s.e	13.58	10.35	11.50	13.27	11.00	10.95	10.96	11.06	10.20
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	9.77	9.24	8.99	8.24	8.61	8.75	9.02	8.93	9.07
s.e	3.04	2.95	3.01	2.82	2.94	2.75	2.51	2.78	2.07

Table 5: Statistics of MSE and model complexity in Examples 4 - 6 for the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules.

Example 4									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	91.35	82.10	96.72	97.10	85.57	89.14	91.20	93.80	93.01
$q_{0.25}$	120.73	101.68	110.70	117.29	114.07	112.27	110.81	115.57	113.98
$q_{0.50}$	159.63	121.98	133.92	149.74	131.13	134.04	135.09	130.81	137.05
$q_{0.75}$	215.05	164.23	192.14	203.67	182.88	183.73	185.95	171.37	172.18
$q_{0.90}$	262.98	215.98	247.62	235.53	228.50	209.97	219.98	205.89	227.13
s.e	14.64	12.55	13.81	15.40	13.14	12.70	13.28	11.51	12.91
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	13.04	12.99	13.07	12.91	13.04	13.42	12.96	12.51	13.04
s.e	4.98	5.01	4.86	4.92	5.02	5.28	4.85	4.91	4.86
Example 5									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	33.31	37.06	34.26	34.09	34.48	30.69	30.77	32.17	30.80
$q_{0.25}$	41.46	46.13	41.90	43.87	41.95	43.15	41.12	40.38	40.10
$q_{0.50}$	53.39	58.69	52.89	54.55	51.25	52.80	50.36	49.79	49.52
$q_{0.75}$	75.76	75.99	73.03	69.30	63.92	66.62	65.58	60.05	63.03
$q_{0.90}$	117.17	105.06	98.25	96.88	80.05	79.60	76.06	75.95	78.46
s.e	12.31	9.05	8.95	12.37	8.58	8.16	7.68	7.50	7.27
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	8.01	7.56	7.28	7.12	6.62	7.23	7.04	7.61	7.30
s.e	1.96	2.06	2.23	2.15	1.83	2.11	2.27	2.02	2.28
Example 6									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	46.64	47.75	46.73	47.16	44.34	45.84	45.06	46.70	46.72
$q_{0.25}$	58.39	61.10	59.82	54.67	55.08	54.40	58.08	54.30	58.84
$q_{0.50}$	76.85	73.08	72.99	75.80	68.56	68.70	69.54	66.93	69.41
$q_{0.75}$	104.86	92.37	87.21	101.83	83.40	80.86	78.56	81.05	79.29
$q_{0.90}$	143.34	111.43	110.85	137.73	96.81	93.47	95.07	97.48	95.60
s.e	10.98	8.77	9.68	10.72	8.15	8.49	8.02	9.87	8.80
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	8.08	7.86	8.01	7.83	7.79	7.38	7.51	7.63	7.86
s.e	2.27	1.98	1.93	1.85	1.82	1.68	1.70	1.86	1.90

Table 6: Statistics of MSE and model complexity in Examples 7 - 9 for the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules.

Example 7									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	76.11	68.68	68.51	76.32	69.08	62.72	72.32	67.91	66.50
$q_{0.25}$	93.04	87.30	89.06	97.40	86.27	88.19	88.14	92.95	87.18
$q_{0.50}$	122.72	106.33	119.84	125.39	118.64	113.06	117.50	112.35	115.58
$q_{0.75}$	155.41	135.84	155.58	161.71	143.56	146.27	146.07	142.97	142.84
$q_{0.90}$	210.43	167.07	176.14	235.44	168.76	174.21	171.15	166.08	164.09
s.e	12.92	10.86	11.46	13.07	10.96	11.69	11.12	10.47	10.50
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	9.14	9.35	8.18	8.56	9.01	8.98	8.75	9.04	9.34
s.e	2.83	2.89	2.76	2.96	2.83	2.74	2.64	2.70	2.54

Example 8									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	90.87	88.86	92.94	92.04	86.58	92.72	95.53	92.25	82.54
$q_{0.25}$	117.10	105.62	118.25	118.20	115.63	113.88	112.52	111.45	114.53
$q_{0.50}$	162.50	133.56	150.75	160.32	144.40	144.68	143.75	144.68	146.24
$q_{0.75}$	202.85	166.43	187.07	204.87	185.46	186.43	182.22	182.54	183.36
$q_{0.90}$	242.28	195.23	229.28	239.60	222.36	213.98	211.59	211.32	204.64
s.e	13.85	11.09	12.71	12.81	12.54	12.31	12.14	12.57	11.81
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	13.01	12.88	13.24	12.90	12.83	13.01	12.84	12.78	13.03
s.e	4.10	4.75	4.65	4.98	4.62	4.35	4.98	4.86	4.24

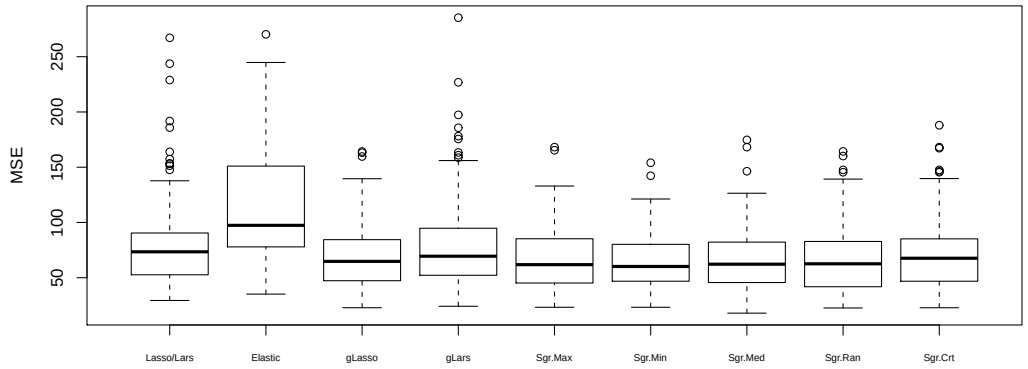
Example 9									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	49.65	47.55	51.80	50.29	47.12	49.68	45.64	46.28	45.63
$q_{0.25}$	58.29	57.25	62.03	65.25	54.26	55.50	54.96	55.78	56.52
$q_{0.50}$	71.03	70.38	76.24	79.75	68.85	69.10	67.69	66.10	68.15
$q_{0.75}$	95.78	89.04	94.09	108.81	84.15	87.70	83.14	81.38	85.64
$q_{0.90}$	127.35	107.58	117.56	140.36	103.76	114.54	102.52	105.43	101.91
s.e	10.76	8.20	9.65	10.84	9.15	9.61	8.06	8.84	8.73
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	8.28	7.90	8.21	7.86	7.86	7.23	8.15	7.93	8.01
s.e	1.97	1.87	1.83	1.75	1.62	1.72	1.80	1.63	1.78

Table 7: Statistics of MSE and model complexity in Examples 10 - 12 for the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules.

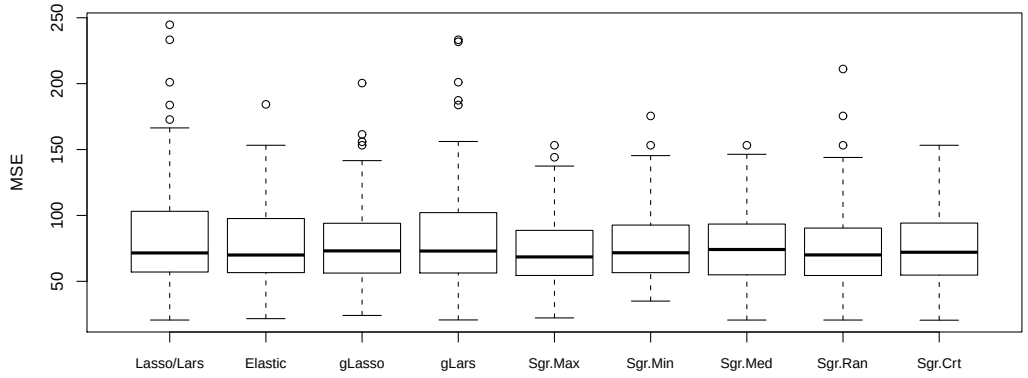
Example 10									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	49.31	48.50	50.91	50.25	48.38	46.58	44.79	47.04	47.47
$q_{0.25}$	61.74	59.95	64.00	62.45	57.83	57.19	54.50	53.27	57.68
$q_{0.50}$	79.46	78.83	81.70	81.81	71.57	75.16	70.43	71.60	72.66
$q_{0.75}$	103.57	96.99	102.12	117.77	87.98	98.70	89.09	91.95	92.83
$q_{0.90}$	165.10	124.30	135.95	165.78	102.56	124.97	106.83	115.97	124.78
s.e	12.66	9.67	9.95	12.76	8.95	10.12	9.06	9.67	11.65
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	8.36	8.01	8.23	7.98	8.03	8.15	7.93	7.68	7.80
s.e	1.98	1.76	1.88	1.93	1.93	1.87	1.87	1.63	1.70
Example 11									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	48.16	48.17	48.15	53.57	42.04	44.86	43.45	44.77	43.87
$q_{0.25}$	55.77	58.75	59.61	59.55	53.21	53.87	52.70	51.49	56.48
$q_{0.50}$	74.83	76.03	74.57	80.80	67.53	68.90	68.80	67.57	67.92
$q_{0.75}$	98.96	90.41	104.58	104.51	87.36	83.52	89.15	85.30	84.13
$q_{0.90}$	137.70	109.26	118.59	126.61	105.28	115.23	107.67	102.98	107.52
s.e	12.32	9.26	9.31	9.60	8.33	10.17	8.57	8.95	8.66
Model					PR-SGR				
Complexity	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
mean	8.04	7.68	8.31	7.95	7.88	7.24	7.66	7.36	7.91
s.e	2.04	1.96	1.98	2.01	1.97	1.55	1.62	1.41	1.90
Example 13									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	45.10	46.59	48.77	54.71	41.99	43.55	42.83	40.15	47.15
$q_{0.25}$	56.66	60.69	58.72	63.00	53.42	54.87	51.92	52.52	58.59
$q_{0.50}$	77.26	84.48	76.85	88.41	68.37	71.89	66.77	71.70	72.28
$q_{0.75}$	100.09	108.41	98.59	113.01	87.11	93.91	85.32	93.02	89.84
$q_{0.90}$	148.64	132.15	130.96	163.15	102.86	116.94	102.71	105.46	114.00
s.e	12.56	9.92	10.69	13.91	8.73	10.13	8.66	9.81	11.26

Table 8: Statistics of MSE and model complexity in Examples 14-16 for the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules.

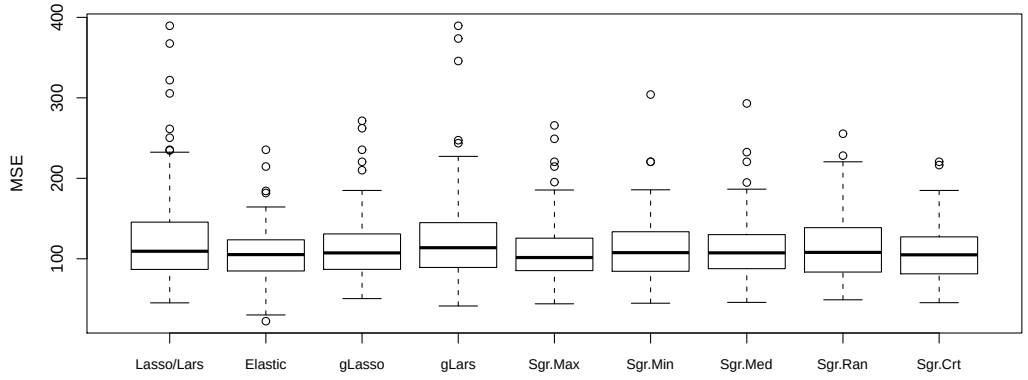
Example 14									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	42.53	46.78	44.62	46.61	40.31	42.73	40.53	42.32	44.34
$q_{0.25}$	55.43	58.98	56.09	58.52	52.95	55.78	51.01	51.07	54.91
$q_{0.50}$	66.36	77.23	67.44	69.94	64.92	67.78	63.52	66.02	65.36
$q_{0.75}$	89.98	92.01	88.13	99.15	77.22	81.59	79.98	84.11	78.77
$q_{0.90}$	128.75	103.65	117.81	158.50	92.10	92.85	95.15	100.61	94.33
s.e	11.59	8.64	10.06	12.34	8.13	10.07	8.33	9.44	8.23
Example 15									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	44.27	53.46	48.63	45.90	43.71	45.11	45.64	43.45	44.01
$q_{0.25}$	58.16	62.58	62.48	60.49	57.63	57.36	56.49	53.99	55.14
$q_{0.50}$	73.01	78.82	77.02	83.54	73.17	71.01	70.57	70.17	70.10
$q_{0.75}$	95.90	98.43	106.65	103.87	90.70	85.44	86.43	88.89	86.70
$q_{0.90}$	136.42	116.44	135.53	153.51	113.77	111.94	113.81	107.37	106.47
s.e	11.95	8.87	10.32	11.25	8.95	9.26	8.92	11.13	9.80
Example 16									
					PR-SGR				
MSE	Lasso/Lars	Elastic	gLasso	gLars	MAX	MIN	MED	RAN	CRT
$q_{0.10}$	44.06	45.21	44.06	45.15	45.92	44.32	44.03	44.32	42.83
$q_{0.25}$	61.13	57.89	60.12	61.69	55.34	54.46	54.35	55.93	54.96
$q_{0.50}$	80.43	73.93	77.88	81.19	71.29	74.15	70.09	77.52	72.66
$q_{0.75}$	104.96	89.04	96.50	101.71	85.00	94.19	85.86	94.45	86.64
$q_{0.90}$	134.09	110.85	118.09	150.32	102.53	109.08	103.37	109.46	117.93
s.e	11.88	9.45	8.91	11.70	8.07	9.49	9.13	9.36	10.22



(a) Example 1

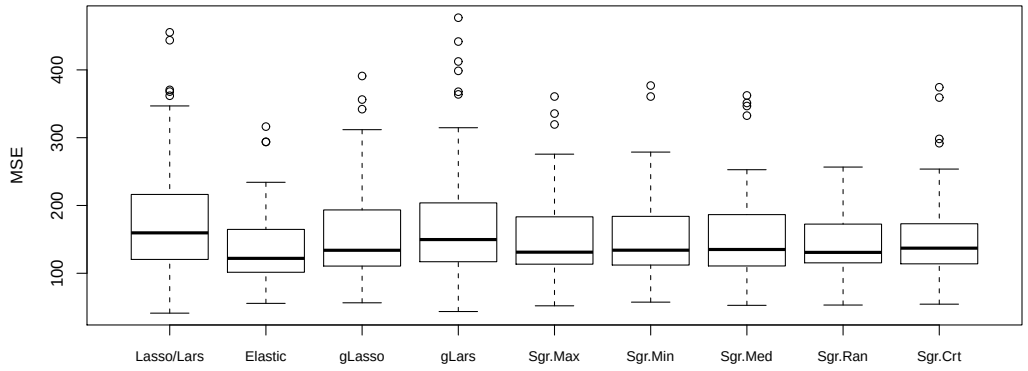


(b) Example 2

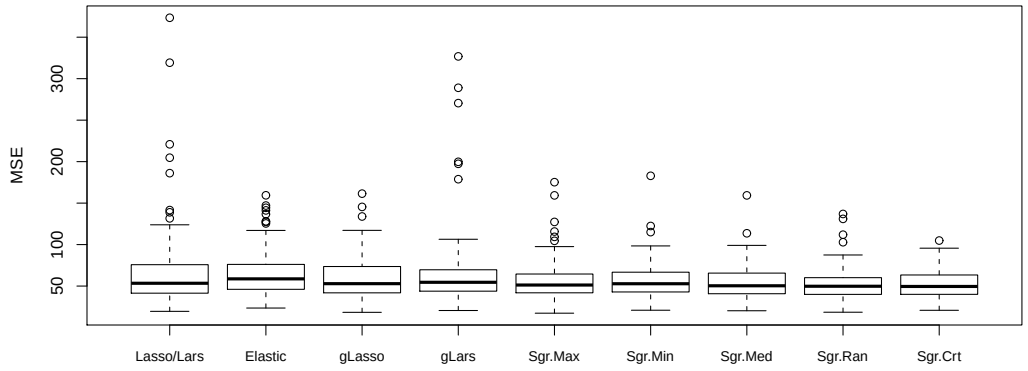


(c) Example 3

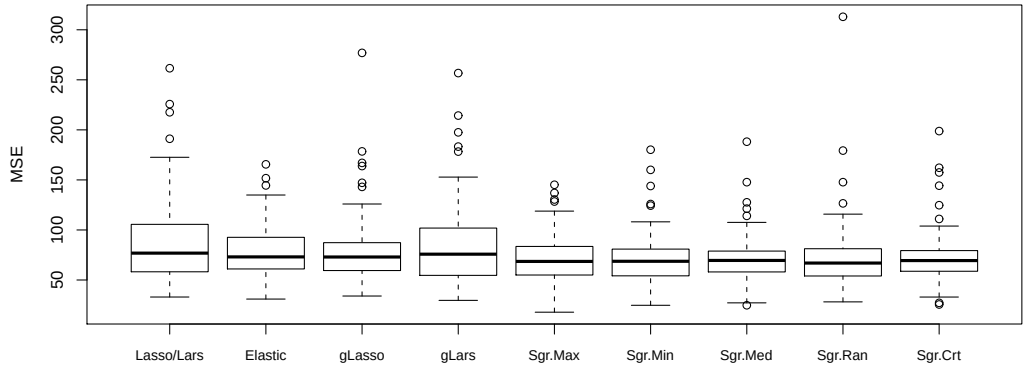
Figure 5: Comparing the accuracy of prediction of the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules for Examples 1 - 3.



(a) Example 4

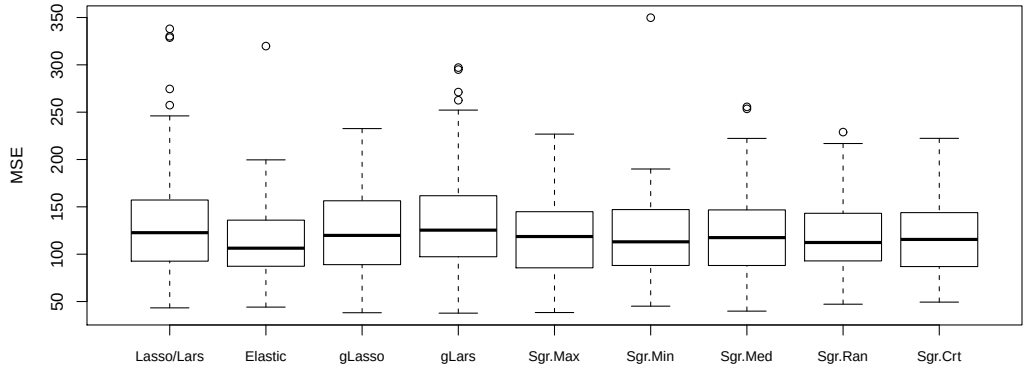


(b) Example 5

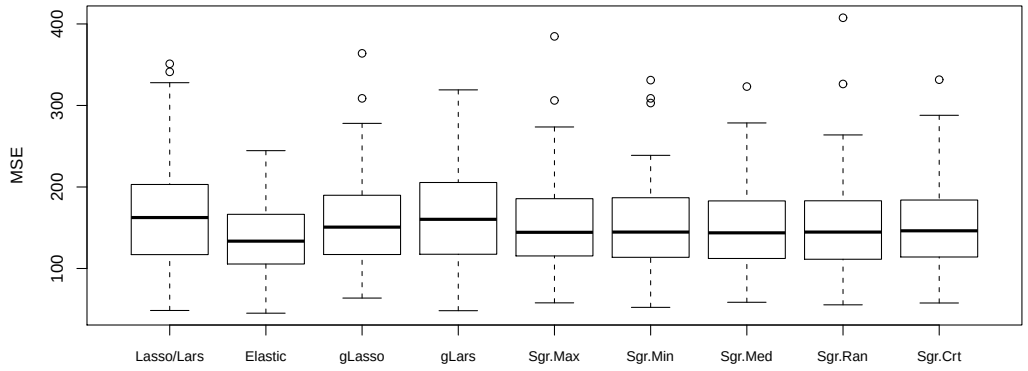


(c) Example 6

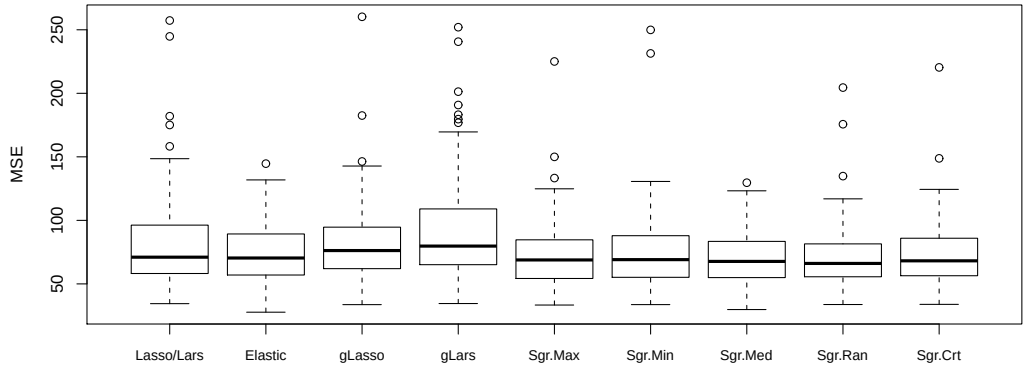
Figure 6: Comparing the accuracy of prediction of the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules for Examples 4 - 6.



(a) Example 7

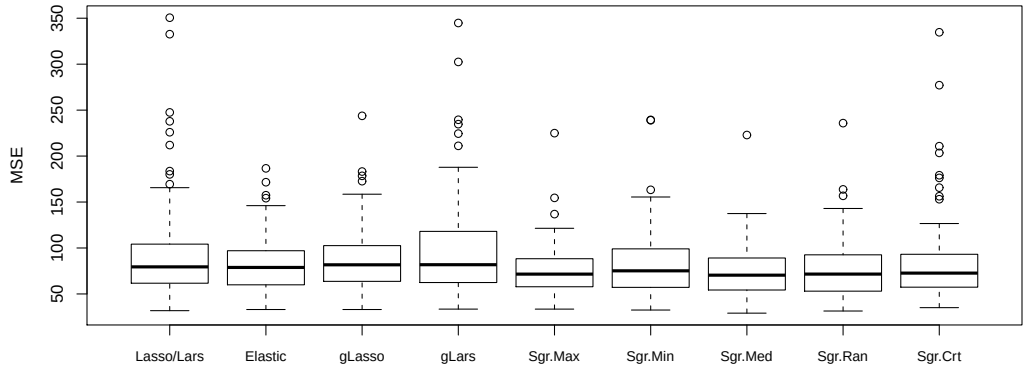


(b) Example 8

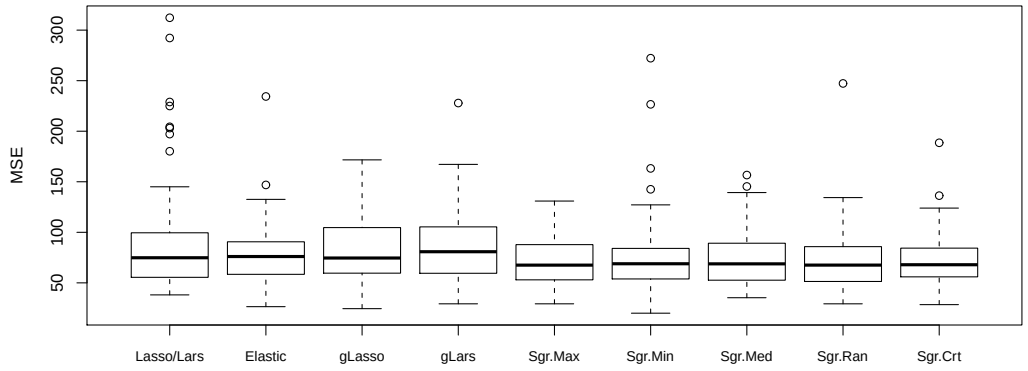


(c) Example 9

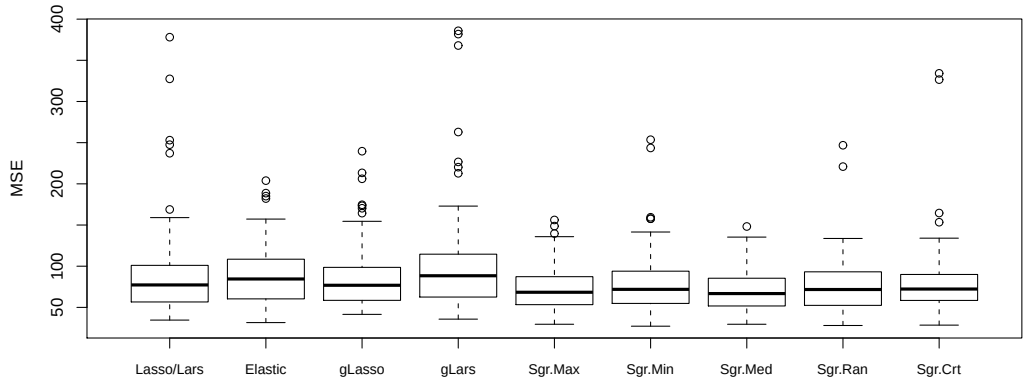
Figure 7: Comparing the accuracy of prediction of the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules for Examples 7 - 9.



(a) Example 10

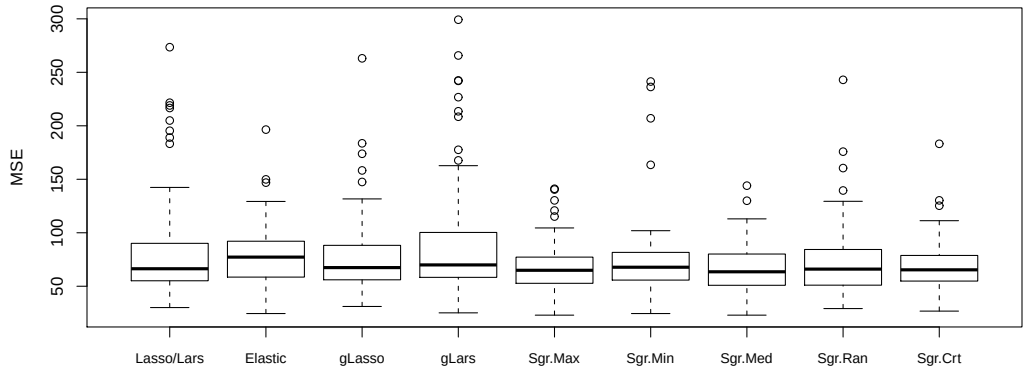


(b) Example 11

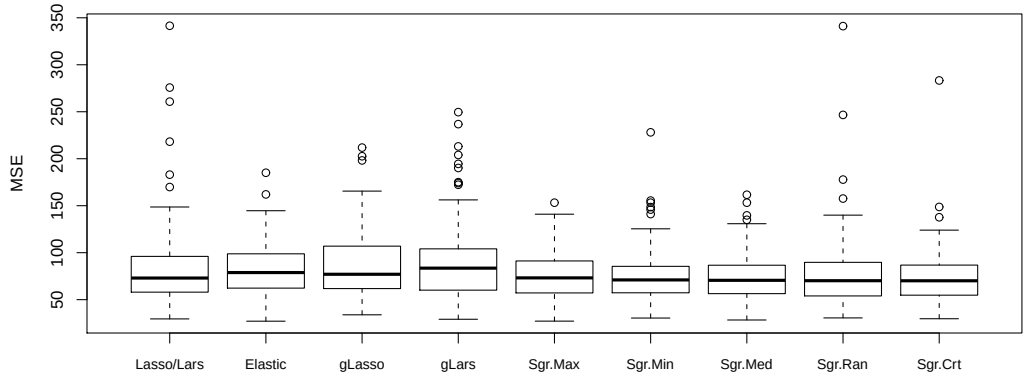


(c) Example 13

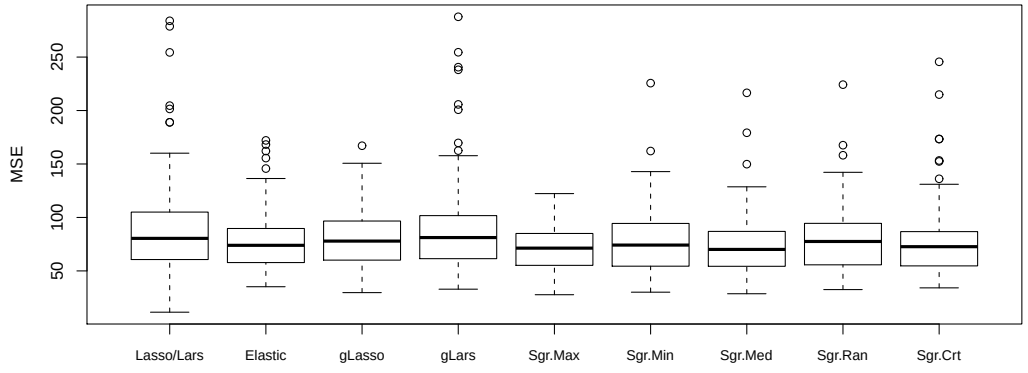
Figure 8: Comparing the accuracy of prediction of the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules for Examples 10, 11 and 13.



(a) Example 14



(b) Example 15



(c) Example 16

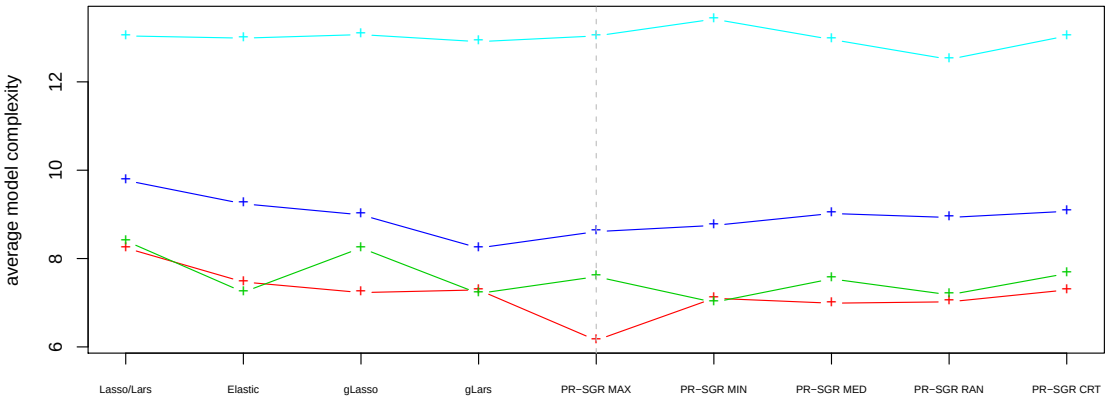
Figure 9: Comparing the accuracy of prediction of the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars) and PR-SGR with 5 representative selection rules for Examples 14 - 16.

Table 9: Reduction rates based on the median MSEs for the best method among the Lasso/Lars, Elastic-Net, grouped-Lasso and grouped-Lars and the best version of the PR-SGR in each example.

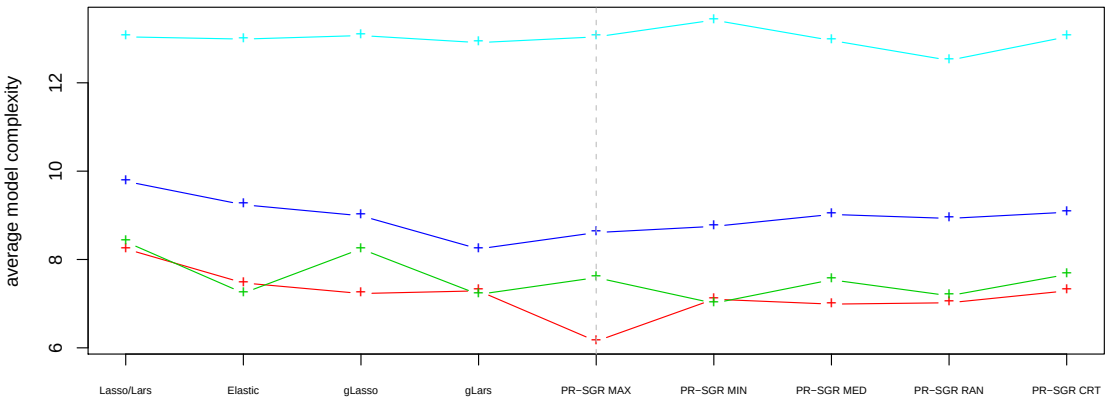
Example	min(median MSEs)		Reduction
	Others	PR-SGR	Rate(%)
1	64.78	68.52	-5.77
2	69.97	68.52	2.07
3	105.22	101.58	3.46
4	121.98	130.81	-7.24
5	52.89	49.52	6.37
6	72.99	66.93	8.30
7	106.33	112.35	-5.66
8	133.56	143.75	-7.63
9	70.38	66.10	6.08
10	78.83	70.43	10.66
11	74.57	67.53	9.44
13	76.85	66.77	13.12
14	66.36	63.52	4.28
15	73.01	70.01	4.11
16	73.93	70.09	5.19

Table 10: Average values of the median MSEs, average ranks based on the median MSEs, average values of the model complexities, and average ranks based on the model complexities for all the representative selection rules of the PR-SGR.

Average Value					
PR-SGR	MAX	MIN	MED	RAN	CRT
median(MSE)	75.88	75.82	75.28	74.12	75.64
model complexity	8.57	8.46	8.55	8.58	8.75
Rank					
PR-SGR	MAX	MIN	MED	RAN	CRT
median(MSE)	1.40	1.40	1.40	1.40	1.40
model complexity	1.40	1.40	1.40	1.40	1.40



(a) Example 1-4



(b) Example 5-8

Figure 10: Comparison of the model complexity of the Lasso/Lars, Elastic-Net, grouped-Lasso (gLasso), grouped-Lars (gLars), and PR-SGR with various representative selection rules for Examples 1-8. Red line for five nonzero coefficients, green for 10 , blue for 20 and cyan for 30.

Table 11: Reduction rates based on the median MSEs for the PR-SGR with ρ selected by 10−CV and the PR-SGR with a fixed ρ ($\rho = 0.9,\ 0.8,\ \cdots, 0.1$).

	Median MSE					Reduction Rate				
ρ	MAX	MIN	MED	RAN	CRT	MAX	MIN	MED	RAN	CRT
CV	67.53	68.90	68.80	67.57	67.92	0.00	0.00	0.00	0.00	0.00
0.9	71.93	69.94	70.89	69.01	70.35	-6.52	-1.51	-3.04	-2.13	-3.58
0.8	68.37	71.89	66.77	71.70	72.28	-1.24	-4.34	2.95	-6.11	-6.42
0.7	64.92	67.78	63.52	66.02	65.36	3.86	1.63	7.67	2.29	3.77
0.6	73.17	71.01	70.57	70.17	70.10	-8.35	-3.06	-2.57	-3.85	-3.21
0.5	71.29	74.15	70.09	74.20	72.66	-5.57	-7.62	-1.88	-9.81	-6.98
0.4	68.26	71.54	72.66	73.70	69.40	-1.08	-3.83	-5.61	-9.07	-2.18
0.3	64.84	70.92	69.72	68.87	67.30	3.98	-2.93	-1.34	-1.92	0.91
0.2	71.39	75.48	80.43	76.43	74.58	-5.72	-9.55	-16.90	-13.11	-9.81
0.1	72.95	79.14	80.04	78.70	76.50	-8.03	-14.86	-16.34	-16.47	-12.63

used in Examples 13-16 are 0.8, 0.7, 0.6 and 0.5, respectively. Table 9 shows that the PR-SGR with those fixed ρ values still performs better than other competitors, resulting in the best reduction rate of 13.12%.

The sensitivity of the PR-SGR to the selection of ρ (0.9, 0.8, $\cdots$, 0.1) is examined in Examples 11-20. While the value of ρ is selected by 10-fold CV in Example 11, ρ values in Examples 12-20 are fixed as 0.9, 0.8, $\cdots$, 0.1, respectively. Table 11 provides the reduction rates of the PR-SGR for each representative selection rule for different ρ values. All the reduction rates for $\rho \geq 0.3$ are within ± 10 % for all the representative selection rules, when compared to the MSE with the ρ value selected by 10−CV. However, when $\rho = 0.2$ or 0.1, the PR-SGR tends to build a large cluster that contains many predictors, and thus the PR-SGR results in poor reduction rates for the small ρ's. Overall, the PR-SGR with the MAX representative selection rule is the least sensitive to the changes in ρ among all the representative selection rules.

Section 1.2.1.1 shows that the PR is equivalent to the Lasso when all the clusters contain only one predictor (i.e., $p_1 = \cdots = p_m = 1$). Similarly, the PR-SGR

Algorithm is shown to be equivalent to the Lars Algorithm in Section 1.2.4. Appendix B tests the performance of the PR-SGR Algorithm when all the predictors are assumed to be independent and when there is no group of highly correlated predictors. In this case, the Elastic-Net perform poorly among all the methods compared. Although there is no group of highly correlated predictors, the PR-SGR builds a cluster of some predictors with given ρ and uses a representative of the cluster in the model. Therefore, the PR-SGR loses the information contained in the predictors that are not selected as representatives. When ρ is selected by $10-$CV or ρ is fixed as $0.9,\ 0.8,\ 0.7,\ 0.6$ or 0.5, the PR-SGR for most ρ values performs better than the Lasso/Lars and Elastic-Net. For details, see Appendix B.

1.4 Real Data Study

The data in Table 12 for illustrating the proposed method are from the mental health diagnosis on depression ([6]). In the study, there were 15 patients and 1794 predictors of depression symptoms.

Figure 11 provides a graphical representation of the correlation matrix of the 40 predictors randomly selected. If a pairwise correlation is negative, its absolute value is taken. The magnitude of each pairwise correlation is presented by a block of blue-red scale image. The lowest pairwise correlation displayed in Figure 11 is 0.2851 and the highest correlation between two different predictors is 0.9944. Some of the pairwise correlations among 1794 predictors are as high as 0.9997. Since the dataset has only $n = 15$ samples and $p = 1794$ predictors ($p \gg n$) and contains many groups of highly correlated predictors, it shows a good illustration of the PR-SGR behavior.

In this experiment, the PR-SGR with all the representative selection rules (MAX, MIN, MED, RAN and CRT) are compared with the Lasso/Lars and Elastic-Net. The predictors are standardized and the response is centered to satisfy the initial settings described in Section 1.2.4.1. We estimate the average prediction errors by 10

Table 12: The data from the mental health diagnosis on depression. $p = 1794$ predictors of depression symptoms are examined for each of the $n = 15$ patients.

	V1	V2	V3	V4	$\cdots$	
Patient	FP1DAP1	FP2DAP1	FP3DAP1	FP4DAP1	$\cdots$	Response
1	-1.669598	-1.482294	-1.209087	-1.138544	$\cdots$	13
2	-2.827492	-2.901544	-3.929796	-3.714740	$\cdots$	8
3	-1.633667	-1.755907	-1.757161	-1.920382	$\cdots$	24
4	-1.974651	-2.019680	-3.246894	-2.952959	$\cdots$	17
5	-2.082676	-2.328255	-2.976314	-3.178645	$\cdots$	11
6	-1.892157	-1.253171	-2.578960	-2.372285	$\cdots$	14
7	-0.631407	-0.479997	-0.203112	0.012080	$\cdots$	9
8	-0.803371	-0.933103	-0.351946	-0.494320	$\cdots$	14
9	-1.564549	-1.789021	-1.305570	-1.590910	$\cdots$	-1
10	-2.891508	-2.914447	-3.100432	-3.224776	$\cdots$	11
11	-1.770761	-1.665579	-1.561414	-1.276832	$\cdots$	14
12	-1.745057	-2.057707	-2.795231	-2.314760	$\cdots$	15
13	-2.227326	-2.389061	-1.830231	-1.514457	$\cdots$	19
14	-3.036525	-2.916400	-4.182050	-3.863695	$\cdots$	22
15	-0.155327	0.136238	-1.012903	-0.461372	$\cdots$	20

replications of the 5−CV. For each CV, we fit a model on its training dataset with four-fifth of the observations (12 patients) and then calculated a prediction error on the test dataset (3 patients). Table 13 provides the average prediction errors on the test datasets. The PR-SGR with all the representative selection rules performs better than the Lasso/Lars and Elastic-Net. The PR-SGR with the MAX representative selection rule is the best, and the PR-SGR with the MED representative selection rule is the second best. The PR-SGR reduces its average prediction error at least 11.58 % compared to the one from the Lasso/Lars. The reduction of the prediction error from the PR-SGR with the best representative selection rule is 32.91 %, and from the second best representative selection rule is 30.74 %.

A regression model on the real-life data is fitted for the PR-SGR with the MAX and MED representative selection rules. The ρ value used to build clusters is chosen as 0.91 by 5−CV. The value of $s = t/\|\hat{\beta}^{OLS}\|_1$, which determines the number of

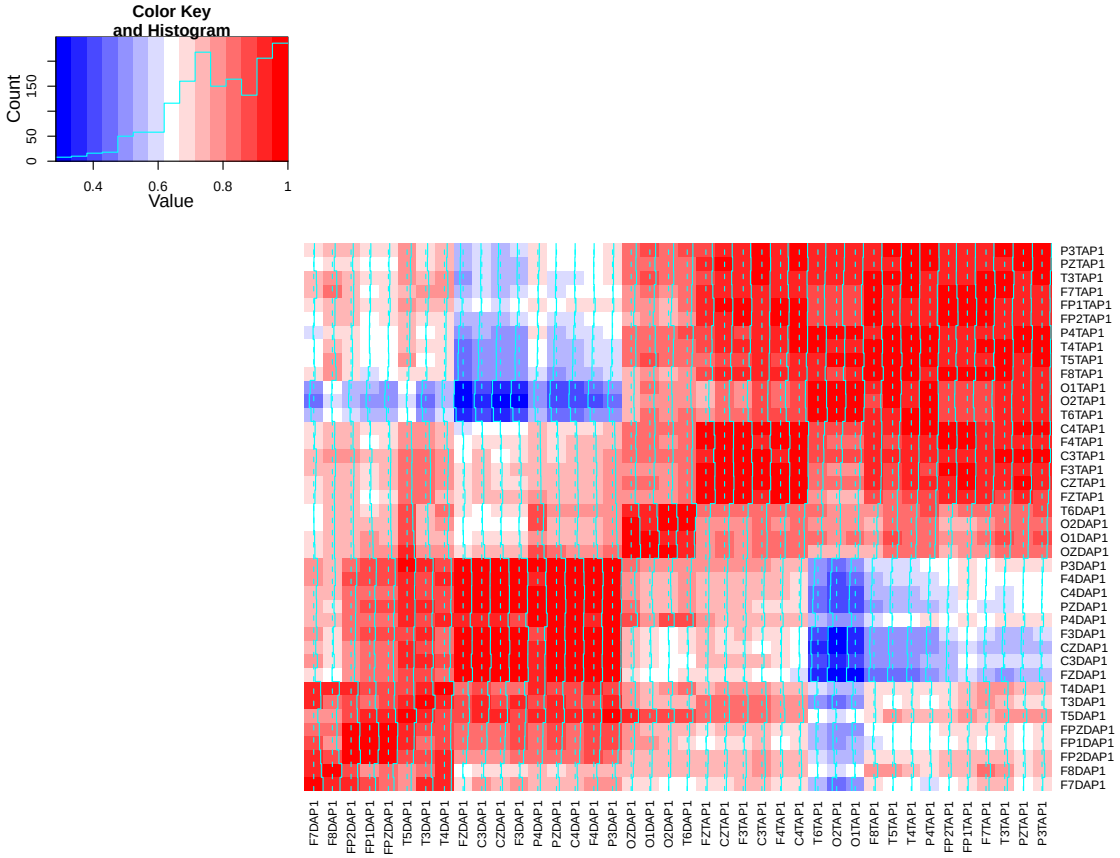


Figure 11: Graphical representation of the correlation matrix of the 40 predictors randomly selected from the real data. If a pairwise correlation is negative, we take its absolute value. The color scale is presented in the upper-left plot.

Table 13: Average prediction errors and their standard errors for the Lasso/Lars, Elastic-Net and PR-SGR with the five representative selection rules on the test datasets generated by 10 replications of the 5−CV. The reduction rate is calculated with the largest prediction error of the Lasso/Lars, so that the reduction rate for the Lasso/Lars is zero.

Method	Average Prediction Error	Standard Error	Reduction Rate (%)
Lasso/Lars	14.25	2.85	0.00
Elastic-Net	13.28	2.34	6.81
PR-SGR MAX	9.56	2.01	32.91
PR-SGR MIN	12.60	2.19	11.58
PR-SGR MED	9.87	2.02	30.74
PR-SGR RAN	10.79	2.16	24.28
PR-SGR CRT	11.77	2.13	17.40

nonzero coefficients in the final model, is also calculated by 5−CV. The final model fitted by the PR-SGR with the MAX representative selection rule includes nine predictors. Table 14 displays the clusters built, their representatives and their estimated coefficients. On the other hand, the final model fitted by the PR-SGR with the MED representative selection rule includes eleven predictors in Table 15. Profiles of the coefficients of the PR-SGR with the MED and MED representative selection rules are drawn in Figures 12 and 13, respectively, as the tuning parameter t varies. The final models are determined with the t value chosen by 5−CV, and the vertical red lines in the figures are drawn at the selected s to indicate the final model and the estimated coefficients. Both methods build the same clusters up to Step 9, but the PR-SGR with the MED representative selection rule finds two more clusters in Steps 10 and 11. The two clusters built with the PR-SGR with the MED representative selection rule are generated due to the representative at Step 9, resulting in the change of the coefficients.

Table 14: The clusters built by the PR-SGR with the MAX representative selection rule, their representatives and their estimated coefficients.

Step	Cluster	Cluster Size	Representative	Estimated Coefficient
1	C3CZ/C4CZOVRL2	1	C3CZ/C4CZOVRL2	6.9760
2	C3CZ/C4CZCCOH2	1	C3CZ/C4CZCCOH2	2.0678
3	ALLLRP3, RHEMLRP3	2	RHEMLRP3	2.0327
4	F4AMF2, C3AMF2 C4AMF2, T4AMF2 FZAMF2, CZAMF2 PZAMF2	7	CZAMF2	1.5279
5	T3/T4TASY2	1	T3/T4TASY2	1.5613
6	BACKDCOH3	1	BACKDCOH3	2.0747
7	T3TRP1	1	T3TRP1	1.1024
8	F3OMF2, F7OMF2 T3OMF2, FPZOMF2	4	T3OMF2	1.3803
9	BACKCCOH2	1	BACKCCOH2	0.6873
10	F3OMF1, C3OMF1 FZOMF1, CZOMF1	4	FZOMF1	0
11	BACKTRP2	1	BACKTRP2	0
12	FP1/FP2DCOH2	1	FP1/FP2DCOH2	0
13	F3/F4TCOH2	1	F3/F4TCOH2	0
14	P3OMF1, P4OMF1 PZOMF1	3	PZOMF1	0

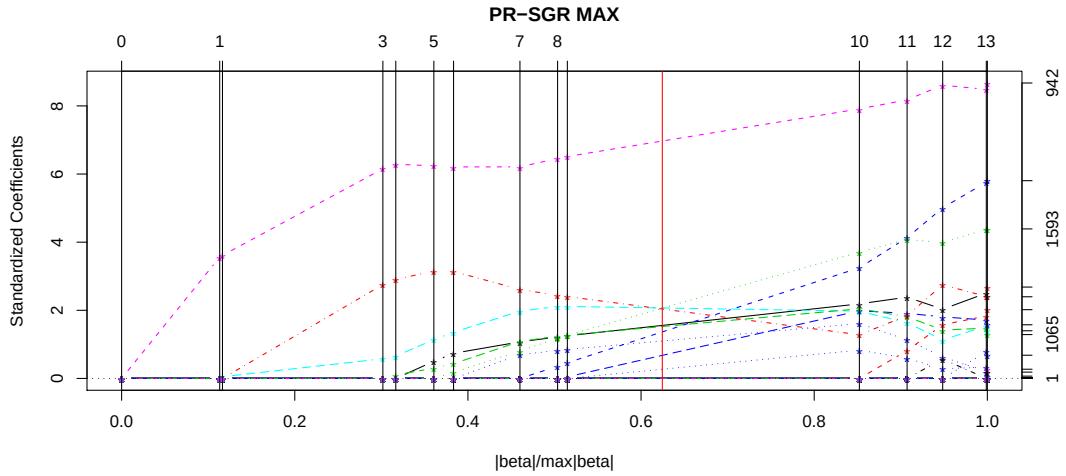


Figure 12: Profiles of the coefficients of the PR-SGR with the MAX representative selection rule, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\|\hat{\beta}^{OLS}\|_1$. The vertical red line is drawn and its value is chosen by 10−fold Cross-Validation.

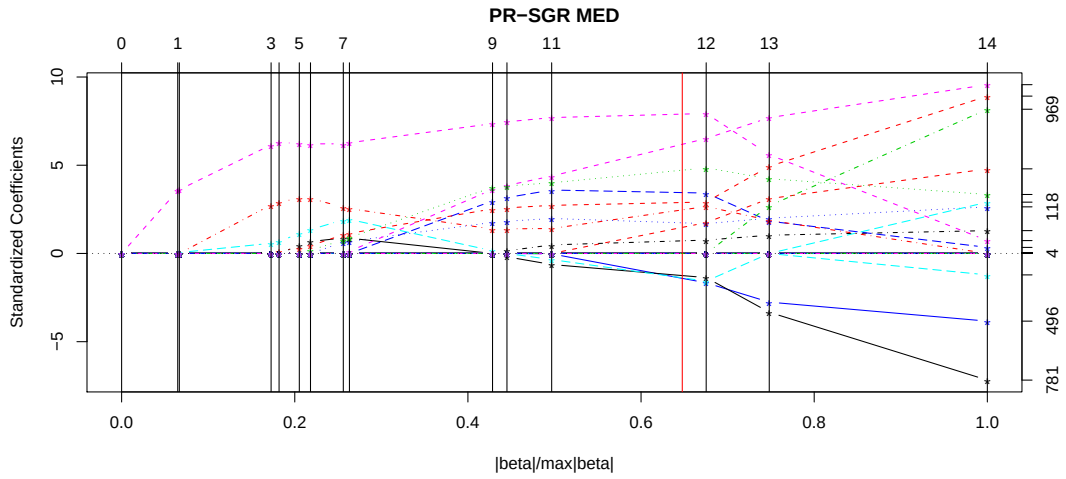


Figure 13: Profiles of the coefficients of the PR-SGR with the MED representative selection rule, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\|\hat{\beta}^{OLS}\|_1$. The vertical red line is drawn and its value is chosen by 10−fold Cross-Validation.

Table 15: The clusters built by the PR-SGR with the MED representative selection rule, their representatives and their estimated coefficients.

Step	Cluster	Cluster Size	Representative	Estimated Coefficient
1	C3CZ/C4CZOVRL2	1	C3CZ/C4CZOVRL2	7.8952
2	C3CZ/C4CZCCOH2	1	C3CZ/C4CZCCOH2	-1.3717
3	ALLLRP3, RHEMLRP3	2	RHEMLRP3	2.4608
4	F4AMF2, C3AMF2 C4AMF2, T4AMF2 FZAMF2, CZAMF2 PZAMF2	7	C3AMF2	2.8855
5	T3/T4TASY2	1	T3/T4TASY2	-1.2485
6	BACKDCOH3	1	BACKDCOH3	4.7140
7	T3TRP1	1	T3TRP1	1.7374
8	F3OMF2, F7OMF2 T3OMF2, FPZOMF2	4	F3OMF2	6.1716
9	BACKCCOH2	1	BACKCCOH2	3.4511
10	FP1/FP2DCOH2	1	FP1/FP2DCOH2	0.7131
11	F3OMF1, C3OMF1 FZOMF1, CZOMF1	4	C3OMF1	-1.4002
12	BACKTRP2	1	BACKTRP2	0
13	ALLBRP2	1	ALLBRP2	0
14	P3O1/P4O2CCOH3 BACKCCOH3	2	P3O1/P4O2CCOH3	0

1.5 Discussion and Summary

In this chapter, we have studied the problem of variable selection in the high dimensional data for the linear regression model. We have proposed a penalized method based on representatives (the PR) as extensions of the Lasso. Clusters of highly pairwise correlated predictors are formulated and representatives of predictors in clusters are selected. The selected representatives are candidates for Lasso variable selection and the model fitting. We have developed an efficient algorithm (the PR-SGR) to solve the PR estimates as a generalization of the Lars. The PR-SGR produces a sparse model having less than n predictors with good prediction accuracy, while grouping highly correlated predictors and selecting representatives of the groups. The simulations and empirical results demonstrate its good performance and its superiority over the Lasso/Lars, Elastic-Net, grouped-Lasso and grouped-Lars.

This study promotes the following open topics. We consider them as potential future works that deserve further attention and study. First, the PR solution can have an interpretation as the posterior mode for a particular prior distribution like other penalized regression techniques have, although it is expected to be more complicated due to the representative selection concept. Second, there are various possibilities to extend the PR by combining the idea of representative selection with other sparse penalties such as SCAD ([10]), adaptive Lasso ([34]) or fused Lasso ([35]). Additionally, we can extend the PR for a linear model to the family of generalized linear models (GLM). Lastly, we may consider selecting two or more variables from an important group (when the data size is large enough). If one variable is not enough to represent its group, that is, two of predictors in the group have similar values of c_j's and if the group improves the regression fitting error, including more than one variable from the group may improve the overall performance.

1.6 Summary of Contributions

As discussed earlier, handling of high dimensional data has been one of the most challenging tasks over the recent years. Due to high dimensionality, overfitting might occur resulting in questionable prediction quality. To improve the prediction accuracy, penalized methods with L_1-penalty have been popular by effectively identifying a subset of important predictors. Here we focused on the case that the number of predictors p is much larger than the sample size n in the high dimensional data.

A challenge with high dimensional data is how to select only important predictors which can be spuriously highly correlated with some or many unimportant ones, as we discussed in Section 1.1. The existing penalized methods such as the Elastic-net select all correlated predictors simultaneously (when one of them would give nonzero coefficient). Some methods such as the grouped-Lasso also select the whole set of predictors belonging to a pre-defined group/cluster. We discussed why including all spuriously correlated predictors may not be meaningful. In this chapter, we suggested a way to take advantage of clustering scheme to reduce the high dimensionality and take advantage of the L_1 penalized regression method to fit a sparse model.

We have developed a new method using L_1-penalty and the representative concept as a way to handle a high dimensional problem with strong ungenuine correlations. As we explained the details in Section 1.2, the modeling based on the representatives of clusters handled the spurious correlation problem successfully. The algorithm PR-SGR introduced in Section 1.2.4 helps build a PR regression model efficiently.

CHAPTER II

NONPARAMETRIC ANALYSIS OF GAP DATA

2.1 Introduction and Motivation

This chapter considers a problem of analyzing time-to-event (survival) data, when missing time intervals referred as gaps ([14, 15, 38]) are possibly detected prior to the first observed event. The gap data raises a concern in statistical analysis, because no data is recorded in those intervals and it is unknown whether an interesting event happens in the gaps or not. In turn, the gap data causes complications in estimating the survival function of the occurrence time of an interesting event such as the first event time.

Green *et al.* ([14, 15]) motivated and studied the gap data analysis with the data from a heart disease study carried out at Duke University Medical Center. The medical data was obtained when using a ST-segment heart monitoring method to detect reperfusion (reopening) of the infarct artery (partially closed or closed artery) in patients. The monitoring devices recording the ST-segment levels may experience periodic failures to store amplitude and result in time intervals of missing data (multiple gaps) ([15]). Also, a patient can experience more than one recovery (multiple events), where the recovery indicates the time of the first reperfusion when the ST-segment amplitude falls to 50 % of the peak deviation. Such data are abundant in many studies from medical, industrial, and economic fields. For example, gap data can be obtained when monitoring a patient's heart activities in biomedical studies or when evaluating a manufacturer's product performance in an industrial quality improvement research.

Left-, right-, or interval-censoring happens when the value of an observation is not

completely known but partially known. Left-censoring occurs when a data point is below a certain value but the exact time is unknown. Similarly, right-censoring occurs when a data point is above a certain value but the exact time is unknown. Interval-censoring occurs when a data point is somewhere of a time interval between two values. Peto ([31]) suggested a censoring scheme using a constrained Newton-Raphson search for locating the maximum of the log likelihood. However,its optimization might not be feasible with a large number of pseudoparameters and the Newton-Raphson method does not guarantee to find a global maximum. Turnbull ([36]) introduced a general scheme and showed that the maximization of the likelihood function is equivalent to the solution of a self-consistent equation, solved using an Expectation-Maximization (EM) algorithm for computing the Non-Parametric MLE (NPMLE). To compute the NPMLE, Groeneboom ([16, 17]) introduced the Iterative Convex Minorant (ICM) algorithm. Since the EM algorithm might be very slow even with a moderate number of parameters, Groeneboom and Wellner ([18]) suggested the ICM algorithm which is much more efficient to compute the NPMLE than the EM algorithm, especially with a large sample size. Gentleman and Geyer ([13]) proposed a method to ensure that the solution is a global maximum.

In censored data analysis, the independence between censoring and event is a crucial assumption. Several statisticians such as Cox ([7]), Efron ([8]), Kalbfleisch and Mackay ([24]), and Kalbfleisch and Prentice ([25]) have studied more general censoring processes with relaxation on the independence between censoring and event in one-side (left- or right-) censored data. Their common idea is to model the censoring-event process as the process unfolds in time. However, the likelihood function derived from the independent assumption is valid only if quasi-independence between censoring and event exists. Therefore, the extension of studies to the dependent censored data becomes limited.

At first glance, the gap data seems to be in the broad category of the interval-censored data, because the missing gaps are half-open intervals and one single time point can be considered as a closed interval (for example, $t = [t, t],\ t > 0$). However, the assumption of the independence between censoring intervals and events in the censored data is not applicable to the gap data. In the gap data, the observed first event time can be viewed as a censoring time, but the observed first event and the first true event are highly dependent. Therefore, the gap data becomes a new type of interval censored data with a mixture of independent and dependent censoring intervals. As Yang ([38]) pointed out, there is no quasi-independent relationship in the gap data between the first true events and the censoring intervals happening to include both gaps and the observed first event times, because the observed first events have positive probability of being the first true events. For example, if a subject record contains only its observed first event time and a missing gap, all information we can extract is that the first true event time may be in the gap or exactly same as the observed first event time. So this is not a censored data but a gap data.

Green *et al.* ([14, 15]) proposed a parametric likelihood function, the Gap Likelihood Function (GLF), to estimate the distribution of the true event time W_1 with a gap data where the first observed event times are given with several missing data intervals. Section 2.2 describes its detailed methodology. The GLF utilizes the data by comparing standard failure time methods for right-censoring data such as Kaplan-Meier analysis method ([26]) and analyzing a real data from the Duke Medical School and several simulation studies. It is shown that the GLF is more efficient and less biased than right censoring methods. Some prefer parametric methods because they provide unique information based on their distribution assumptions, but others look for nonparametric ways because nonparametric methods require less restrictive distribution assumptions and work well with relatively smaller data.

Yang ([38]) proposed a nonparametric method, the Imputed Empirical Estimating

method (IEE), which uses the imputation idea when calculating the probability of the first true event being in the gap, to construct an empirical estimate of the survival function of the first true event time. It is shown that the IEE method outperforms other traditional nonparametric methods such as the classical estimating approach ([32]), applied to the observed first event time while ignoring gaps in simulations studies. However, it is difficult to derive its probabilistic and asymptotic properties, because the IEE formula uses the ordered subject records sorted by their observed first event times. Moreover, Yang claimed that the IEE estimate is unbiased and robust via simulation studies, but it is shown as a biased estimate in Appendix D and it might not be robust under some conditions provided in Section 2.3.2.2. This chapter proposes a new nonparametric method, a Non-Parametric Estimate for the Gap data (NPEG), with clear statistical properties and definition. Based on the analytical work on the bias of the NPEG and simulation studies, we discuss situations in which the bias can be reduced.

Section 2.2 defines notations, assumptions and definitions used in deriving the proposed method. With these notations and assumptions, the detailed methodologies of the GLF ([14, 15]) and the IEE ([38]) are reviewed. Section 2.3 proposes a new estimation method and will provide its basic statistical properties such as its mean and variance. The subsequential section compares the proposed method with the IEE using a numerical example. The proposed method will be illustrated with simulation examples in Section 2.4 and a real dataset in Section 2.5. Section 2.6 concludes this chapter with discussions and summaries.

2.2 Notation, Assumption and Relevant Literature

This section introduces the notations and provides assumptions, used to describe the gap data. Let T_1 denote a random variable of the observed first event time, W_k be a random variable of the k^{th} true event time, and $G = (B, E]$ be a gap where B and

E are random variables denoting the beginning and ending of the gap, respectively. The random variables W_k, B, and E are assumed to be independent.

Given n subjects, $T_{1,i}$ and $G_i = (B_i, E_i]$, $i = 1, \cdots, n$ are assumed independently and identically distributed (*i.i.d.*). If no gap has occurred before the observed first event time $T_{1,i}$ of the i^{th} observation (*no gap* case), $T_{1,i} = W_1$. In the no gap case,B_i and E_i are defined as infinity and G_i as an empty set for simplicity. If the i^{th} observation has a gap, data collection is continued until the observed first event time $T_{1,i}$ is detected. If there are any cases containing more than one gap prior to the observed first event time, such a case is discarded in this analysis so that cases with either no gap or only one gap are considered. To simplify the study, it is assumed that the chance to have more than one event in a gap is negligible.

As a result of the notations and assumptions described above, the i^{th} gap G_i becomes

$$G_i = \begin{cases} (B_i, E_i] & , \text{ with a gap} \\ \emptyset & , \text{ with no gap, i.e., } (B_i, E_i] = (\infty, \infty] \end{cases}$$

And the observed first event time $T_{1,i}$, $i = 1, \cdots, n$ becomes

$$T_{1,i} = \begin{cases} W_{1,i} & , \text{ with no gap} \\ W_{1,i} & , \text{ with a gap and } W_{1,i} \notin G_i \\ W_{2,i} & , \text{ with a gap and } W_{1,i} \in G_i. \end{cases} \quad (5)$$

Note that our goal here is estimating the survival function $S_{W_1}(w) = P(W_1 > w)$ of the first true event time W_1 for $w \in (0, \infty)$.

Here, two existing methods developed to handle the gap data are reviewed in detail. First, the parametric modeling method GLF proposed by Green *et al.* ([14, 15]) is reviewed, only when the models are relevant to the study setting of this chapter, although more complicated cases are considered in the GLF. The joint probability density function (*pdf*) of B and E are defined as $f_{(B,E)}(b, e)$, and *pdf*s of W_1 and W_2 as $f_{W_1}(t)$ and $f_{W_2}(t)$, respectively. When no gap has occurred before the observed

first event time $T_{1,i} = t$, the observed first event time $T_{1,i}$ becomes the first true event time $W_{1,i}$ and its likelihood becomes $f_{W_1}(t),\ 0 < t < \infty$. In this case, it is unnecessary to include the information of the second true event time $W_{2,i}$. With one gap before T_1, the conditional *pdf* of the observed first event time T_1 given the gap $G = (B, E] = (b, e]$ is derived as follows:

$$f_{T_1|(b,e)}(t) = f_{W_1}(t) + \int_b^e f_{W_2}(t-x) f_{W_1}(x) dx, \quad e < t < \infty \tag{6}$$

where W_2 is the second true event time. Without the assumption that the probability of two events occurring in a gap is zero, the conditional *pdf* becomes

$$\begin{aligned} f_{T_1|(b,e)}(t) &= f_{W_1}(t) + \int_b^e f_{W_2}(t-x) f_{W_1}(x) dx \\ &\quad + \int_b^e \int_x^e f_{W_3}(t-y) f_{W_2}(y-x) f_{W_1}(x) dy dx, \quad e < t < \infty, \end{aligned}$$

where f_{W_3} is the *pdf* of the third true event time W_3. Therefore, the conditional likelihood for the random samples $T_{1,i} = t_{1,i},\ G_i = (B_i, E_i] = (b_i, e_i],\ i = 1, \cdots, n$ is defined as

$$L = \prod_{i=1}^{n} \left[f_{W_1}(t_{1,i}) \right]^{\delta_i} \left[f_{T_1|(b_i,e_i)}(t_{1,i}) f_{(B,E)}(b_i, e_i) \right]^{1-\delta_i},$$

where δ_i is an indicator which equals to 1 with no gap or equals to 0 with a gap.

A nonparametric estimate of gap data, IEE ([38]) starts with estimated probability that W_1 is in the gap. The method simply considers $n-1$ subjects as the sample data for the i^{th} subject and estimates the probability p_{wig} of $T_{1,i}$ being located in a known gap $G_i = (B_i, E_i]$. The following is the algorithm for the IEE estimate: Given n subjects with $(T_{1,i}, B_i, E_i) = (t_{1,i}, b_i, e_i),\ i = 1, \cdots, n$,

1. Sort the data according to their observed first event time $t_{1,i}$ to get the ordered observed first event time $t_{1,(i)},\ i = 1, \cdots, n$.

2. For each ordered time point $t_{1,(i)}$, two probabilities $p_{wig}(i, j)$ and $p_{wap}(i)$ are estimated. $p_{wig}(i, j),\ j \neq i,\ j = 1, \cdots, n$ for the i^{th} ordered subject, is the

imputed probability that the first true event time W_1 is located at $t_{1,(j)}$ falling in the ordered gap $G_{(i)} = (B_{(i)}, E_{(i)}] = [b_{(i)}, e_{(i)})$. $p_{wap}(i)$ is the probability that W_1 is at the observed first event time $T_{1,(i)} = t_{1,(i)}$. The IEE calculates $\hat{p}_{wig}(i,j)$ and $\hat{p}_{wap}(i)$ as

$$\begin{aligned}
\hat{p}_{wap}(1) &= 1, \\
\hat{p}_{wig}(1,1) &= \frac{\hat{p}_{wap}(1)}{n-n_1} = \frac{1}{n-n_1}, \\
\hat{p}_{wap}(2) &= \begin{cases} 1-\hat{p}_{wig}(1,1) & , \text{ if } t_{1,(1)} \in (b_{(2)}, e_{(2)}] \\ 1 & , \text{ otherwise} \end{cases}, \\
&\vdots
\end{aligned}$$

where n_i is the number of observed gap including $t_{1,(i)}$, that is, $n_i = \sum_{j=i+1}^{n} I(t_{1,(i)} \in (b_{(j)}, e_{(j)}])$. Therefore, the above equations are written as

$$\begin{aligned}
\hat{p}_{wap}(i) &= 1 - \sum_{j=1}^{i-1} \hat{p}_{wig}(i,j) I(T_{1,(j)} \in (b_{(i)}, e_{(i)}]),\ i = 2, \cdots, n, \text{ and} \\
\hat{p}_{wig}(i,j) &= \frac{\hat{p}_{wap}(i)}{n-n_i},\ i = 1, \cdots, n.
\end{aligned}$$

3. The IEE of $S_{W_1}(t)$ is defined as

$$\hat{S}_{IEE}(t) = 1 - \sum_{i:t_{1,(i)}<t} \frac{\hat{p}_{wap}(i)}{n-n_i}.$$

By letting $h_{ij} = I(t_{1,(i)} \in G_{(j)})$, the calculations of the IEE are simplified as follows:

$$\begin{aligned}
\hat{S}_{IEE}(t) &= 1 - \sum_{i:t_{1,(i)}<t} \frac{\hat{p}_{wap}(i)}{n-n_i} \\
&= 1 - \sum_{i:t_{1,(i)}<t} \frac{\hat{p}(i)}{m_i} \\
&= 1 - \sum_{i=1}^{n} \frac{\hat{p}(i)}{m_i} I(t_{1,(i)} \le t),
\end{aligned}$$

where $\hat{p}(i) = 1 - \sum_{j=1}^{i-1} \frac{h_{ji}}{m_j} \hat{p}(j)$ and $m_i = n - n_i = n - \sum_{j=i+1}^{n} h_{ij}$. The IEE is extended with multiple gaps with this simplified form, in Appendix C.

Nonparametric methods are sometimes preferred because they require less restrictive distribution assumptions than parametric methods and work well with relatively smaller data. On the other hand, the nonparametric method IEE is built based on the ordered subject records and it has a limitation in deriving some probabilistic properties. The subsequent section proposes a new nonparametric method (the NPEG) with better-defined estimate and clear statistical properties.

2.3 Proposed Estimate

Two main problems are considered in the gap data. First, since the first true event time W_1 is not observable, it needs to be estimated from the observable variables, the observed first event T_1 and observable variable gap $G = (B, E]$. Another problem is that it is unknown if W_1 is in the gap G and/or equals to T_1. Therefore, we need to study the following to construct a nonparametric estimated survival function for the first true event time W_1: first discuss (a) the relationship between $I(W_1 \leq t)$and $I(T_1 \leq t)$, show (b) the relationship between S_{W_1} and S_{T_1} based on (a), and then provide (c) a new estimate $\hat{S}_{NPEG}$ of S_{W_1} based on (b) in terms of only the observable variables T_1 and $G = (B, E]$.

When estimating the survival function of the first true event, $S_{W_1}(t)$ should be expressed in terms of T_1 and G, because only T_1's and G's are observed with the incomplete information of whether or not $T_1 = W_1$. Based on the relationship between $W_1,\ W_2,\ T_1$ and G, the relationship between $I(W_1 \leq t)$ and $I(T_1 \leq t)$ is written as $I(W_1 \leq t) = I(T_1 \leq t) + I(B < W_1 \leq \min(E, t))I(T_1 > t)$ for a given censoring time t. Then, the relationship between $P(W_1 \leq t)$ and $P(T_1 \leq t)$ is shown, by simply taking expectations on both sides of the equation, as $P(W_1 \leq t) = P(T_1 \leq t) + P(T_1 > t,\ B < W_1 \leq \min(E, t))$. The definition of the survival function $S_T(t) = P(T > t) = 1 - P(T \leq t)$ gives $S_{W_1}(t) = S_{T_1}(t) - P(T_1 > t,\ B < W_1 \leq \min(E, t))$. $S_{W_1}(t)$ can be written as $S_{W_1}(t) = S_{T_1}(t)\{1 - P(B < W_1 \leq \min(E, t)) \mid T_1 > t\})$ by using the

definition of a conditional expectation. Due to the lack of knowledge in the gap data, the W_1 cannot be removed from the equation of $S_{W_1}(t)$. Thus, $n-1$ subjects are considered as sample data to estimate the first true event time for the i^{th} subjects, that is, the unobserved value $W_{1,i}$ is estimated based on $T_{1,j},\ j \neq i,\ j = 1, \cdots, n$. Therefore, the new estimate $\hat{S}_{NPEG}(t)$ is defined as

$$\hat{S}_{NPEG}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_{1,i} > t) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i, t)]).$$

The details of each relationship and of the new estimate are as follows.

First, $I(W_1 \leq t)$ can be written in terms of other variables as

$$I(W_1 \leq t) = I(T_1 \leq t) + I(B < W_1 \leq \min(E, t))I(T_1 > t). \tag{7}$$

Proof:

$$\begin{aligned}
I(T_1 \leq t) &= I(T_1 \leq t)I(W_1 \in G) + I(T_1 \leq t)I(W_1 \notin G) \\
&= I(W_2 \leq t)I(W_1 \in G) + I(W_1 \leq t)I(W_1 \notin G) \\
&\qquad \because\ T_1 = \begin{cases} W_2 & \text{, if } W_1 \in G \\ W_1 & \text{, if } W_1 \notin G \end{cases} \\
&= I(W_2 \leq t)I(W_1 \in G) + I(W_1 \leq t)\{1 - I(W_1 \in G)\} \\
&= I(W_1 \leq t) - I(W_1 \in G)\{I(W_1 \leq t) - I(W_2 \leq t)\} \\
&= I(W_1 \leq t) - I(W_1 \in G)I(W_1 \leq t < W_2) \\
&\qquad \because\ I(W_1 \leq t) - I(W_2 \leq t) = \begin{cases} 1 & \text{, if } W_1 \leq t \ \&\ W_2 > t \\ 0 & \text{, otherwise} \end{cases} \\
&= I(W_1 \leq t) - I(W_1 \in G)I(W_1 \leq t)I(W_2 > t) \\
&= I(W_1 \leq t) - I(B < W_1 \leq \min(E, t))I(W_2 > t) \\
&\qquad \because\ I(W_1 \in G)I(W_1 \leq t) = I(B < W_1 \leq E)I(W_1 \leq t) \\
&\qquad\qquad = I(B < W_1 \leq \min(E, t)) \\
&= I(W_1 \leq t) - I(B < W_1 \leq \min(E, t))I(T_1 > t)
\end{aligned}$$

$\because\ T_1 = W_2$ if $W_1 \in G$. $\square$

Now,

$$S_{W_1}(t) = S_{T_1}(t) - P(T_1 > t,\ B < W_1 \leq \min(E,t)) \tag{8}$$

or

$$= S_{T_1}(t)\{1 - P(B < W_1 \leq \min(E,t) \mid T_1 > t)\} \tag{9}$$

is shown based on Equation (7).

Proof: Since $S_{W_1}(t) = 1 - P(W_1 \leq t)$ and $S_{T_1}(t) = 1 - P(T_1 \leq t)$, the following relationship between $P(W_1 \leq t)$ and $P(T_1 \leq t)$ is driven by taking the expectation of Equation (7).

$$\begin{aligned}
P(T_1 \leq t) &= \mathbb{E}\{I(T_1 \leq t)\} \\
&= \mathbb{E}\{I(W_1 \leq t) - I(B < W_1 \leq \min(E,t))I(T_1 > t)\} \\
&= P(W_1 \leq t) - \mathbb{E}\{I(B < W_1 \leq \min(E,t),\ T_1 > t)\} \\
&= P(W_1 \leq t) - P(B < W_1 \leq \min(E,t),\ T_1 > t) \qquad (10) \\
&= P(W_1 \leq t) - P(T_1 > t) \cdot P(B < W_1 \leq \min(E,t) \mid T_1 > t). \qquad (11)
\end{aligned}$$

Therefore,

$$\begin{aligned}
S_{W_1}(t) &= 1 - P(W_1 \leq t) \\
&= 1 - P(T_1 \leq t) - P(B < W_1 \leq \min(E,t),\ T_1 > t) \qquad \text{from (10)} \\
&= S_{T_1}(t) - P(T_1 > t,\ B < W_1 \leq \min(E,t))
\end{aligned}$$

or

$$\begin{aligned}
&= 1 - P(T_1 \leq t) - P(T_1 > t) \cdot P(B < W_1 \leq \min(E,t) \mid T_1 > t) \qquad \text{from (11)} \\
&= S_{T_1}(t) - S_{T_1}(t) \cdot P(B < W_1 \leq \min(E,t) \mid T_1 > t) \\
&= S_{T_1}(t)\{1 - P(B < W_1 \leq \min(E,t) \mid T_1 > t)\}. \qquad \square
\end{aligned}$$

Based on the above relationship and Equation (8), the new estimate is defined as

$$\hat{S}_{NPEG}(t) = \hat{S}_{T_1}(t) - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i, t)]), \qquad (12)$$

where

$$\begin{aligned} \hat{S}_{T_1}(t) &= \frac{1}{n} \sum_{i=1}^{n} I(T_{1,i} > t), \\ \hat{P}(T_1 \geq t,\ B < W_1 \leq \min(E, t)) &= \frac{1}{n} \sum_{i=1}^{n} \hat{P}(T_{1,i} > t,\ B_i < W_{1,i} \leq \min(E_i, t)), \end{aligned}$$

and

$$\begin{aligned} & \hat{P}(T_{1,i} > t,\ B_i < W_{1,i} \leq \min(E_i, t)) \\ = \ & \hat{P}(T_{1,i} > t,\ T_{1,j},\ \text{for any } j \text{ is located in } (B_i, \min(E_i, t)]) \\ & \because\ W_1 \text{ is not observable, so use } T_{1,i}\text{'s from } n-1 \text{ subjects} \\ = \ & \sum_{\substack{j=1 \\ j \neq i}}^{n} \hat{P}(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i, t)]) \\ & \because\ T_{1,j}\text{'s are independent } \ \& \ \ T_{1,j} \notin (B_i, \min(E_i, t)] \subseteq G_i \text{ for } j = i \\ = \ & \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i, t)]). \end{aligned}$$

Remarks: The estimate $\hat{S}_{NPEG}(t)$ is a nonparametric estimate of the survival function of the first true event time W_1. $\hat{S}_{NPEG}(0) = 0$ and $\hat{S}_{NPEG}(t) \longrightarrow 0$ as $t \longrightarrow \infty$.

2.3.1 Example

This section uses a numerical example introduced in ([38]) to illustrate the NPEG and IEE estimates. The dataset contains eight subjects, where each observation consists of gap information and the observed first event time. The observed data and their ordered data are given in Table 16.

The new estimate $\hat{S}_{NPEG}(t)$ is calculated first. Fix the value of censoring time t before calculating $\hat{S}_{NPEG}(t)$. For $0 < t \leq 10$, $t_{1,i} > t$ for any $i = 1, \cdots, 8$ and

Table 16: Example introduced in Yang's thesis up to one gap.

Original data			Ordered data		
i	$g_i = (b_i, e_i]$	$t_{1,i}$	(i)	$g_{(i)} = (b_{(i)}, e_{(i)}]$	$t_{1,(i)}$
1	(21,52]	57	1	(3,5]	10
2	(12,26]	31	2	(7,16]	20
3	(17,25]	63	3	(12,26]	31
4	(3,5]	10	4	$(\infty, \infty]$	35
5	(29,37]	47	5	(13,24]	45
6	$(\infty, \infty]$	35	6	(29,37]	47
7	(7,16]	20	7	(21,52]	57
8	(13,24]	45	8	(17,25]	63

$t_{1,j} \notin (b_i, \min(e_i, t)]$ for any $i, j = 1, \cdots, 8,\ i \neq j$. Therefore, $\hat{S}_{NPEG}(t) = 1$. For $10 < t \leq 20$, the subject 4 should be ignored because $t_{1,4} < t$. For the adjusted gaps $(b_i, \min(e_i, t)]$, only the adjusted gap of the subject 7 contains the observed first event time of the subject 4. Therefore, $\hat{S}_{NPEG}(t) = \frac{7}{8} - \frac{1}{7 \cdot 8} = 0.8571$.For $20 < t \leq 31$, the subjects 4 and 7 are ignored because $t_{1,i} < t$ for $i = 4$ and 7. For a fixed $i \neq 4$ or 7, only three adjusted gaps for the subjects 2, 3, and 8 contain other observed first event times. That is, for $i = 2,\ 3$, or 8, $t_{1,7} = 20 \in (b_i, \min(e_i, t)]$. Consequently, $\hat{S}_{NPEG}(t) = \frac{6}{8} - \frac{3}{7 \cdot 8} = 0.6964$. For $31 < t \leq 35$, $i = 2,\ 4$ and 7 are ignored. For $i = 1$, the adjusted gap $(21, t]$ contains $t_{1,2} = 31$. When $i = 3$, the adjusted gap $(17, 25]$ contains $t_{1,7} = 20$. When $i = 5$, the adjusted gap $(29, t]$ contains $t_{1,2} = 31$. For $i = 8$, the adjusted gap $(13, 24]$ contains $t_{1,7} = 20$. Therefore, $\hat{S}_{NPEG}(t) = \frac{5}{8} - \frac{4}{7 \cdot 8} = 0.5536$. Similarly, the rest $\hat{S}_{NPEG}(t)$ values can be calculated. For $35 < t \leq 45$, only half of the subjects are considered with $i = 1,\ 3,\ 5$ and 8. The adjusted gap $(b_1, \min(e_1, t)] = (21, t]$ contains $t_{1,2} = 31$ and $t_{1,6} = 35$, the adjusted gap $(b_3, \min(e_3, t)] = (17, 25]$ contains $t_{1,7} = 20$, the adjusted gap $(b_5, \min(e_5, t)] = (29, \min(37, t)]$ contains $t_{1,2} = 31$ and $t_{1,6} = 35$, and the adjusted gap $(b_8, \min(e_8, t)] = (13, 24]$ contains $t_{1,7} = 20$. Thus, $\hat{S}_{NPEG}(t) = \frac{4}{8} - \frac{6}{7 \cdot 8} = 0.3929$. For $45 < t \leq 47$, only three subjects with $i = 1,\ 3$, or 5 are considered. $t_{1,2},\ t_{1,6}$ and $t_{1,8}$ fall in the adjust

gap $(b_1, \min(e_1, t)] = (21, t]$. $t_{1,7}$ is in $(b_3, \min(e_3, t)] = (17, 25]$. $t_{1,2}$ and $t_{1,6}$ are falling into the adjust gap $(b_5, \min(e_5, t)] = (29, 37]$. Thus, $\hat{S}_{NPEG}(t) = \frac{3}{8} - \frac{6}{7 \cdot 8} = 0.2679$. For $47 < t \leq 57$, only two subjects with $i = 1$ and 3 are considered. The adjusted gap $(b_1, \min(e_1, t)] = (21, \min(52, t)]$ contains $t_{1,j}$ for $j = 2,\ 5,\ 6$ or 8 and the adjusted gap $(b_3, \min(e_3, t)] = (17, 25]$ contains $t_{1,7} = 20$. Thus, $\hat{S}_{NPEG}(t) = \frac{2}{8} - \frac{5}{7 \cdot 8} = 0.1607$. For $57 < t \leq 63$, only one subject with $i = 3$ is considered. Its adjusted gap becomes $(b_3, \min(e_3, t)] = (17, 25]$ where $t_{1,7}$ falls into. The estimated $\hat{S}_{W_1}(t)$ is calculated as $\hat{S}_{NPEG}(t) = \frac{1}{8} - \frac{1}{7 \cdot 8} = 0.1071$. Finally, for $t > 63$, there is no subject to be considered. Therefore, $\hat{S}_{NPEG}(t) = \frac{0}{8} = 0$. Table 18 provides the estimated survival function based on the NPEG, $\hat{S}_{NPEG}(t)$ for a given censoring time t.

Since Yang ([38]) provided a numerical calculation with its original algorithm, the calculation with the simplified form is provided below. First, calculate $\hat{p}(i)$ for each ordered subject i. With the first ordered subject with $t_{1,(1)} = 10$, the estimate for $p(1)$ is one as defined. Then, for the next ordered subject with $t_{1,(2)} = 20$, $h_{12} = 1$ and $n_1 = 1$, because $t_{1,(1)} \in g_{(2)}$ and only the second ordered gap covers $t_{1,(1)}$. Therefore, $\hat{p}(2) = 1 - \frac{h_{12}}{m_1} = 1 - \frac{1}{8-1} = 0.8571$. For the third ordered subjects with $t_{1,(3)} = 31$, the gap $g_{(3)} = (12, 26]$ of the third ordered subject does not cover $t_{1,(1)} = 10$ but cover $t_{1,(2)} = 20$. Therefore, $h_{13} = 0$ and $h_{23} = 1$. Since the third, fifth and eighth ordered gaps cover $t_{1,(2)}$, $n_2 = 3$, $\hat{p}(3) = 1 - \frac{h_{13}}{m_1} - \frac{h_{23}}{m_2} = 0.8286$. The fourth ordered subject with $t_{1,(4)}$ has no gap, so $h_{i4} = 0,\ i = 1, 2,$ or 3 and $\hat{p}(4) = 0$. For the fifth ordered subject with $t_{1,(5)}$, $\hat{p}(5) = 0.8286$ because $n_3 = 2,\ n_4 = 2, h_{15} = h_{35} = h_{45} = 0$ and $h_{25} = 1$. The gap of the next ordered subject contains $t_{1,(3)}$ and $t_{1,(4)}$ and its observed first event $t_{1,(6)}$ is falling in the gap $g_{(7)}$. Therefore, the estimate of $p(6)$ is $\hat{p}(6) = 1 - \frac{h_{16}}{m_1} - \frac{h_{26}}{m_2} - \frac{h_{36}}{m_3} - \frac{h_{46}}{m_4} - \frac{h_{56}}{m_5} = 0.6952$. For the seventh ordered subject, $h_{17} = h_{27} = 0,\ h_{37} = h_{47} = h_{57} = h_{67} = 1$ and $n_6 = 1$. So $\hat{p}(7) = 0.4776$. Finally, since the eighth ordered subject' gap contains only $t_{1,(2)}$, h_{28} becomes one and other h_{ij} values are zero and $\hat{p}(8) = 0.8286$. These calculation procedures are summarized

Table 17: Procedure to calculate the IEE with the simplified formulation.

(i)	$g_{(i)} = [b_{(i)}, e_{(i)})$	$t_{1,(i)}$	h_{1i}	h_{2i}	h_{3i}	h_{4i}	h_{5i}	h_{6i}	h_{7i}	n_i	$\hat{p}(i)$
1	(3,5]	10								1	1
2	(7,16]	20	1							3	0.8571
3	(12,26]	31	0	1						2	0.8286
4	$(\infty, \infty]$	35	0	0	0					2	1
5	(13,24]	45	0	1	0	0				1	0.8286
6	(29,37]	47	0	0	1	1	0			1	0.6952
7	(21,52]	57	0	0	1	1	1	1		0	0.4776
8	(17,25]	63	0	1	0	0	0	0	0	0	0.8286

Table 18: The IEE and NPEG estimates on the example dataset in Table 16.

t	$\hat{S}_{IEE}(t)$	$\hat{S}_{NPEG}(t)$
$0 < t \leq 10$	1	1
$10 < t \leq 20$	0.8571	0.8571
$20 < t \leq 31$	0.6857	0.6964
$31 < t \leq 35$	0.5476	0.5536
$35 < t \leq 45$	0.3810	0.3929
$45 < t \leq 47$	0.2626	0.2679
$47 < t \leq 57$	0.1633	0.1607
$57 < t \leq 63$	0.1036	0.1071
$63 < t$	0	0

in Table 17. Table 18 also provides the estimated survival function based on the IEE method, $\hat{S}_{IEE}(t)$ for a given censoring time t.

Figure 14 is drawn to compare the NPEG with IEE and traditional empirical survival function $\hat{S}_{T_1}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_{1,i} > t)$, denoted as S_3. The NPEG (solid black line) is very similar with the IEE (dashed red line), but both estimates are smaller than the traditional nonparametric method S_3 (dotted blue line). The simulation studies in the later section also show that S_3 overestimates because the first observed event time T_1 may correspond to the second true event time W_2 for records with gaps.

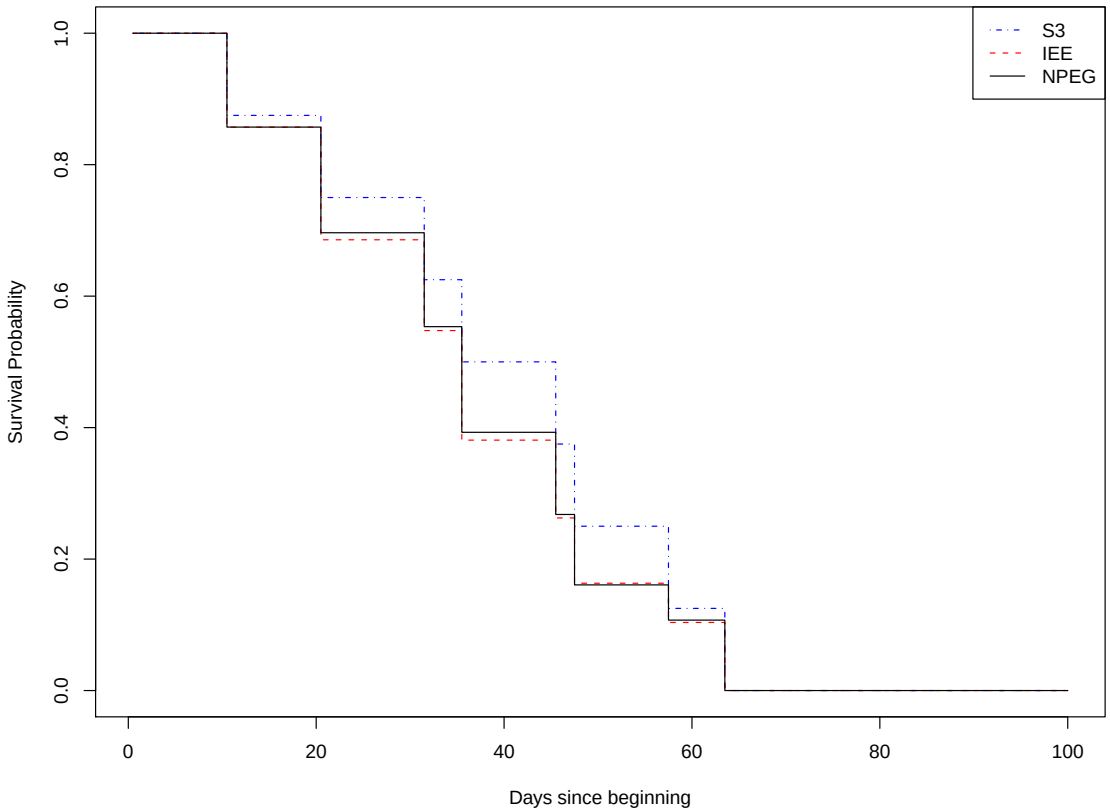


Figure 14: The estimated survival functions based on the IEE and NPEG. The dotted blue line represents the traditional empirical survival function $\hat{S}_{T_1}$ based on the observed first event times when ignoring the gaps. The dashed red line represents the estimated survival function based on the IEE method, and the solid black line represents the estimated survival function for the first true event time W_1 based on the NPEG.

2.3.2 Basic Statistics and Properties

2.3.2.1 Expectation and Bias of $\hat{S}_{NPEG}(t)$

The expectation of $\hat{S}_{NPEG}(t)$ is

$$\begin{aligned}\mathbb{E}\hat{S}_{NPEG}(t) &= S_{W_1}(t) + P(T_1 > t,\ B < W_1 \leq \min(E,t)) \\ &\quad -\mathbb{E}_{(B,E)}[P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (B, \min(E,t)])].\end{aligned}$$

Proof:

$$\begin{aligned}\mathbb{E}\hat{S}_{NPEG}(t) &= \mathbb{E}\left[\hat{S}_{T_1}(t) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i,t)])\right] \\ &= \mathbb{E}\hat{S}_{T_1}(t) - \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i,t)])\right].\end{aligned}$$

Note that $\mathbb{E}[\hat{S}_{T_1}(t)] = \frac{1}{n}\sum_{i=1}^{n} P(T_{1,i} > t) = P(T_1 > t) = S_{T_1}(t)$, because $T_{1,i}$ are *i.i.d.*. And note that

$$\begin{aligned}&\mathbb{E}[I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i,t)])] \\ =\ &\mathbb{E}\mathbb{E}[I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i,t)]) \mid (B_i, E_i)] \\ =\ &\mathbb{E}_{(B,E)}[P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (B_1, \min(E_1,t)])] \quad (\because\ (13)) \\ =\ &\mathbb{E}_{(B,E)}[P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (B, \min(E,t)])],\end{aligned}$$

because

$$\begin{aligned}&\mathbb{E}[I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i,t)]) \mid (B_i, E_i) = (b_i, e_i)] \\ =\ &\mathbb{E}[I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i,t)]] \\ =\ &P(T_{1,i} > t,\ T_{1,i'} \in (b_i, \min(e_i,t)],\ i' \neq i,\ 1 \leq i' \leq n) \quad (\because\ T_{1,i}\text{'s}\ \ i.i.d) \\ =\ &P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (b_1, \min(e_1,t)]). \qquad (13)\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\hat{S}_{NPEG}(t) &= S_{T_1}(t) - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbb{E}[I(T_{1,i} > t,\ T_{1,j} \in (B_i, \min(E_i, t)])] \\
&= S_{T_1}(t) - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbb{E}_{(B,E)}[P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (B, \min(E, t)])] \\
&= S_{T_1}(t) - \mathbb{E}_{(B,E)}[P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (B, \min(E, t)])] \\
&= S_{W_1}(t) + P(T_1 > t,\ B < W_1 \leq \min(E, t)) \\
&\quad -\mathbb{E}_{(B,E)}[P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (B, \min(E, t)])]. \quad \text{(from (8))}
\end{aligned}$$

□

Therefore, its bias is calculated as

$$\begin{aligned}
\mathbb{B}ias[\hat{S}_{NPEG}(t)] &= \mathbb{E}\hat{S}_{NPEG}(t) - S_{W_1}(t) \\
&= P(T_1 > t,\ W_1 \in (B, \min(E, t)]) \\
&\quad -\mathbb{E}_{(B,E)}[P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (B, \min(E, t)])].
\end{aligned}$$

2.3.2.2 Investigation on the Bias of the NPEG

The bias of the NPEG and its properties are studied in this section.

1. If $E - B = G_{length}$ is small, then $\mathbb{B}ias[\hat{S}_{NPEG}]$ is small.

   When $\mid G \mid$ is small, $P(W_1 \in (B, \min(E, t)])$ is also small, resulting in small $\mathbb{B}ias[\hat{S}_{NPEG}]$. The simulation results with Setting 6 show that the curve for $\hat{S}_{NPEG}$ is very close to the curve for S_1 and hence small bias is obtained, when the expectation of the gap length is small.

2. If T_1 and W_1 are close, then $\mathbb{B}ias[\hat{S}_{NPEG}]$ is small.

   If $W_1 = T_1$, Equation (9) becomes $S_{W_1}(t) = S_{T_1}(t)$, because $P(B < W_1 \leq \min(E, t) \mid T_1 > t) = 0$. Therefore, the bias of the NPEG becomes zero. When the value of W_1 is close to the value of T_1, $P(B < W_1 \leq \min(E, t),\ T_1 > t)$

will be also small, and $S_{W_1}(t)$ is almost the same as $S_{T_1}(t)$. From Equation (8), small bias of the NPEG is obtained in this case.

The simulation studies also support this. For all the settings except for Settings 4 and 7 with high portion of T_1's, which are not actually the first true ones (at least 26%), the new estimate NPEG works well regardless of the size of n. As $\mid G \mid$ gets smaller, the chance of $T_1 = W_1$ increases from Equation (5) in Section 2.2. As studied in the simulations, many T_1 values can be exactly the same as W_1 even with gaps.

3. In the gap data, the first true event time W_1 and the observed first event time T_1 are dependent from Equation (5). The dependency is inevitably caused by the incomplete knowledge of *whether* W_1 *has happened before* T_1 due to the existence of missing interval, gaps G. Naturally, the bias of the NPEG also depend on the relationship between G, W_1 and T_1, and the bias cannot be removed without the knowledge on the relationship. Appendix E studies the relationship between $P(T_1 > t)$ and $P(W_1 > t)$ based on the relationship among T_1, W_1, W_2 and G using the GLF ([14, 15]). Unless the true distributions of W_1, W_2 and G are known, the differences between $P(T_1 > t)$ and $P(W_1 > t)$ cannot be calculated. In practice, the true distributions of T_1, W_1 and W_2 are expected to be unknown and only the first event time T_1, which can be either W_1 or W_2, is observable. This study based on the GLF also supports that the lack of knowledge described above makes the handling of the bias difficult and out-of-control in empirical studies. Note that the value of the difference calculated in Appendix E can be large if the underlying distributions are misspecified.

4. The bias of the NPEG does not depend on the sample size n. Therefore, even if n is increased, the bias under certain situations may not be removed.

   In the simulation studies, experiments with a large sample of 500 are conducted,

but their results are very similar with those even with $n = 20$. It shows that the bias is not removed even if n increases.

5. For censoring time $t \leq B$, $(B, \min(E,t)] = \emptyset$ and the bias of the NPEG is zero. If the censoring time t is large enough to be $P(T_1 > t) = 0$, then $P(T_1 > t,\ W_1 \in (B, \min(E,t)]) \leq P(T_1 > t) = 0$ and hence the bias of the NPEG becomes zero. So the proposed estimate works well for small or large censoring time t. On the other hand, the IEE tends to be underestimated for such censoring times as seen in the simulation studies.

2.3.2.3 Standard Deviation

The variance of $\hat{S}_{NPEG}(t)$ is

$$\mathbb{V}ar[\hat{S}_{NPEG}(t)] \;=\; \mathbb{E}[\hat{S}^2_{NPEG}(t)] - \{\mathbb{E}[\hat{S}_{NPEG}(t)]\}^2 = F(T_1, B, E, n),$$

where

$$\begin{aligned} F(T_1,B,E,n) \;=\; & \frac{1}{n}\{S_{T_1}(t)(1-S_{T_1}(t)) \\ & -\frac{2n-3}{n-1}S_{T_1}(t)\mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B,\min(E,t)])] \\ & +\frac{n-2}{n-1}S_{T_1}(t)\mathbb{E}_{(B,E)}[\{P_{T_1}(T_1 \in (B,\min(E,t)])\}^2] \\ & +\frac{n-2}{n-1}\{S_{T_1}(t)\}^2 \\ & \qquad \times \mathbb{E}_{(B,E)}[P_{T_1}(T_{1,l} \in (B_1,\min(E_1,t)] \bigcap (B_2,\min(E_2,t)])] \\ & -\frac{2n-3}{n-1}\{S_{T_1}(t)\}^2\mathbb{E}_{(B,E)}[\{P_{T_1}(T_1 \in (B,\min(E,t)])\}^2]\Big\}. \end{aligned}$$

Proof: First, consider $\mathbb{E}[\hat{S}^2_{NPEG}(t) \mid (B,E) = (b,e)]$.

$$\begin{aligned} & \mathbb{E}[\hat{S}^2_{NPEG}(t) \mid (B,E) = (b,e)] \\ = \;& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(T_{1,i} > t) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i,t)])\right]^2 \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(T_{1,i} > t)\right]^2 \\
&\quad + \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t)])\right]^2 \\
&\quad - 2\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} I(T_{1,k} > t)\right)\right. \\
&\qquad \left.\times\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t)])\right)\right] \\
&= (A) + (B) - 2(C) \\
&\quad \text{, where } (A) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(T_{1,i} > t)\right]^2, \\
&\quad (B) = \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t)])\right]^2, \text{ and} \\
&\quad (C) = \mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} I(T_{1,k} > t)\right)\right. \\
&\qquad \left.\times\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t)])\right)\right] \\
&= \left\{\{S_{T_1}(t)\}^2 + \frac{1}{n}S_{T_1}(t)(1 - S_{T_1}(t))\right\} \\
&\quad + \frac{1}{n^2(n-1)^2}\left\{(n-1)S_{T_1}(t)\sum_{k=1}^{n} P_{T_1}(T_1 \in (b_k, \min(e_k, t)])\right. \\
&\quad + (n-2)(n-1)S_{T_1}(t)\sum_{k=1}^{n}\{P_{T_1}(T_1 \in (b_k, \min(e_k, t)])\}^2 \\
&\quad \left. + \{S_{T_1}(t)\}^2\sum_{k=1}^{n}\sum_{\substack{l=1\\ l\neq k}}^{n}\sum_{\substack{i=1\\ i\neq k}}^{n} P_{T_1}(T_{1,l} \in (b_k, \min(e_k, t)] \bigcap (b_i, \min(e_i, t)])\right.
\end{aligned}
$$

$$
+\{S_{T_1}(t)\}^2 \sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} \sum_{\substack{i=1 \\ i\neq k}}^{n} \sum_{\substack{j=1 \\ j\neq i \\ j\neq l}}^{n} \{P_{T_1}(T_1 \in (b_k, \min(e_k, t))])\}^2 \Bigg\}
$$

$$
-\frac{2}{n^2} \Bigg\{ S_{T_1}(t) \sum_{k=1}^{n} P(T_1 \in (b_k, \min(e_k, t))])
$$

$$
+\{S_{T_1}(t)\}^2 \sum_{k=1}^{n} \sum_{\substack{i=1 \\ i\neq k}}^{n} P(T_1 \in (b_i, \min(e_i, t))]) \Bigg\},
$$

because

$$
(A) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(T_{1,i} > t)\right]^2
$$

$$
= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n} I^2(T_{1,i} > t) + 2\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j>i}}^{n} I(T_{1,i} > t) I(T_{1,j} > t)\right]
$$

$$
= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n} I(T_{1,i} > t)\right] + \frac{2}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j>i}}^{n} I(T_{1,i} > t) I(T_{1,j} > t)\right]
$$

$$
= \frac{1}{n^2}\sum_{i=1}^{n} P(T_{1,i} > t) + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j>i}}^{n} P(T_{1,i} > t,\ T_{1,j} > t)
$$

$$
= \frac{1}{n^2}\sum_{i=1}^{n} P(T_{1,i} > t) + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j>i}}^{n} P(T_{1,i} > t) P(T_{1,j} > t) \quad (\because\ T_{1,i}\ i.i.d.)
$$

$$
= \frac{1}{n^2}\sum_{i=1}^{n} P(T_1 > t) + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j>i}}^{n} P(T_1 > t) P(T_1 > t) \quad (\because\ T_{1,i}\ i.i.d.)
$$

$$
= \frac{1}{n} P(T_1 > t) + \frac{(n-1)}{n}\{P(T_1 > t)\}^2
$$

$$
= \frac{1}{n} S_{T_1}(t) + \{S_{T_1}(t)\}^2 - \frac{1}{n}\{S_{T_1}(t)\}^2
$$

$$
= S_{T_1}(t)^2 + \frac{1}{n} S_{T_1}(t)(1 - S_{T_1}(t)),
$$

$$
(B) = \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t))])\right]^2
$$

$$
\begin{aligned}
&= \frac{1}{n^2(n-1)^2}\mathbb{E}\left[\left\{\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n} I(T_{1,k}>t,\ T_{1,l}\in(b_k,\min(e_k,t)])\right\}\right.\\
&\quad\left.\times\left\{\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} I(T_{1,i}>t,\ T_{1,j}\in(b_i,\min(e_i,t)])\right\}\right]\\
&= \frac{1}{n^2(n-1)^2}\mathbb{E}\,[\\
&\quad\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i=k}}^{n}\sum_{\substack{j=1\\j\neq i\\j=l}}^{n} I(T_{1,k}>t,\ T_{1,l}\in(b_k,\min(e_k,t)])\\
&\qquad\qquad\times I(T_{1,i}>t,\ T_{1,j}\in(b_i,\min(e_i,t)])\\
&\quad+\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i=k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n} I(T_{1,k}>t,\ T_{1,l}\in(b_k,\min(e_k,t)])\\
&\qquad\qquad\times I(T_{1,i}>t,\ T_{1,j}\in(b_i,\min(e_i,t)])\\
&\quad+\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j=l}}^{n} I(T_{1,k}>t,\ T_{1,l}\in(b_k,\min(e_k,t)])\\
&\qquad\qquad\times I(T_{1,i}>t,\ T_{1,j}\in(b_i,\min(e_i,t)])\\
&\quad+\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n} I(T_{1,k}>t,\ T_{1,l}\in(b_k,\min(e_k,t)])\\
&\qquad\qquad\times\ I(T_{1,i}>t,\ T_{1,j}\in(b_i,\min(e_i,t)])]\\
&= \frac{1}{n^2(n-1)^2}\mathbb{E}\,[(B1)+(B2)+(B3)+(B4)]\\
&= \frac{1}{n^2(n-1)^2}\left[(n-1)P(T_1>t)\sum_{k=1}^{n}P_{T_1}(T_1\in(b_k,\min(e_k,t)])\right.\\
&\quad+(n-2)(n-1)P(T_1>t)\sum_{k=1}^{n}\{P_{T_1}(T_1\in(b_k,\min(e_k,t)])\}^2\\
&\quad\left.+\{P_{T_1}(T_1>t)\}^2\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}P_{T_1}(T_{1,l}\in(b_k,\min(e_k,t)]\bigcap(b_i,\min(e_i,t)])\right.
\end{aligned}
$$

$$+\{P_{T_1}(T_1 > t)\}^2 \sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} \sum_{\substack{i=1 \\ i\neq k}}^{n} \sum_{\substack{j=1 \\ j\neq i \\ j\neq l}}^{n} \{P_{T_1}(T_1 \in (b_k, \min(e_k, t)])\}^2 \Bigg],$$

$$\begin{aligned}
& \mathbb{E}(B1) \\
= \; & \mathbb{E}\Bigg[\sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} \sum_{\substack{i=1 \\ i=k}}^{n} \sum_{\substack{j=1 \\ j\neq i \\ j=l}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)]) \\
& \qquad \times I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t)])] \\
= \; & \mathbb{E}\Bigg[\sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)]) \\
& \qquad \times I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)])] \\
= \; & \mathbb{E}\Bigg[\sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)])\Bigg] \\
= \; & \sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} P(T_1 > t) P(T_1 \in (b_k, \min(e_k, t)]) \\
= \; & (n-1) P(T_1 > t) \sum_{k=1}^{n} P_{T_1}(T_1 \in (b_k, \min(e_k, t)]),
\end{aligned}$$

$$\begin{aligned}
& \mathbb{E}(B2) \\
= \; & \mathbb{E}\Bigg[\sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} \sum_{\substack{i=1 \\ i=k}}^{n} \sum_{\substack{j=1 \\ j\neq i \\ j\neq l}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)]) \\
& \qquad \times I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t)])] \\
= \; & \mathbb{E}\Bigg[\sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} \sum_{\substack{j=1 \\ j\neq i \\ j\neq l}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)]) \\
& \qquad \times I(T_{1,k} > t,\ T_{1,j} \in (b_k, \min(e_k, t)])]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)],\ T_{1,j} \in (b_k, \min(e_k, t)])\right] \\
&= \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n} P(T_1 > t)P(T_1 \in (b_k, \min(e_k, t)])P(T_1 \in (b_k, \min(e_k, t)]) \\
&= \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n} P(T_1 > t)\{P(T_1 \in (b_k, \min(e_k, t)])\}^2 \\
&= (n-2)(n-1)P(T_1 > t)\sum_{k=1}^{n}\{P_{T_1}(T_1 \in (b_k, \min(e_k, t)])\}^2,
\end{aligned}
$$

$\mathbb{E}(B3)$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j=l}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)]) \right. \\
&\qquad\qquad \left. \times I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t)])\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)],\ T_{1,i} > t,\ T_{1,l} \in (b_i, \min(e_i, t)])\right] \\
&= \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n} P(T_1 > t)P(T_1 > t)P(T_{1,l} \in (b_k, \min(e_k, t)]\bigcap(b_i, \min(e_i, t)]) \\
&= \{P(T_1 > t)\}^2\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n} P_{T_1}(T_{1,l} \in (b_k, \min(e_k, t)]\bigcap(b_i, \min(e_i, t)]),
\end{aligned}
$$

$\mathbb{E}(B4)$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n} I(T_{1,k} > t,\ T_{1,l} \in (b_k, \min(e_k, t)]) \right. \\
&\qquad\qquad \left. \times I(T_{1,i} > t,\ T_{1,j} \in (b_i, \min(e_i, t)])\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n} I(T_{1,k}>t,\ T_{1,i}>t,\right.\\
&\qquad\qquad \left. T_{1,l}\in(b_k,\min(e_k,t)],\ T_{1,j}\in(b_i,\min(e_i,t)])\right]\\
&= \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n} P(T_1>t)P(T_1>t)P(T_1\in(b_k,\min(e_k,t)])\\
&\qquad\qquad \times P(T_1\in(b_i,\min(e_i,t)])\\
&= \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n}\{P(T_1>t)\}^2\{P(T_1\in(b_k,\min(e_k,t)])\}^2\\
&= \{P(T_1>t)\}^2\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{i=1\\i\neq k}}^{n}\sum_{\substack{j=1\\j\neq i\\j\neq l}}^{n}\{P_{T_1}(T_1\in(b_k,\min(e_k,t)])\}^2,\ \text{and}\\
&\quad (C)\\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} I(T_{1,k}>t)\right)\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} I(T_{1,i}>t,\ T_{1,j}\in(b_i,\min(e_i,t)])\right)\right]\\
&= \frac{1}{n^2(n-1)}\mathbb{E}\left[\left(\sum_{k=1}^{n} I(T_{1,k}>t)\right)\left(\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} I(T_{1,i}>t,\ T_{1,j}\in(b_i,\min(e_i,t)])\right)\right]\\
&= \frac{1}{n^2(n-1)}\mathbb{E}\left[\sum_{k=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} I(T_{1,k}>t,\ T_{1,j}\in(b_k,\min(e_k,t)])\right.\\
&\qquad\qquad \left.+\sum_{k=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{i=1}^{n} I(T_{1,k}>t,\ T_{1,i}>t,\ T_{1,j}\in(b_i,\min(e_i,t)])\right]\\
&= \frac{1}{n^2(n-1)}\left\{\sum_{k=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} P(T_1>t)P(T_1\in(b_k,\min(e_k,t)])\right.
\end{aligned}
$$

$$
\left. + \sum_{k=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{i=1}^{n} P(T_1 > t) P(T_1 > t) P(T_1 \in (b_i, \min(e_i, t)]) \right\}
$$

$$
\begin{aligned}
= & \frac{1}{n^2(n-1)} \left[(n-1) P(T_1 > t) \sum_{k=1}^{n} P(T_1 \in (b_k, \min(e_k, t)]) \right. \\
& \left. + (n-1) \{P(T_1 > t)\}^2 \sum_{k=1}^{n} \sum_{\substack{i=1 \\ i \neq k}}^{n} P(T_1 \in (b_i, \min(e_i, t)]) \right] \\
= & \frac{1}{n^2} \left[P(T_1 > t) \sum_{k=1}^{n} P(T_1 \in (b_k, \min(e_k, t)]) \right. \\
& \left. + \{P(T_1 > t)\}^2 \sum_{i=1}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} P(T_1 \in (b_i, \min(e_i, t)]) \right] \\
= & \frac{1}{n^2} \left[P(T_1 > t) \sum_{k=1}^{n} P(T_1 \in (b_k, \min(e_k, t)]) \right. \\
& \left. + (n-1) \{P(T_1 > t)\}^2 \sum_{\substack{k=1 \\ k \neq i}}^{n} P(T_1 \in (b_i, \min(e_i, t)]) \right].
\end{aligned}
$$

Thus, the following is driven:

$$
\begin{aligned}
& \mathbb{E}[\hat{S}^2_{NPEG}(t)] = \mathbb{E}_{(B,E)} \mathbb{E}_{T_1}[\hat{S}^2_{NPEG}(t) \mid (B, E)] \\
= & \{S_{T_1}(t)\}^2 + \frac{1}{n} S_{T_1}(t)(1 - S_{T_1}(t)) \\
& + \frac{1}{n^2(n-1)^2} [(n-1) n S_{T_1}(t) \mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E, t)])] \\
& + (n-2)(n-1) n S_{T_1}(t) \mathbb{E}_{(B,E)}[\{P_{T_1}(T_1 \in (B, \min(E, t)])\}^2] \\
& + (n-2)(n-1) n \{S_{T_1}(t)\}^2 \mathbb{E}_{(B,E)}[P_{T_1}(T_{1,l} \in (B_1, \min(E_1, t)] \bigcap (B_2, \min(E_2, t)])] \\
& + (n-1) n (n^2 - 3n + 3) \{S_{T_1}(t)\}^2 \{\mathbb{E}_{(B,E)} P_{T_1}(T_1 \in (B, \min(E, t)])\}^2 \\
& - \frac{2}{n} \left[S_{T_1}(t) \mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E, t)])] \right. \\
& \left. + (n-1) \{S_{T_1}(t)\}^2 \mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E, t)])] \right] \\
= & \{S_{T_1}(t)\}^2 + \frac{1}{n} S_{T_1}(t)(1 - S_{T_1}(t)) \\
& + \frac{1}{n(n-1)} S_{T_1}(t) \mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E, t)])]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{n-2}{n(n-1)}S_{T_1}(t)\mathbb{E}_{(B,E)}[\{P_{T_1}(T_1 \in (B, \min(E,t)])\}^2] \\
&+\frac{n-2}{n(n-1)}\{S_{T_1}(t)\}^2\mathbb{E}_{(B,E)}[P_{T_1}(T_{1,l} \in (B_1, \min(E_1,t)] \bigcap (B_2, \min(E_2,t)])] \\
&+\frac{n^2-3n+3}{n(n-1)}\{S_{T_1}(t)\}^2\{\mathbb{E}_{(B,E)}P_{T_1}(T_1 \in (B, \min(E,t)])\}^2 \\
&-\frac{2}{n}S_{T_1}(t)\mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E,t)])] \\
&-\frac{2(n-1)}{n}\{S_{T_1}(t)\}^2\mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E,t)])] \\
= \;& \{S_{T_1}(t)\}^2 + \{S_{T_1}(t)\}^2\{\mathbb{E}_{(B,E)}P_{T_1}(T_1 \in (B, \min(E,t)])\}^2 \\
&-2\{S_{T_1}(t)\}^2\mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E,t)])] + F(T_1, B, E, n).
\end{aligned}
$$

Therefore, the variance of $\hat{S}_{NPEG}(t)$ is calculated as

$$
\begin{aligned}
\mathbb{V}ar[\hat{S}_{NPEG}(t)] &= \mathbb{E}[\hat{S}^2_{NPEG}(t)] - \{\mathbb{E}[\hat{S}_{NPEG}(t)]\}^2 \\
&= \{S_{T_1}(t)\}^2 + \{S_{T_1}(t)\}^2\{\mathbb{E}_{(B,E)}P_{T_1}(T_1 \in (B, \min(E,t)])\}^2 \\
&\quad -2\{S_{T_1}(t)\}^2\mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E,t)])] + F(T_1, B, E, n) \\
&\quad -\{S_{T_1}(t) - S_{T_1}(t)\mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E,t)]]\}^2 \\
&= F(T_1, B, E, n). \qquad \square
\end{aligned}
$$

Remarks: The variance of $\hat{S}_{NPEG}(t)$ is $o(1)$, because

$$
\begin{aligned}
& F(T_1, B, E, n) \\
= \;& \frac{1}{n}\Bigg\{ S_{T_1}(t)(1 - S_{T_1}(t)) - \frac{2n-3}{n-1}S_{T_1}(t)\mathbb{E}_{(B,E)}[P_{T_1}(T_1 \in (B, \min(E,t)])] \\
&+\frac{n-2}{n-1}S_{T_1}(t)\mathbb{E}_{(B,E)}[\{P_{T_1}(T_1 \in (B, \min(E,t)])\}^2] \\
&+\frac{n-2}{n-1}\{S_{T_1}(t)\}^2\mathbb{E}_{(B,E)}[P_{T_1}(T_{1,l} \in (B_1, \min(E_1,t)] \bigcap (B_2, \min(E_2,t)])] \\
&-\frac{2n-3}{n-1}\{S_{T_1}(t)\}^2\mathbb{E}_{(B,E)}[\{P_{T_1}(T_1 \in (B, \min(E,t)])\}^2]\Bigg\} \\
= \;& o(1) \cdot O(1) \\
&\Big(\because \; \frac{1}{n} = o(1), \; \frac{2n-3}{n-1} = O(1), \; \frac{n-2}{n-1} = O(1) \\
&\qquad 0 \le S_{T_1}(t) \le 1, \; \& \; 0 \le \mathbb{E}_{(B,E)}P_{T_1}(\cdot) \le 1\Big)
\end{aligned}
$$

$$= \; o(1).$$

$\mathbb{M}SE(\hat{S}_{NPEG}(t)) = \mathbb{V}ar(\hat{S}_{NPEG}(t)) + \{\mathbb{B}ias(\hat{S}_{NPEG}(t))\}^2$ cannot go to zero even though the sample size goes to infinity, because its bias does not go to zero.

2.3.2.4 Other Properties

For the consistency of $\hat{S}_{NPEG}(t)$, consider the following:

$$\begin{aligned}
& P\left(\mid \hat{S}_{NPEG}(t) - S_{W_1}(t) \mid \geq \epsilon\right) \\
\leq \; & P\left(\mid \hat{S}_{NPEG}(t) - \mathbb{E}[\hat{S}_{NPEG}(t)] \mid \geq \frac{\epsilon}{2}\right) + P\left(\mid \mathbb{E}[\hat{S}_{NPEG}(t)] - S_{W_1}(t) \mid \geq \frac{\epsilon}{2}\right) \\
\leq \; & \frac{\mathbb{V}ar[\hat{S}_{NPEG}(t)]}{(\epsilon/2)^2} + \frac{(\mathbb{B}ias[\hat{S}_{NPEG}(t)])^2}{(\epsilon/2)^2} \quad \text{(by Chebyshev's inequality)} \\
= \; & o(1) + \frac{(\mathbb{B}ias[\hat{S}_{NPEG}(t)])^2}{(\epsilon/2)^2}.
\end{aligned}$$

Remarks: If $\mathbb{B}ias[\hat{S}_{NPEG}(t)] \longrightarrow 0$ as $n \longrightarrow \infty$, $\hat{S}_{NPEG}(t)$ becomes consistent. However, as discussed in Section 2.3.2.1, the bias of the NPEG is not $o(1)$ and thus $\hat{S}_{NPEG}$ may not be consistent. Since the NPEG is a biased estimate, it is not meaningful to study its asymptotic distribution.

2.4 Simulation Study

A simulation study is conducted to evaluate the quality of the proposed estimate. The new estimate, NPEG is compared with the following values obtained from existing nonparametric estimation methods:

- S_1 defined as the underlying true survival function $S_{W_1}(t)$ of the first true event time W_1 in each simulation setting.
- $\hat{S}_2$ defined as the classical empirical survival function $\hat{S}_{W_1}(t) = \frac{1}{n}\sum I(W_{1,i} > t)$ based on the first true event time W_1 in each simulation setting. Note that these values cannot be observable in real experiments.

- $\hat{S}_3$ defined as the empirical survival function $\hat{S}_{T_1}(t) = \frac{1}{n}\sum I(T_{1,i} > t)$ based on the observed first event time T_1 which might or might not be the first true event time W_1 in a simulation data. Note that the $\hat{S}_3$ becomes the traditional empirical survival function based on the observed first event time T_1, when the gaps are ignored. We call this method as the ignoring-gap method.

- $\hat{S}_{IEE}$ introduced in Section 2.2. If each simulation uses repetition, the mean of results is used as the IEE.

For each simulation setting, each sample subject will consist of four independent random variables $(G_{begin}, G_{length}, W_1, W_{elapse})$ to generate. Let a random variable G_{begin} be the starting time of gap, and a random variable G_{length} be the length of gap. Then, the ending time of gap, G_{end} becomes $G_{begin} + G_{length}$. Let random variable W_1 be the first true event time, and a random variable W_{elapse} be the elapsed time between the first true event time W_1 and the second true event time W_2. Thus, the W_2 becomes $W_1 + W_{elapse}$. Two sets of simulation experiments are studied based on two different underlying distributions (exponential and Weibull distributions) for the four random variables. The reason to consider only exponential and Weibull distributions is just that they are commonly used in lifetime data analysis. The details of their parameter values are given in Tables 19 and 20. In each set of simulation experiments, two situations (small samples ($n = 20$) and large samples ($n = 500$)) are considered. When n is small, the same setting is repeated $r = 300$ times to compare the mean of all the candidate estimates described in the first paragraph. With the large sample size, only $r = 3$ simulation replications are conducted due to the computational burden to examine the property of large sample consistency of the estimates.

An exponential distribution is defined with a parameter called the rate parameter, while a Weibull distribution is defined with two parameters, the shape parameter and the scale parameter. Note that the Weibull distribution is related to exponential

Table 19: Parameter values of the four random variables introduced in Yang's thesis after correction.

Random Variables	Distribution	Parameter	Distribution	Parameter
G_{begin}	$Exp(\theta_1)$	$\theta_1 = 5$	$Weibull(2, \lambda_1)$	$\lambda_1 = 5.64$
G_{length}	$Exp(\theta_2)$	$\theta_2 = 15$	$Weibull(2, \lambda_2)$	$\lambda_2 = 15.93$
W_1	$Exp(\alpha_1)$	$\alpha_1 = 55$	$Weibull(2, \delta_1)$	$\delta_1 = 65.06$
W_{elapse}	$Exp(\alpha_2)$	$\alpha_2 = 54.99$	$Weibull(2, \delta_2)$	$\delta_2 = 62.05$

distribution, when its shape parameter value is set to 1. Here the shape parameter value of Weibull distributions is set to 2 for comparison with the settings using Exponential distributions. The parameter values used in Yang's thesis ([38]) are displayed in Table 19. Yang generated the random variables G_{begin}, G_{length}, W_1, and W_{elapse} with their expectation 5, 15, 55, and 54.99, respectively. Note that Yang used wrong parameter values for Weibull distributions, so their values are corrected here. These settings Yang provided are used as the baseline in this section. Each parameter value is changed at each time for comparisons. All settings studied are described in Table 20.

In the gap analysis study, it is assumed that the probability of two events happening in the gap is negligible. Since parameters should be selected to satisfy this assumption, the selected parameter values are evaluated based on the following three probabilities:

- $P_1 = Pr(\text{a gap occurs}) = 1 - Pr(W_1 \leq G_{begin})$,
- $P_2 = Pr(\text{the observed first event time is not the first true event time})$

  $= Pr(T_1 \neq W_1,\ T_1 = W_2)$

  $= Pr(W_1 \text{ is in the gap } G \text{ and } W_2 \text{ is outside of the gap})$

  $= Pr(G_{begin} \leq W_1 < G_{end},\ W_2 \geq G_{end})$

  $= Pr(G_{begin} \leq W_1 < G_{begin} + G_{length} \leq W_1 + W_{elapse})$, and

Table 20: Parameter values of the four random variables for each simulation setting.

	Exp											
	Parameter				Expectation				Variance			
	θ_1	θ_2	α_1	α_2	θ_1	θ_2	α_1	α_2	θ_1	θ_2	α_1	α_2
1	5	15	55	54.49	5	15	55	54.49	25	225	3025	2969.16
2	5	15	55	**5**	5	15	55	**5**	25	225	3025	**25**
3	5	15	55	**100**	5	15	55	**100**	25	225	3025	**10000**
4	5	15	**5**	54.49	5	15	**5**	54.49	25	225	**25**	2969.16
5	5	15	**100**	54.49	5	15	**100**	54.49	25	225	**10000**	2969.16
6	5	**5**	55	54.49	5	**5**	55	54.49	25	**25**	3025	2969.16
7	5	**45**	55	54.49	5	**45**	55	54.49	25	**2025**	3025	2969.16
8	**2**	15	55	54.49	**2**	15	55	54.49	**4**	225	3025	2969.16
9	**12.5**	15	55	54.49	**12.5**	15	55	54.49	**156.25**	225	3025	2969.16
	Weibull											
	Parameter				Expectation				Variance			
	λ_1	λ_2	δ_1	δ_2	λ_1	λ_2	δ_1	δ_2	λ_1	λ_2	δ_1	δ_2
1	5.64	16.93	62.06	62.05	5	15	55	54.49	6.83	61.51	826.53	826.26
2	5.64	16.93	62.06	**5.64**	5	15	55	**5**	6.83	61.51	826.53	**6.83**
3	5.64	16.93	62.06	**112.84**	5	15	55	**100**	6.83	61.51	826.53	**2732.50**
4	5.64	16.93	**5.64**	62.05	5	15	**5**	54.49	6.83	61.51	**6.83**	826.26
5	5.64	16.93	**112.84**	62.05	5	15	**100**	54.49	6.83	61.51	**2732.50**	826.26
6	5.64	**5.64**	62.06	62.05	5	**5**	55	54.49	6.83	**6.83**	826.53	826.26
7	5.64	**50.78**	62.06	62.05	5	**45**	55	54.49	6.83	**553.37**	826.53	826.26
8	**2.26**	16.93	62.06	62.05	**2**	15	55	54.49	**1.10**	61.51	826.53	826.26
9	**14.1**	16.93	62.06	62.05	**12.5**	15	55	54.49	**42.66**	61.51	826.53	826.26

- $P_3 = Pr(\text{two true events fall in the gap})$

  $= Pr(G_{begin} \leq W1 \leq W_2 < G_{end})$

  $= Pr(G_{begin} \leq W_1 \leq W_1 + W_{elapse} < G_{begin} + G_{length})$.

These probabilities and the distributions of the four random variables affect the estimates of the survival function, their unbiasness, as well as their consistency. The simulation studies explore how all these factors affect the performance of the estimates. Moreover, they also explore the behaviors of the NPEG biaseness and consistency, by studying the knowledge on the relationship between the first true event time and the observed first event time.

The parameter values for each setting but with different underlying distribution are adjusted to have the same means of the four random variables to compare the effect of distributions and the effect of variances of the random variables. From the 9 simulation settings for fixed underlying distributions, the effects of changing parameter values for the four random variables and three probabilities $P_1 - P_3$ are compared. By studying Settings 2-1-3, the expectation of W_{elapse} increases. Thus, the expectation of W_2 increases so that the values of W_1 and W_2 become further, and the probability P_2 becomes higher but P_3 smaller. Therefore, the effect of P_2 and P_3 can be studied by comparing Settings 1, 2 and 3. Similarly, the expectation of W_1 increases from Settings 4-1-5. Although the expectation of W_1 increases, the parameters of G_{begin} and G_{length} do not change at all. Therefore, it is expected to have bigger P_1 and lower P_3. By examining Settings 6-1-7, the expectation of G_{length} increases and hence more wider gaps are generated. Therefore, the chance P_3 of W_1 and W_2 falling into $(G_{begin}, G_{length}]$ increases. The expectation of G_{begin} increases from Settings 8-1-9 while keeping other expectations same. Thus, P_1 definitely decreases. The empirical probabilities $\hat{P}_1$, $\hat{P}_2$, and $\hat{P}_3$ for each simulation setting are in Table 21. These empirical probabilities have the same patterns as the parameter values are changed.

Each sample data is generated as follows:

1. Generate the four random variables, ($G_{begin}, G_{length}, W_1$, and W_{elapse}) according to a given distribution, their parameter values, sample size n and r for repetitions.

2. Examine the generated data to exclude subjects that do not satisfy the assumptions described in Section 2.2. That is, remove the data points from the simulation dataset, if two values W_1 and $W_1 + W_{elapse}$ fall into the corresponding gap $(G_{begin}, G_{begin} + G_{length}]$. Those deleted data points are not used for simplicity, when calculating $\hat{S}_3$, $\hat{S}_{IEE}$ and $\hat{S}_{NPEG}$.

3. Find out the observed first event time of each subject as follows.

   - If the first true event time W_1 is observed before the gap (i.e., $W_1 < G_{begin}$) or if the gap has zero length (i.e.,$G_{length} = 0$), then it is a no-gap case and the observed first event time T_1 is exactly the first true event time W_1.
   - If the first true event time W_1 falls in the gap (i.e., $G_{begin} < W_1 \leq G_{end}$) and the second true event time W_2 happens after the gap (i.e., $W_2 > G_{end}$), then T_1 becomes W_2.
   - If the first true event time happens after the gap (i.e., $W_1 > G_{end}$), then T_1 becomes W_1.

   For the NPEG and IEE, it is assumed that the probability of two events falling into the gap is negligible. So if the first and second true event times of a subject are in the gap (i.e., $G_{begin} < W_1,\ W_2 \leq G_{end}$), then the subject is ignored in the calculations.

4. Calculate each estimate (S_1, $\hat{S}_2$, $\hat{S}_3$, $\hat{S}_{IEE}$ or $\hat{S}_{NPEG}$). For the IEE, the data should be ordered properly by the observed first event time T_1 before the calculation.

Table 21: Empirical probabilities $\hat{P}_1,\ \hat{P}_2,\ \hat{P}_3$ for all simulation settings.

Exp						
	$(n,r)=(20,300)$			$(n,r)=(500,3)$		
Setting	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}_3$	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}_3$
1	91.2356	16.1474	4.1167	91.2986	15.1998	3.4667
2	89.6977	5.2253	15.2167	89.7139	5.8016	13.8000
3	91.8239	17.5033	2.3000	92.3402	18.2596	2.5333
4	46.4843	33.3631	8.0833	48.3864	31.9450	9.2000
5	95.1179	9.3985	2.8333	95.2059	9.3102	2.6000
6	91.4214	7.0468	0.7500	91.4832	7.1762	0.6000
7	90.5368	26.7245	18.5833	89.4033	26.8294	20.7333
8	96.5172	16.8289	4.2500	96.6526	17.0075	4.4000
9	80.7292	13.8457	3.9500	80.6317	14.5460	3.2667
Weibull						
	$(n,r)=(20,300)$			$(n,r)=(500,3)$		
Setting	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}_3$	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}_3$
1	99.0632	9.9518	0.2500	99.0640	9.0881	0.2667
2	99.2266	3.6016	6.4500	99.0044	4.2115	6.6000
3	99.2491	10.3597	0.0667	99.6663	9.8061	0.0667
4	48.5916	46.2262	2.6667	47.6304	44.6166	2.7333
5	99.7483	3.0684	0.1667	99.7333	2.6667	0.0000
6	99.3667	2.5000	0.0167	99.0667	1.7333	0.0000
7	99.1214	40.1576	8.6500	99.3387	39.9848	9.4667
8	99.9167	8.1500	0.2000	99.7996	7.2812	0.2000
9	94.7867	13.0162	0.5500	94.9551	13.8493	0.8667

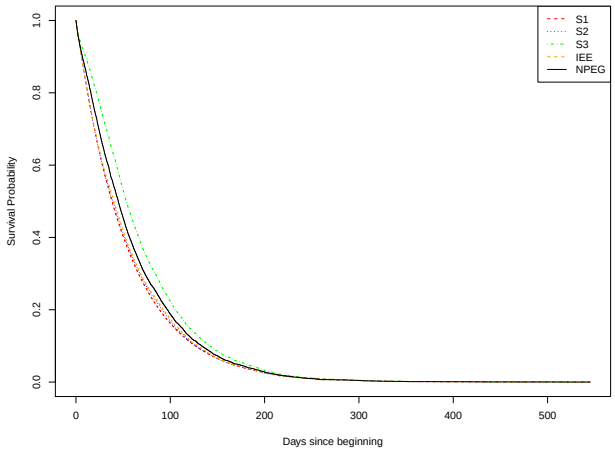
5. Plot all the estimated survival functions along with the true survival curve for S_{W_1}.

Each plot of Figures 15-26 contains a total of five curves. One is for S_1 (dashed red line) which is the underlying true survival function. With the values given in Table 20, the curve for S_1 can be drawn. Another curve is for $\hat{S}_2$ (dotted blue line), the average value of r empirical survival functions for the first true event time W_1. The empirical estimate based on the first true event W_1 is known as the Uniformly Minimum Variance Unbiased Estimate (UMVUE) ([27]). Therefore, the curve for $\hat{S}_2$ in each simulation setting is expected to be close to the curve for S_1. Each plot also

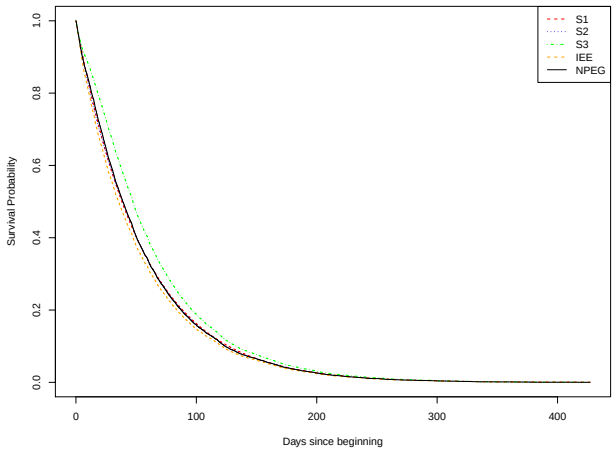
contains the curve for $\hat{S}_3$ (dotted and dashed green line), which is the average value of r empirical survival functions $\hat{S}_{T_1}$ based on the observed first event times T_1. The observed first event time might not be the first true event times but the second true event times due to the possible gaps. Thus, the simulation studies apparently obtain overestimated $\hat{S}_{T_1}$. Moreover, if more observed first event times are not the first true event times, the difference between $\hat{S}_2$ and $\hat{S}_3$ increases. Another curve is for the IEE (dashed orange line). The last curve is for the proposed method, NPEG (solid black line). From the definition of the NPEG, Equation (12), $\hat{S}_{NPEG}$ is definitely smaller than $\hat{S}_3$.

The probability P_1 is calculated to see how many subjects contain gaps. All simulation settings except Setting 4's have at least 80% chance to have gaps. However, a lower chance of having gaps does not mean having better information on the relationship between the first true event time and the observed first event time which can be amounted by P_2.

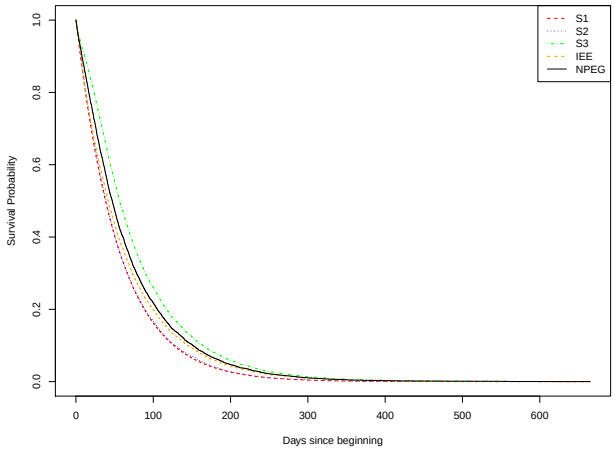
The probability P_2 provide information about how many observed first events are not the first true ones. As P_2 becomes larger, more observed first events T_1 are not the first true ones W_1 but the second true ones W_2. Therefore, when P_2 is large, the empirical survival function $\hat{S}_3$ based on T_1 trivially do not work well as an estimated survival function for W_1, because many T_1's are not W_1 but W_2. Since more information about W_1 is obtained from T_1 as P_2 decreases, the performances of the NPEG and IEE are expected to be improved. All simulation settings except Settings 4 and 7 demonstrate both curves for the NPEG and IEE are very close to the true survival curve even with the small sample size and closer to the true value than the curve for $\hat{S}_3$, nonetheless their $\hat{P}_2$'s are not small (at most 18.26%). But with a little higher $\hat{P}_2$ values (about 25 to 40%) for Setting 7's (with the high expectation of the gap length), the IEE works better than the NPEG and both work better than the ignoring-gap method.



(a) Setting 1

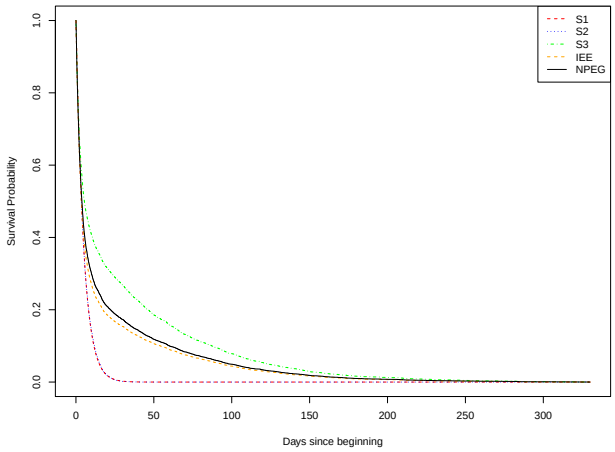


(b) Setting 2

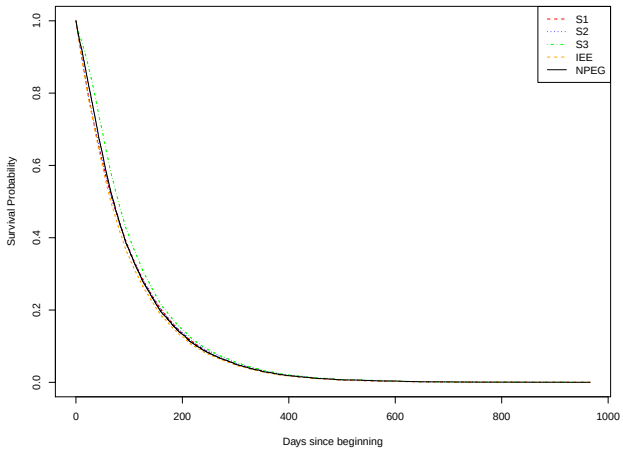


(c) Setting 3

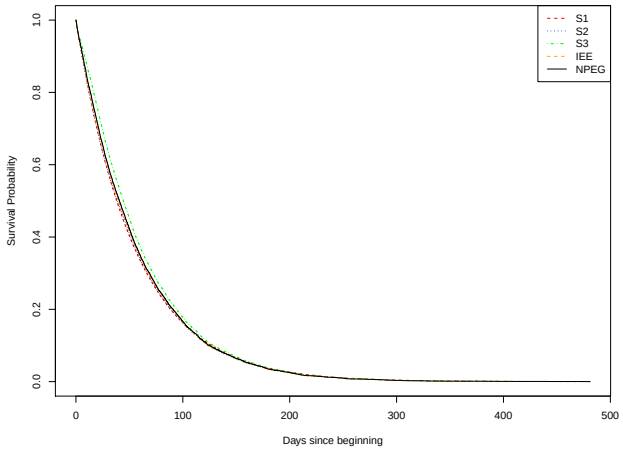
Figure 15: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 1 - 3 with all exponential distributions, $n = 20$ and $r = 300$. S_1 is the true survival function.



(a) Setting 4

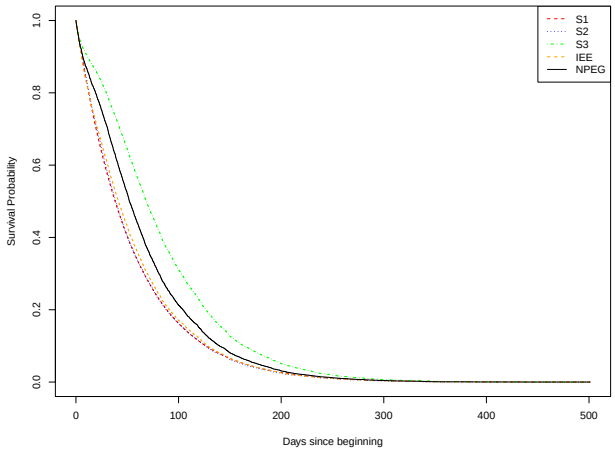


(b) Setting 5

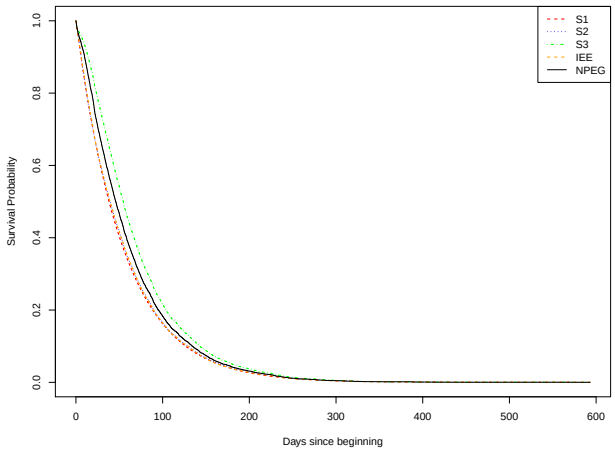


(c) Setting 6

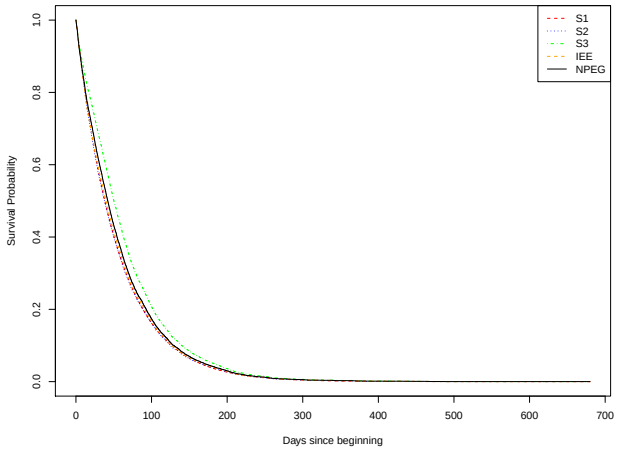
Figure 16: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 4 - 6 with all exponential distributions, $n = 20$ and $r = 300$. S_1 is the true survival function.



(a) Setting 7

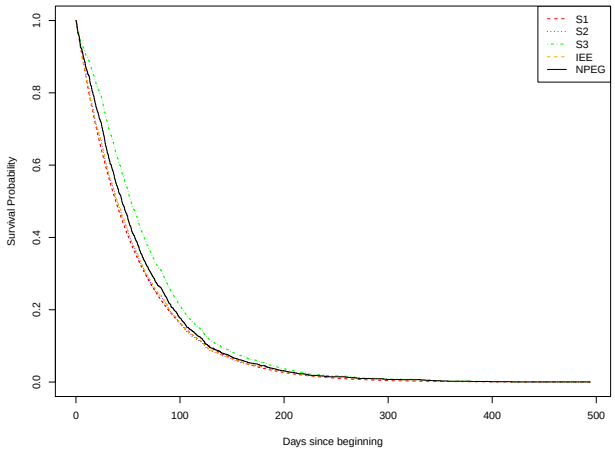


(b) Setting 8

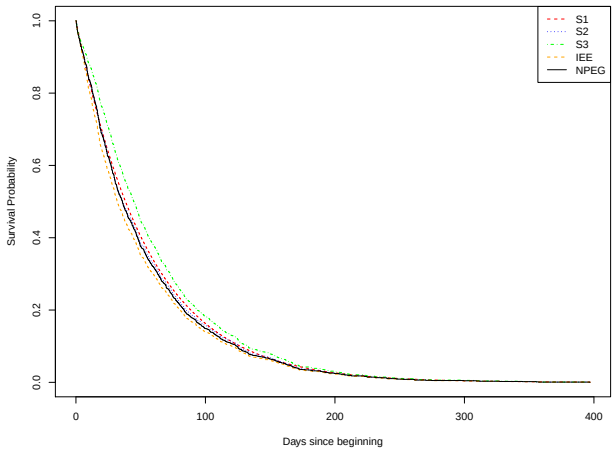


(c) Setting 9

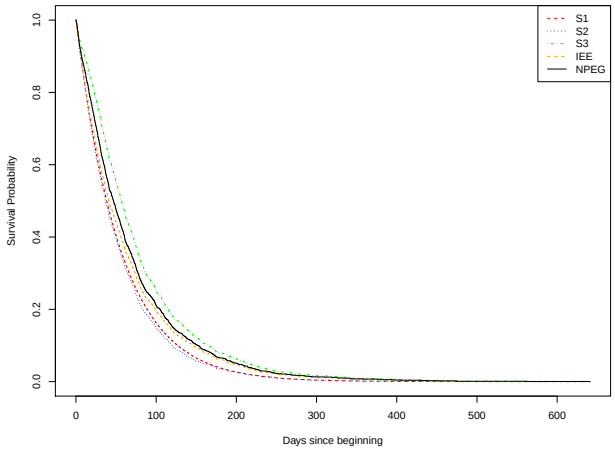
Figure 17: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 7 - 9 with all exponential distributions, $n = 20$ and $r = 300$. S_1 is the true survival function.



(a) Setting 1

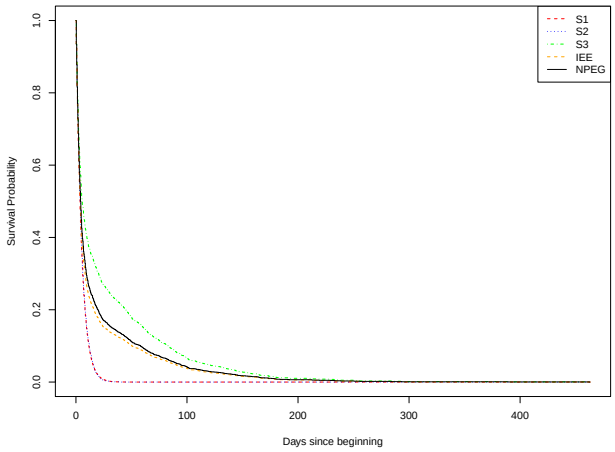


(b) Setting 2

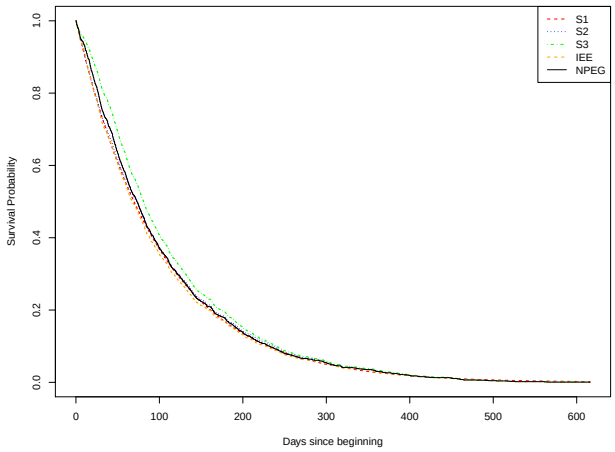


(c) Setting 3

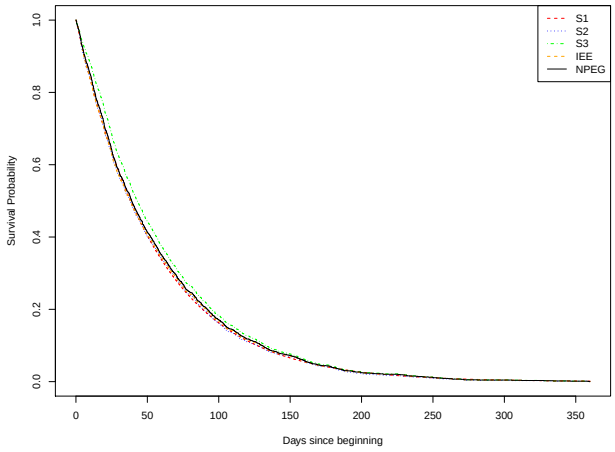
Figure 18: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 1 - 3 with all exponential distributions, $n = 500$ and $r = 3$. S_1 is the true survival function.



(a) Setting 4

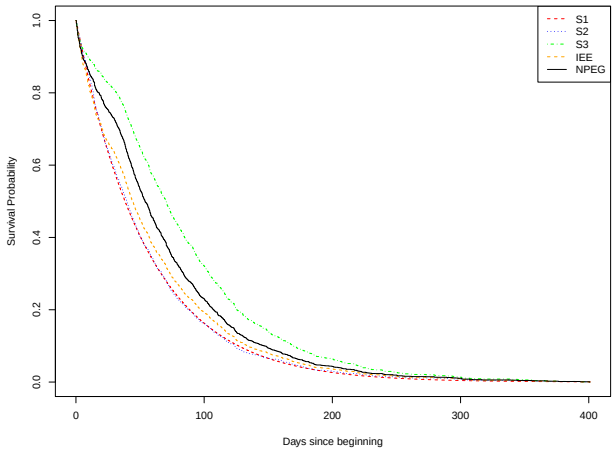


(b) Setting 5

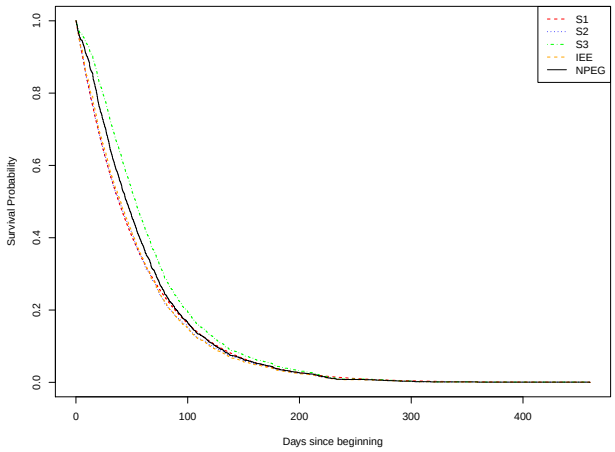


(c) Setting 6

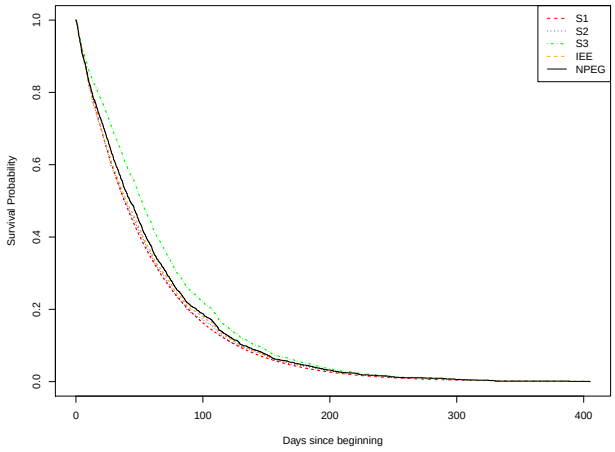
Figure 19: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 4 - 6 with all exponential distributions, $n = 500$ and $r = 3$. S_1 is the true survival function.



(a) Setting 7

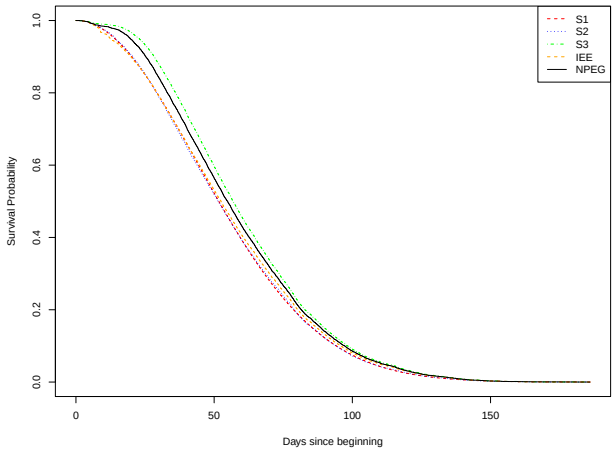


(b) Setting 8

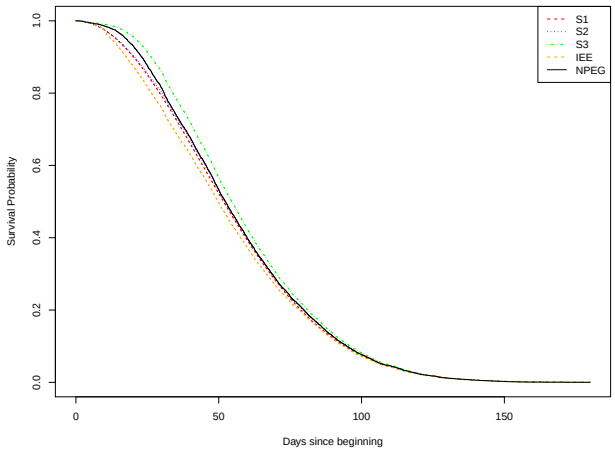


(c) Setting 9

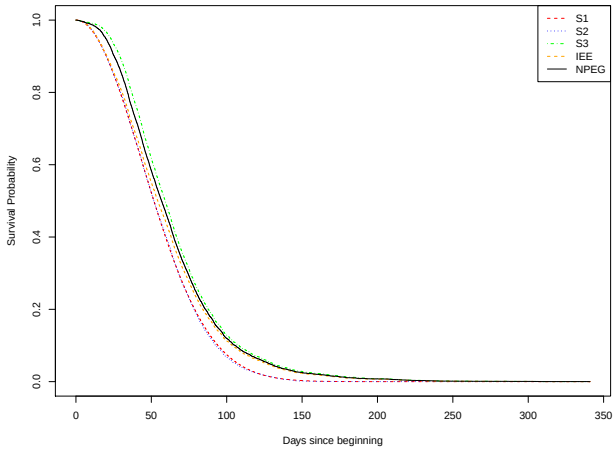
Figure 20: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 7 - 9 with all exponential distributions, $n = 500$ and $r = 3$. S_1 is the true survival function.



(a) Setting 1

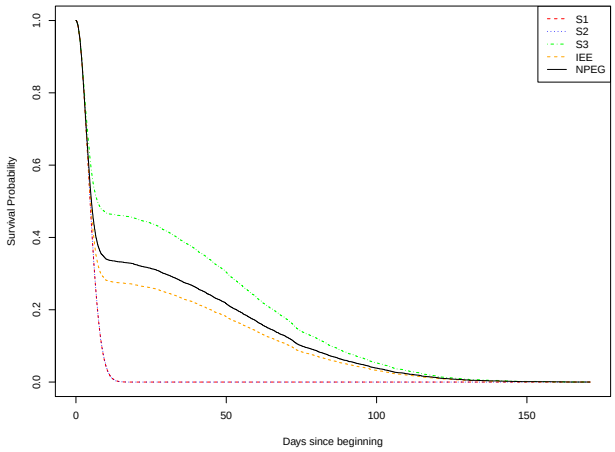


(b) Setting 2

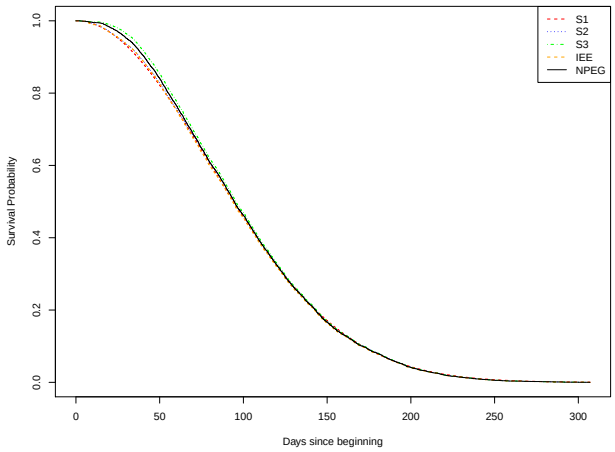


(c) Setting 3

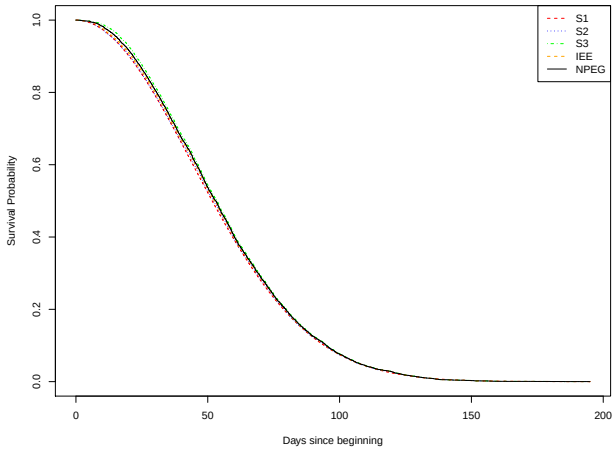
Figure 21: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 1 - 3 with all Weibull distributions, $n = 20$ and $r = 300$. S_1 is the true survival function.



(a) Setting 4

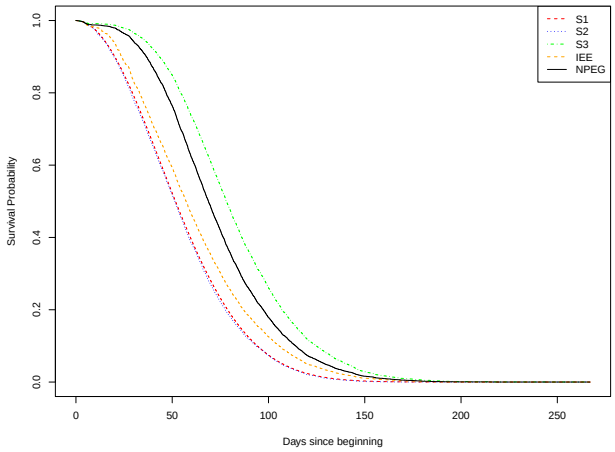


(b) Setting 5

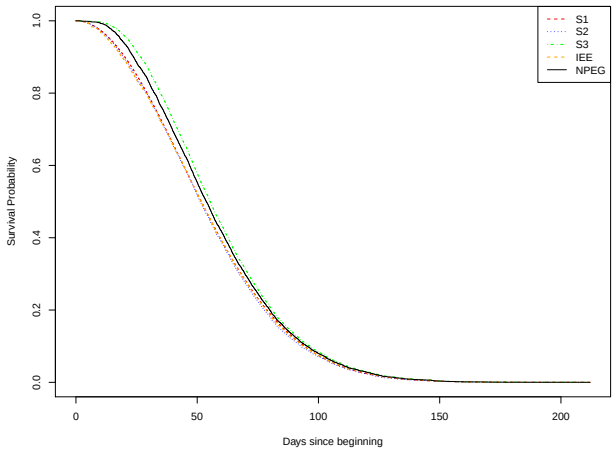


(c) Setting 6

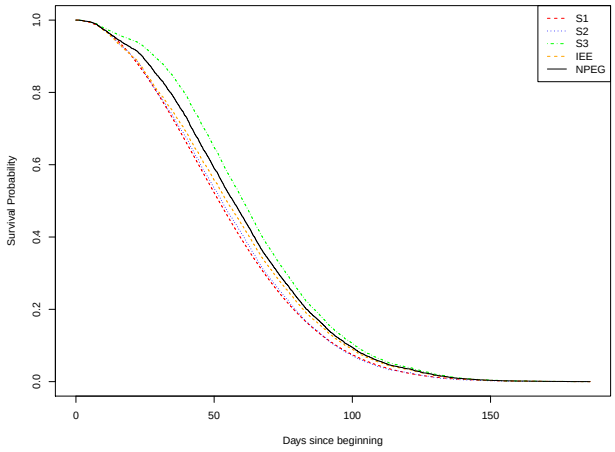
Figure 22: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 4 - 6 with all Weibull distributions, $n = 20$ and $r = 300$. S_1 is the true survival function.



(a) Setting 7

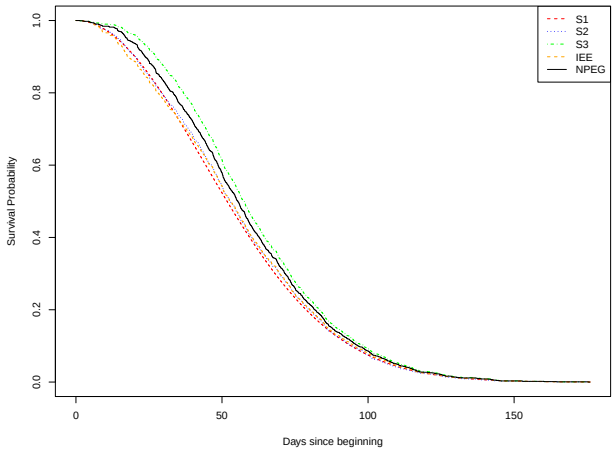


(b) Setting 8

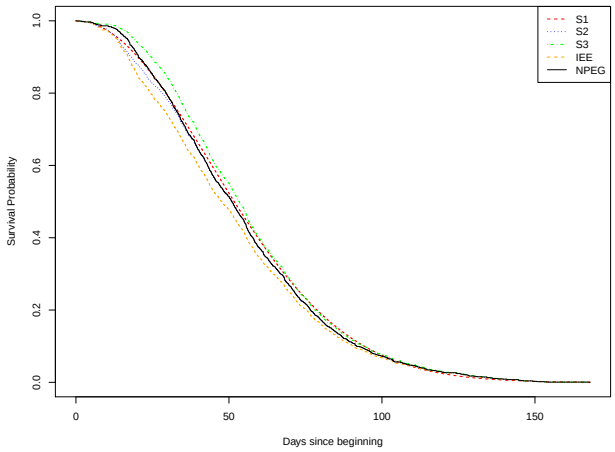


(c) Setting 9

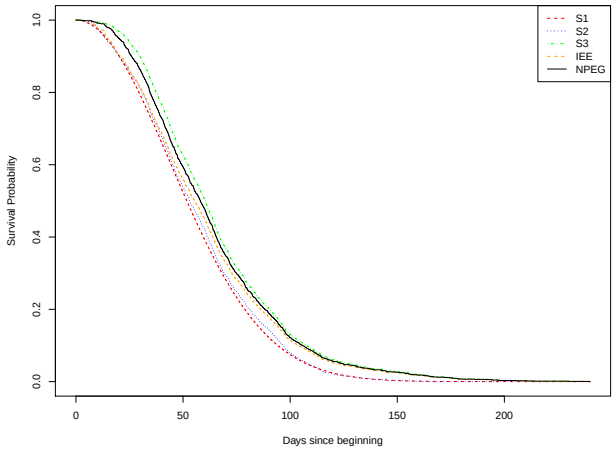
Figure 23: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 7 - 9 with all Weibull distributions, $n = 20$ and $r = 300$. S_1 is the true survival function.



(a) Setting 1

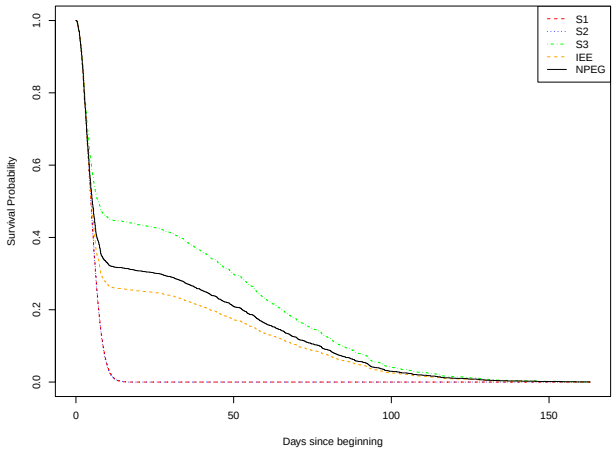


(b) Setting 2

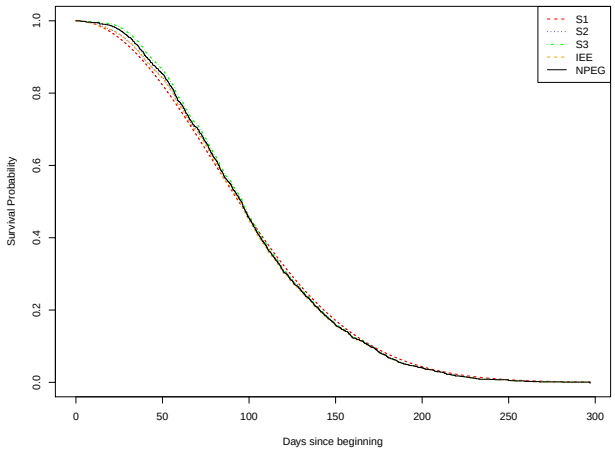


(c) Setting 3

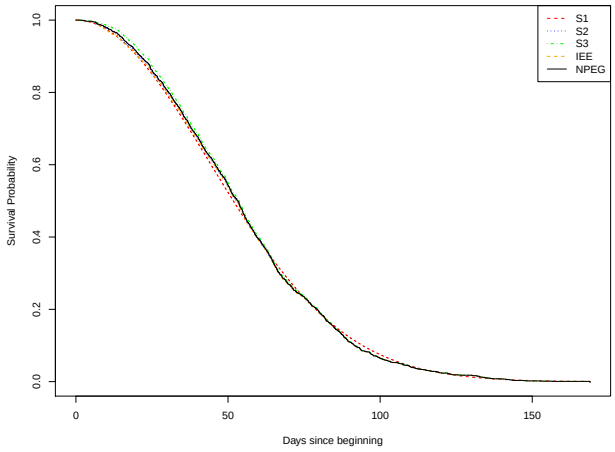
Figure 24: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 1 - 3 with all Weibull distributions, $n = 500$ and $r = 3$. S_1 is the true survival function.



(a) Setting 4

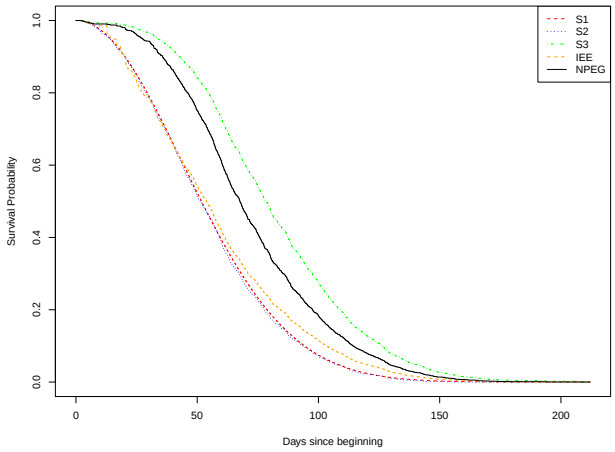


(b) Setting 5

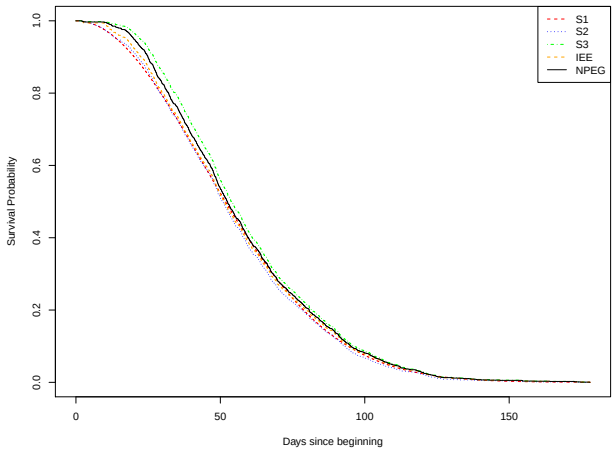


(c) Setting 6

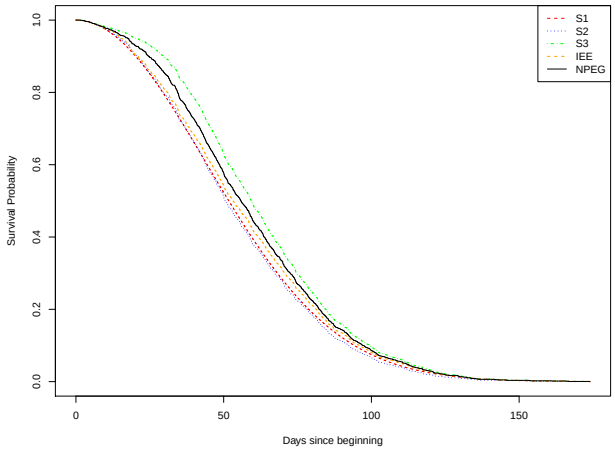
Figure 25: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 4 - 6 with all Weibull distributions, $n = 500$ and $r = 3$. S_1 is the true survival function.



(a) Setting 7



(b) Setting 8



(c) Setting 9

Figure 26: Comparison of the estimated survival functions based on the NPEG, IEE, $\hat{S}_2$ and $\hat{S}_3$ for the settings 7 - 9 with all Weibull distributions, $n = 500$ and $r = 3$. S_1 is the true survival function.

The probability P_3 informs how many subjects will not be considered, because they do not satisfy the assumption described in Section 2.2. Since it is assumed that the chance to have more than one event in a gap is ignored, P_3 is expected to be small. Most of simulation settings have quite small empirical probabilities $\hat{P}_3$. With large n, the setting having $\hat{P}_3 = 20.73\%$ does not become a problem, because 396 subjects are used on average when calculating the estimates $\hat{S}_3$, $\hat{S}_{IEE}$ and $\hat{S}_{NPEG}$. With only 20 subjects, $\hat{P}_3$ values for Settings 2 and 7 with all exponential distributions are 15.22 % and 18.58 %, respectively. Thus, only 15 subjects are used on average to calculate $\hat{S}_3$, $\hat{S}_{IEE}$ and $\hat{S}_{NPEG}$. But, in the other settings, 16 subjects are kept on average for the calculation. Even if only small number of subjects are used, $\hat{P}_3$ is not a problem when comparing the simulation results, because each simulation setting is repeated 300 times and the estimates $\hat{S}_3$, $\hat{S}_{IEE}$ and $\hat{S}_{NPEG}$ are the averages of 300 estimated values.

The IEE underestimates for small or large censoring time t, but overestimates for middle value of censoring time t, as mentioned in Yang's ([38]): the IEE and true survival curves almost merge together in the area where the sample points are very dense, because less information can be provided with the fewer sample points. The simulation results show that the NPEG works better than the IEE with fewer sample data points. The NPEG estimated curve changes slower than the IEE estimated curve for small or large censoring time t.

The simulation results from Settings 7-1-6 give smaller bias by decreasing the expected length of the gap as discussed in Section 2.3.2.2. If $\hat{P}_2$ becomes bigger, T_1 which are not W_1 should be detected more frequently. We can see the curves for $\hat{S}_{NPEG}$ and $\hat{S}_{IEE}$ are above the curves for S_1 and $\hat{S}_2$ with middle values of censoring time t. It means that the NPEG and IEE overestimate. However, the curves are below the curve for the empirical estimate $\hat{S}_3$.

To examine the consistency of the NPEG, simulation settings are conducted with a large sample of 500. Due to high computational burden, each setting is repeated only three times. If the results having $n = 500$ are compared with those having only $n = 20$, the consistencies of the NPEG and IEE are not achieved by the increments of the subject sizes because their results are not different as we discussed in Section 2.3.2.2. Although Yang mentioned the IEE is always robust, the simulation results show that its variance may not go away even if the sample size increases. As remarked in Section 2.3.2.4, the NPEG would be a consistent estimate if its bias is zero or if its bias goes to zero as the sample size increases.

In these simulation studies, two different underlying distributions of the four random variables having same expectations are used. Table 20 shows that the variances of the four variables with the exponential distributions are much larger than those with the Weibull distributions, even though the four variables have the same expectations. A smaller variance means that the corresponding variable is more stable. That is, variables generated based on the Weibull distributions are more stable than those based on the exponential distributions. As the variance of the random variable G_{length} increases, there is a less chance to have overlapping gaps. When gap overlapping happens more frequently, more common missing time periods are obtained. As more subjects have the same missing periods, more information is lost. In an extreme case that all the subjects have the same gap G, all information for the time period G is missed and thus $\hat{S}_{NPEG}$ and $\hat{S}_{IEE}$ become $\hat{S}_3$, the empirical survival function when the gaps are ignored. By comparing Settings 6-1-7, the effect of missing intervals due to the variances of the gap length can be checked. Since the settings with the exponential distributions have larger variances than those with the Weibull distributions, the estimates from the settings with the exponential distributions overestimate more than those from the settings with the Weibull distribution as the plots show.

In order to evaluate the performance of the NPEG, IEE and the ignoring-gap method, their estimated Mean Square Errors (MSE's) and biases are computed. All the settings use only $n = 500$ and the Weibull distributions for this evaluation. To obtain the estimates of the MSE and bias for several percentiles, $p = 5,\ 25,\ 50,\ 75$ and 95, every simulation setting is repeated 10 times. The estimates considered here are all discrete stepwise functions, even if n is large. Therefore, estimating p might sometimes be impossible because the stepwise function does not have its inverse value. In this case, the following approximation is used to calculate the inverse value: if there is no inverse value for a given p, find out the generated sample point x that makes the difference between p and the estimated value $\hat{S}(x)$ less than 0.01 (1%). This approximated percentile is treated as an estimated percentile p.

Tables 22 and 23 present the estimated bias and MSE of the NPEG, IEE and ignoring-gap method. See Table 20 for the underlying distributions used to generate data. The values in Tables 22 and 23 are rounded to two decimal places for simplicity. The simulation error is calculated by repeating each simulation 5 times to check whether those values can be rounded. Table 24 shows the simulation errors for each percentile. It shows that the simulation error is small enough to round the estimated values without losing accuracy. The small SE enables us to compare the MSE and bias of different estimates. Tables 22 and 23 show that the NPEG estimated values tend to have smaller Bias and MSE than the IEE and ignoring-gap method for all the simulation settings except Settings 4 and 7. It is detected in this bias study that the IEE underestimates for small or large censoring time t but overestimates for middle ranged time t as discussed earlier. Table 23 shows that the MSE values of the NPEG and IEE become larger as their corresponding biases become greater. Therefore, when the biases of the NPEG and IEE are not small, consistent estimates cannot be obtained. This comparison on the Bias and MSE values provides a strong evidence that the NPEG is reliable and robust if the proportion of the observed first

Table 22: Estimated biases of the NPEG, IEE and ignoring-gap methods for $p = 5,\ 25,\ 50,\ 75$ and 95.

	Setting 1			Setting 2			Setting 3		
p	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap
5	2.13	0.81	4.05	-0.25	-0.48	1.96	16.93	13.35	20.51
25	1.18	3.10	4.42	0.10	1.66	3.58	5.75	8.10	9.78
50	1.29	3.68	5.72	0.92	4.42	1.47	3.35	6.13	8.45
75	0.09	4.02	7.59	0.53	4.80	2.58	1.77	5.24	7.75
95	4.08	1.12	7.33	0.34	-2.43	5.43	1.39	6.33	9.26
	Setting 4			Setting 5			Setting 6		
p	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap
5	82.66	78.66	88.14	-3.30	-4.92	-4.96	-2.56	-3.32	-3.32
25	33.29	12.53	50.26	-1.81	-2.01	-3.16	-0.38	-0.87	-0.92
50	0.49	0.18	2.88	0.27	1.45	2.25	0.85	1.39	1.81
75	-0.08	-0.10	0.14	1.12	2.73	3.76	0.36	0.92	1.97
95	-0.03	-0.03	-0.01	5.89	2.42	9.12	0.46	1.38	2.82
	Setting 7			Setting 8			Setting 9		
p	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap
5	21.30	11.30	33.26	2.96	1.38	4.17	2.69	3.53	5.602
25	17.47	4.12	29.60	-0.28	1.30	2.23	2.38	3.94	6.870
50	15.64	1.69	26.41	-0.44	0.72	2.13	1.39	4.44	7.560
75	17.26	-1.49	25.15	0.42	2.56	4.48	1.29	4.76	9.229
95	13.03	0.46	19.89	2.45	5.54	7.98	2.91	1.00	5.623

events that are not actually the first true first events is moderate.

2.5 Real Data Study

This section studies the NPEG to analyze a real life data from the medical study at the Duke Medical School mentioned in Section 2.1. The original data contains 404 patients (subjects). Only 129 subjects among them satisfy the assumptions described in Section 2.2, because those patients do not contain more than one gap before the observed first event time. Although we tried to get access to the original data, we did not get the permission to publish the analysis of the actual medical data. However,

Table 23: Estimated biases of the NPEG, IEE and ignoring-gap methods for $p =$ 5, 25, 50, 75 and 95.

	Setting 1			Setting 2			Setting 3		
p	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap
5	13.29	4.70	28.47	17.68	18.00	14.75	293.52	192.30	422.49
25	3.15	11.55	21.80	0.92	2.89	14.13	40.78	73.15	105.00
50	1.96	13.82	33.49	1.46	26.14	2.23	17.86	49.79	75.62
75	0.43	17.29	58.49	4.48	46.30	7.74	6.61	28.46	61.75
95	17.22	3.57	53.95	3.24	11.46	33.91	10.33	42.40	89.04
	Setting 4			Setting 5			Setting 6		
p	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap
5	453.86	129.19	1111.25	34.37	14.91	52.46	9.72	17.07	33.28
25	312.12	20.36	877.54	3.16	3.74	7.97	7.38	17.60	49.26
50	249.13	11.72	700.38	1.95	1.91	6.73	4.67	20.22	59.36
75	301.75	13.87	633.88	0.73	7.01	21.68	2.43	22.75	85.59
95	188.37	4.71	404.52	6.73	45.12	64.00	10.04	2.69	32.20
	Setting 7			Setting 8			Setting 9		
p	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap	NPEG	IEE	ignoring gap
5	128.06	455.19	1111.25	27.46	21.81	52.46	9.72	17.07	33.280
25	23.78	308.76	877.54	3.15	3.74	7.97	7.38	17.60	49.259
50	7.33	253.53	700.38	1.95	1.90	6.72	4.66	20.22	59.360
75	5.79	309.83	633.87	0.73	7.00	21.68	2.43	22.75	85.590
95	18.80	174.28	404.52	6.73	45.12	64.00	2.56	10.16	32.198

Table 24: Simulation errors for each percentile p.

NPEG	Settings								
p	1	2	3	4	5	6	7	8	9
5	0.004	0.003	0.000	0.010	0.005	0.005	0.004	0.002	0.005
25	0.005	0.004	0.001	0.009	0.003	0.003	0.005	0.004	0.003
50	0.002	0.005	0.002	0.008	0.005	0.002	0.003	0.003	0.005
75	0.002	0.002	0.001	0.009	0.002	0.002	0.002	0.002	0.002
95	0.006	0.002	0.001	0.005	0.006	0.006	0.004	0.002	0.006
IEE	Settings								
p	1	2	3	4	5	6	7	8	9
5	0.008	0.003	0.001	0.009	0.005	0.003	0.006	0.003	0.003
25	0.001	0.004	0.001	0.008	0.005	0.003	0.001	0.004	0.004
50	0.003	0.004	0.002	0.009	0.006	0.006	0.003	0.004	0.006
75	0.004	0.002	0.004	0.007	0.003	0.005	0.004	0.002	0.005
95	0.003	0.002	0.002	0.007	0.002	0.002	0.003	0.006	0.002
ignoring-gap	Settings								
p	1	2	3	4	5	6	7	8	9
5	0.004	0.008	0.006	0.008	0.006	0.005	0.004	0.005	0.005
25	0.000	0.003	0.020	0.006	0.005	0.004	0.000	0.003	0.004
50	0.001	0.003	0.003	0.007	0.003	0.003	0.001	0.004	0.003
75	0.002	0.005	0.004	0.007	0.004	0.004	0.002	0.005	0.003
95	0.006	0.004	0.003	0.005	0.006	0.006	0.005	0.004	0.007

Table 25: First ten subjects of a real life data from Duke Medical School using hour as the unit.

i	$b_{1,i}$	$e_{1,i}$	$t_{1,i}$
1	0.825824	24.341285	37.917731
2	0.547296	5.503642	23.694427
3	0.548477	80.330046	91.634539
4	2.308849	109.117011	133.634564
5	0.523838	55.938910	58.304398
6	∞	∞	13.693960
7	2.365197	16.691791	21.470155
8	0.730642	4.498762	47.635809
9	2.361725	7.914084	21.470155
10	1.729372	10.680830	14.917731

Yang's thesis provides the first 10 subjects and thus the small data is studied in this section. Table 25 provides the beginning and ending of the gap, and the observed first event time for each patient. Only one patient has no gap, thus the observed first event time for the patient is the first true event time.

The estimated survival functions of the first true event time are drawn in Figure 27 for the NPEG and IEE. The solid black line represents the estimated survival function of the first true event time based on the NPEG, the dashed blue line represents the estimate of the survival function based on the IEE method, the dotted green line represents the traditional empirical survival function $\hat{S}_3$ which ignores all gaps, and the solid red line represents a survival function whose underlying distribution is $Weibull(\hat{\alpha}, \hat{\beta})$ with the estimated parameter values $\hat{\alpha} = 1$ and $\hat{\beta} = 53.4$ by the GLF ([14, 15]). Since the GLF estimated the parameter α as 1, the distribution of the first true event time becomes an exponential distribution with $\lambda = \hat{\beta}$. Similarly, Yang mentioned that the estimated IEE survival function seems to follow exponential distribution based on Figure 2.12 in ([38]). However, with only 10 subjects, the distribution of the first true event time does not seem to follow an exponential distribution. The estimated functions based on both the NPEG and IEE are a little bit

farther from the estimated GLF survival function.

2.6 Discussion and Summary

In this chapter, we studied an estimation method for analyzing time-to-event data called a gap data, when missing time intervals can possibly happen prior to the first observed event. We have proposed a nonparametric estimate NPEG, which estimates the survival function of the first true event time up to one gap. We studied analytical properties of the proposed estimate. We show that, the new estimate NPEG is almost unbiased under some situations discussed in Section 2.3.2.1, and that it is less biased than the existing nonparametric method IEE. The simulation studies demonstrated that the NPEG is powerful and robust in certain circumstances. Additionally, the simulation studies showed that the NPEG is superior over the traditional nonparametric method based on the observed first events when ignoring gaps.

This gap data study raises the following open topics which we consider as potential future works. First, the proposed method can be extended to the multiple gaps as the GLF method and the IEE which is extended in Appendix C. We can extend the NPEG, the IEE and the GLF with the weaker assumption that the two true events can fall into a gap. Additionally, we can also extend them to study the second, third, $\cdots$, k^{th} true event times. Finally, we can use a bayesian approach to get a unbiased estimate. Using a bayesian formulation, better knowledge on the relationship between the first true event and the second true event can be attained, and the knowledge can help obtain a unbiased estimate for the gap data.

2.7 Summary of Contributions

This thesis research is motivated by a special type of time-to-event data with missing data called the gap data which was encountered in a heart disease study conducted at the Duke University Medical Center. This type of data consists of multiple event time observations. For example, the event time for the medical data

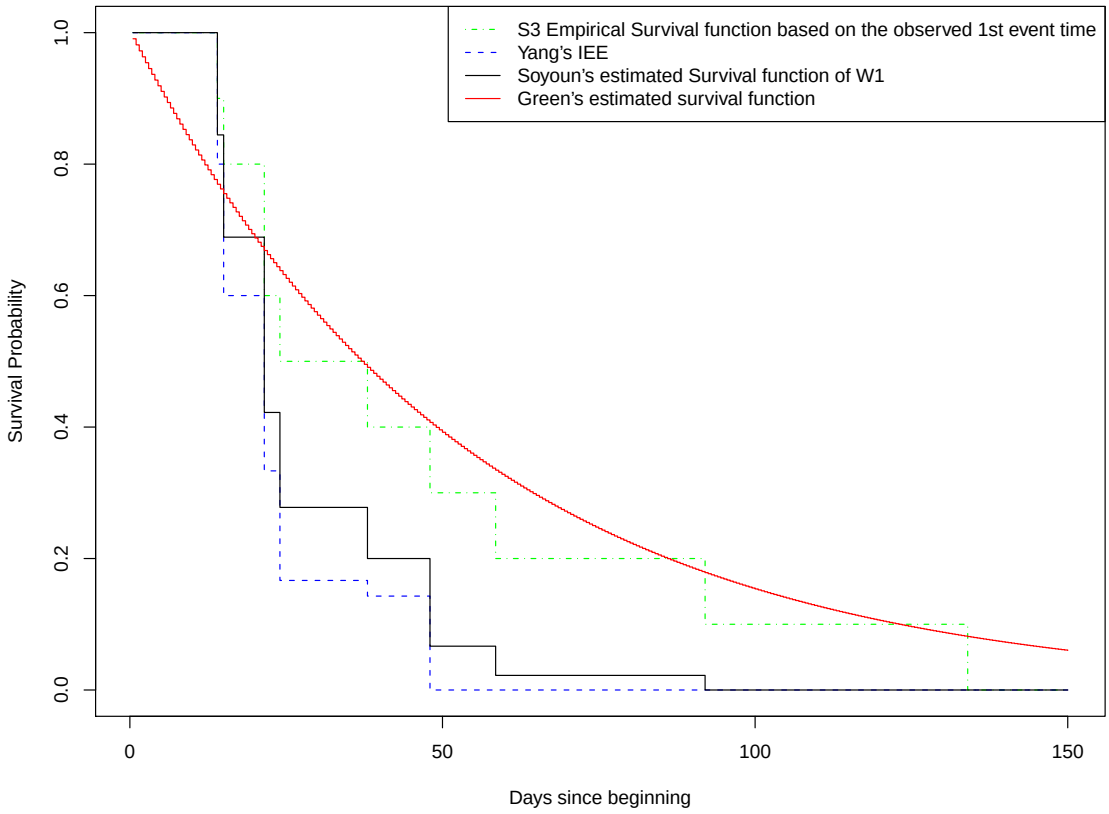


Figure 27: The estimated survival functions based on the NPEG and IEE. The dotted green line represents the traditional empirical survival function $\hat{S}_{T_1}$ based on the observed first event times when ignoring the gaps. The dashed blue line represents the estimated survival function based on the IEE method, and the solid black line represents the estimated survival function for the first true event time W_1 based on the NPEG. The solid red line represents the survival function S_1 where the distribution of W_1 is $Weibull(\hat{\alpha}, \hat{\beta})$ with the estimated parameter values $\hat{\alpha} = 1$ and $\hat{\beta} = 53.4$ by the GLF.

is the time that the ST-segment amplitude has fallen to 50% of the peak deviation. However, the observations may have one or more missing time periods called gaps before observing the first events. Therefore, the observed first event may or may not be the first true event due to the possibility that the first true event may have happened in one of the gaps. Here, we focused on such dataset up to one gap for simplicity.

As we discussed in Section 2.1, this type of data cannot be handled with methods for the right/left censored data. There are few studies done on this type of data ([14, 15, 38]). An existing parametric method, GLF ([14, 15]) estimates the parameters for the distribution of the true event time with multiple gaps by maximizing the gap likelihood function. Yang ([38]) proposed a nonparametric method called the IEE for the gap data up to one gap. The IEE has a limitation in deriving its analytical property due to its complicated formula. We introduced a new nonparametric method, NPEG, which is well defined and simple.

We have studied a new nonparametric method to estimate the survival function of the first true event time in the gap data up to one possible gap. As we illustrated with a simple example in Section 2.3.1, the calculation is simply and easy to understand. Several analytical works for the proposed estimate are finished and their properties are studied in Section 2.3.2. In the simulation studies, the proposed method is shown to be robust and powerful. We also extended the IEE to the case with multiple gaps in Appendix C.

APPENDIX A

INVESTIGATION OF THE CLUSTERING GENERATION PROCEDURE IN THE PR-SGR ALGORITHM

We conduct a simulation study to evaluate the clustering generation procedure used in the PR-SGR Algorithm described in Section 1.2.4.1. A cluster is generated as following: for a predictor X_i and a given ρ,

$$\mathcal{C}_i = \mathcal{C}_i(\rho) = \{j \ : \ Corr(X_i, X_j) =\mid X_i^T X_j \mid \geq \rho, \ \forall j \in \mathcal{A}^c\} \cup \{i\}, \text{ for } i \in \mathcal{A}.$$

The active set $\mathcal{A}$ contains all the predictors contained in the clusters that are constructed so far. Therefore, there is no information about a cluster that can contain X_k for $k \notin \mathcal{A}$.

We use the same setting as in Example 11, in Section 1.3 and repeat 100 times to evaluate the clustering generation procedure. In Example 11, a cluster $\mathcal{C}_1$ consists of three predictors X_1, X_2 and X_3 having their pairwise correlations at least 0.9, and the other cluster $\mathcal{C}_2$ consists of two predictors X_4 and X_5 having their correlation at least 0.8. Additionally, the predictors which are not in the same cluster are generated independently, and thus their pairwise correlation should be small. We use the PR-SGR with the MAX representative selection rule in this appendix.

For $\rho = 0.9$, we always obtain a cluster containing all the predictors in $\mathcal{C}_1$, whenever one of the predictors in $\mathcal{C}_1$ is selected in the model. As $\mathcal{C}_2$ consists of two predictors having at least 0.8 correlation, there is a low chance of generating a cluster that contains both X_4 and X_5 with $\rho = 0.9$. Empirically, we have only 3% chance of obtaining a cluster that contains X_4 and X_5. As ρ decreases down to 0.8, the chance of obtaining a cluster consisting of X_4 and X_5 increases. For $\rho = 0.8$, we always

have two clusters which are identical to $\mathcal{C}_1$ and $\mathcal{C}_2$ whenever their representatives are selected in the model. However, if ρ decreases, there is more chance of generating different clusters from the clusters $\mathcal{C}_1$ and $\mathcal{C}_2$. For example, for $\rho = 0.5$, we have 45 cases among 100 repetitions in which the generated clusters are different from $\mathcal{C}_1$ and $\mathcal{C}_2$ (e.g., a generated cluster contains only X_1 and X_3 and another cluster contains X_2 and some other predictors). In this thesis, we focus on the case that there are spuriously highly correlated predictors. Therefore, considering high ρ values (for example, $\rho = 0.9$) is reasonable. As we discussed in Section 1.2.3, the tuning parameter ρ in the simulation and real-life example studies is selected from relatively large values (0.99, 0.98, $\cdots$, 0.91, 0.90, 0.85, 0.80, 0.70, 0.60). All the representative selection rules result in similar conclusion in this investigation.

APPENDIX B

PR-SGR WITH INDEPENDENT PREDICTORS

A simulation study is carried out to evaluate the performance of the PR-SGR Algorithm when all the predictors are assumed to be independent and when there is no group of highly correlated predictors. We keep the same simulation setting used in Section 1.3, but generate n independent predictors.

Table 26 summarizes the prediction results in terms of median MSE values for various ρ values. The results in the first two rows are based on the ρ value selected by $10-$CV and the results in other rows are based on fixed ρ values (0.9, 0.8, 0.7, 0.6 and 0.5). When there are only independent predictors, the Elastic-Net performs poorly among all the methods. The PR-SGR with various representative selection rules for most ρ values performs better than the Lasso/Lars and Elastic-Net, and in particular, the performance of the PR-SGR is the best when the ρ value is selected by $10-$CV. Although there is no group of highly correlated predictors, the PR-SGR builds a cluster of some predictors and uses a representative of the cluster in the model. Therefore, the PR-SGR loses the information contained in the predictors that are not selected as representatives. The cluster built for $\rho = 0.5$ tends to include more predictors than for higher ρ values, and thus the PR-SGR with the MIN, RAN or CRT representative selection rules performs worse than the Lasso/Lars and the PR-SGR with other representative selection rules. The MAX and MED representative selection rules give relatively stable MSE values.

Table 26: Statistics of MSE and model complexity for the Lasso/Lars, Elastic-Net, and PR-SGR with 5 representative selection rules.

				PR-SGR				
ρ	MSE	Lasso/Lars	Elastic	MAX	MIN	MED	RAN	CRT
10−CV	Median	69.96	109.75	60.94	62.75	62.96	65.28	62.09
	s.e	13.25	12.14	9.76	9.91	9.18	9.48	9.95
0.90	Median	79.79	103.54	63.76	65.12	66.03	68.11	62.21
	s.e	12.77	11.70	10.75	9.47	9.78	9.87	10.03
0.80	Median	79.44	108.93	67.19	69.47	68.95	70.00	69.27
	s.e	11.02	11.05	9.79	8.88	9.03	8.67	9.59
0.70	Median	69.10	103.37	67.21	65.80	66.73	68.36	72.17
	s.e	11.22	10.10	8.66	8.67	8.57	9.69	9.02
0.60	Median	73.44	105.80	63.90	70.71	63.71	70.00	70.67
	s.e	11.18	12.07	9.96	11.36	9.95	9.34	10.93
0.50	Median	73.49	106.55	66.08	79.66	67.17	76.16	76.66
	s.e	16.13	11.32	9.97	10.16	9.66	11.79	13.11

APPENDIX C

EXTENSION OF THE IEE WITH MULTIPLE GAPS

Yang ([38]) proposed the IEE to estimate the first true event time W_1 when up to one gap is detected. This appendix extend the idea of the IEE with multiple gaps. To present this extension, start with the simplified IEE formulation driven in Section 2.2. The IEE calculates the survival function of the first true event time W_1 up to one gap as

$$\begin{aligned}
\hat{S}_{IEE}(t) &= 1 - \sum_{i:t_{(i)}<t} \frac{\hat{p}_{wap}(i)}{n - n_i} \\
&= 1 - \sum_{i:t_{(i)}<t} \frac{\hat{p}(i)}{m_i} \\
&= 1 - \sum_{i=1}^{n} \frac{\hat{p}(i)}{m_i} I(t_{1,(i)} \leq t),
\end{aligned}$$

where $\hat{p}(i) = 1 - \sum_{j=1}^{i-1} \frac{h_{ji}}{m_j} \hat{p}(j)$, $m_i = n - n_i = n - \sum_{j=i+1}^{n} h_{ij}$, and $h_{ij} = I(t_{1,(i)} \in G_{(j)})$. Note that $p(i)$ is the probability that the first true event time W_1 is at the observed first event time $T_{1,(i)}$. Its estimated probability is interpreted in a different way as follows:

$$\begin{aligned}
\hat{p}(i) &= \hat{Pr}\{W_1 \text{ is at the observed time point } T_{1,(i)}\} \\
&= 1 - \hat{Pr}\{W_1 \text{ is not at the observed time point } T_{1,(i)}\} \\
&= 1 - \sum_{j=1,\ j\neq i}^{n} \hat{Pr}\{W_1 \text{ is at the observed time point } T_{1,(j)} \text{ and } T_{1,(j)} \in G_{(i)}\} \\
&= 1 - \sum_{j=1,\ j\neq i}^{n} \hat{Pr}(T_{1,(j)} \in G_{(i)})\ \hat{Pr}\{W_1 \text{ is at the observed time point } T_{1,(j)}\} \\
&= 1 - \sum_{j=1,\ j\neq i}^{n} \hat{Pr}(T_{1,(j)} \in G_{(i)})\ \hat{p}(j)
\end{aligned}$$

$$
\begin{aligned}
&= 1 - \sum_{j=1}^{i-1} \hat{Pr}(T_{1,(j)} \in G_{(i)})\ \hat{p}(j) \quad (\because\ \{T_{1,(j)} \in G_{(i)}\} = \emptyset\ for\ j > i) \\
&= 1 - \sum_{j=1}^{i-1} \frac{I(T_{1,(j)} \in G_{(i)})}{n - n_j}\ \hat{p}(j) \quad \text{where } n_j = \sum_{k=j+1}^{n} h_{jk} \\
&= 1 - \sum_{j=1}^{i-1} \frac{h_{ji}}{m_j}\ \hat{p}(j) \quad \text{where } h_{ji} = I(T_{1,(j)} \in G_{(i)}) \text{ and } m_j = n - n_j.
\end{aligned}
$$

Using the same arguments, $p(i)$ can be estimated with two gaps $G_{1,i}$ and $G_{2,i}$ with a new assumption that the chance to have more than one event in each gap is negligible. The $\hat{p}(i)$ can be written as:

$$
\begin{aligned}
\hat{p}(i) &= \hat{Pr}\{W_1 \text{ is at the observed time point } T_{1,(i)}\} \\
&= 1 - \hat{Pr}\{W_1 \text{ is not at the observed time point } T_{1,(i)}\} \\
&= 1 - \sum_{j=1,\ j\neq i}^{n} \hat{Pr}\{W_1 \text{ is at the observed time point } T_{1,(j)} \\
&\qquad\qquad \text{and } T_{1(j)} \in G_{1,(i)} \cup G_{2,(i)}\} \\
&= 1 - \sum_{j=1,\ j\neq i}^{n} \hat{Pr}(T_{1,(j)} \in G_{1,(i)} \cup G_{2,(i)}) \\
&\qquad \times \hat{Pr}\{W_1 \text{ is at the observed time point } T_{1,(j)}\} \\
&= 1 - \sum_{j=1,\ j\neq i}^{n} \hat{Pr}(T_{1,(j)} \in G_{1,(i)} \cup G_{2,(i)})\ \hat{p}(j) \\
&= 1 - \sum_{j=1}^{i-1} \hat{Pr}(T_{1,(j)} \in G_{1,(i)} \cup G_{2,(i)})\ \hat{p}(j) \\
&\quad (\because\ \{T_{1,(j)} \in G_{1,(i)} \cup G_{2,(i)}\} = \emptyset\ for\ j > i) \\
&= 1 - \sum_{j=1}^{i-1} \frac{I(T_{1,(j)} \in G_{1,(i)} \cup G_{2,(i)})}{n - n_j}\ \hat{p}(j) \quad \text{where } n_j = \sum_{k=j+1}^{n} h_{jk} \\
&= 1 - \sum_{j=1}^{i-1} \frac{h_{ji}}{m_j}\ \hat{p}(j) \quad \text{where } h_{ji} = I(T_{1,(j)} \in G_{1,(i)} \cup G_{2,(i)}) \text{ and } m_j = n - n_j.
\end{aligned}
$$

Similarly, the IEE of $S_{W_1}(t)$ with m gaps $G_1,\ G_2, \cdots, G_m$ can be written as:

$$\hat{S}_{IEE}(t) = 1 - \sum_{i:t_{1(i)} \le t} \frac{\hat{p}(i)}{m_i} = 1 - \sum_{i=1}^{n} \frac{\hat{p}(i)}{m_i} I(t_{1,(i)} \le t),$$

where $\hat{p}(k) = 1 - \sum_{i=1}^{k-1} \frac{h_{ik}}{m_i}\hat{p}(i)$, $h_{ji} = I(T_{1,(j)} \in G_{1,(i)} \cup G_{2,(i)} \cup \cdots \cup G_{m,(i)}) = I\left(T_{1,(j)} \in \bigcup_{k=1}^{m} G_{k,(i)}\right)$ and $m_j = n - n_j = n - \sum_{k=j+1}^{n} h_{jk}$.

The following numerical example is displayed in Table 27, which is modified from the example used in Section 2.3.1 as an illustration to describe the IEE method with multiple gaps. First, calculate $\hat{p}(i)$ for each ordered subject i. With the first ordered subject with $t_{1,(1)} = 10$, our estimate for $p(1)$ becomes one as defined. Then, at the next ordered subject with $t_{1,(2)} = 20$, $h_{12} = 1$, $h_{14} = 1$ and $n_1 = 2$ because $t_{1,(1)} \in g_{1,(2)} \cup g_{2,(2)}$ and the second and fourth ordered gaps cover $t_{1,(1)}$. Therefore, $\hat{p}(2) = 1 - \frac{h_{12}}{m_1} = 1 - \frac{1}{8-2} = 0.8333$. Now, let's move to the third ordered subjects with $t_{1,(3)} = 31$. The gaps $g_{1,(3)} = (12, 26]$ and $g_{2,(3)} = (17, 19]$ of the third ordered subject does not cover $t_{1,(1)} = 10$ but cover $t_{1,(2)} = 20$, so $h_{13} = 0$ and $h_{23} = 1$. Since the third, fourth, fifth and eighth ordered gaps cover $t_{1,(2)}$, $n_2 = 4$, $\hat{p}(3) = 1 - \frac{h_{13}}{m_1} - \frac{h_{23}}{m_2} = 0.7917$. The fourth ordered subject with $t_{1,(4)}$ has one gap $g_{2,(4)} = (6, 21]$ which covers $t_{1,(1)}$ and $t_{1,(2)}$, so $h_{14} = h_{15} = 1$. Because $t_{1,(3)}$ falls into the gaps of the fifth, sixth, and seventh, $n_3 = 3$ and hence $\hat{p}(4) = 0.6250$. For the fifth ordered subject with $t_{1,(5)}$, $\hat{p}(5) = 0.6333$ with $n_4 = 2, h_{15} = h_{45} = 0$ and $h_{25} = h_{35} = 1$. The gap of the next ordered subject contains $t_{1,(3)}$, $t_{1,(4)}$, $t_{1,(5)}$ and $t_{1,(6)}$ and its observed first event $t_{1,(6)}$ is falling in the gap $g_{2,(7)}$. Therefore, the estimate of $p(6)$ is $\hat{p}(6) = 1 - \frac{h_{16}}{m_1} - \frac{h_{26}}{m_2} - \frac{h_{36}}{m_3} - \frac{h_{46}}{m_4} - \frac{h_{56}}{m_5} = 0.7375$. For the seventh ordered subject, $h_{17} = h_{27} = 0$, $h_{37} = h_{47} = h_{57} = h_{67} = 1$ and $n_6 = 2$. So $\hat{p}(7) = 0.5090$. Finally, because the eighth ordered subject' gap contains $t_{1,(5)}$ and $t_{1,(6)}$, and $h_{28} = h_{58} = h_{68}$ are one, $\hat{p}(8) = 0.5632$. These calculation procedures are summarized in Table 28.

Table 29 provides the IEE estimated survival function $\hat{S}_{IEE}(t)$ up to two gaps,

Table 27: Modified example from Table 16 up to two gaps.

Original data				Ordered data			
i	$g_{1,i}$ $((b_{1,i}, e_{1,i}])$	$g_{2,i}$ $((b_{2,i}, e_{2,i}])$	$t_{1,i}$	(i)	$g_{1,(i)}$ $((b_{1,(i)}, e_{1,(i)}])$	$g_{2,(i)}$ $((b_{2,(i)}, e_{2,(i)}])$	$t_{1,(i)}$
1	(21,52]	$(\infty, \infty]$	57	1	(3,5]	(8,10]	10
2	(12,26]	(28,29]	31	2	(7,16]	(17,19]	20
3	(17,25]	(40,52]	63	3	(12,26]	(28,29]	31
4	(3,5]	(8,10]	10	4	(6,21]	$(\infty, \infty]$	35
5	(29,37]	$(\infty, \infty]$	47	5	(13,24]	(30,32]	45
6	(6,21]	$(\infty, \infty]$	35	6	(29,37]	$(\infty, \infty]$	47
7	(7,16]	(17,19]	20	7	(21,52]	$(\infty, \infty]$	57
8	(13,24]	(30,32]	45	8	(17,25]	(40,52]	63

Table 28: Procedure to calculate the IEE up to two gaps.

(i)	$g_{1,(i)}$	$g_{2,(i)}$	$t_{1,(i)}$	h_{1i}	h_{2i}	h_{3i}	h_{4i}	h_{5i}	h_{6i}	h_{7i}	n_i	$\hat{p}(i)$
1	(3,5]	(8,10]	10								2	1
2	(7,16]	(17,19]	20	1							4	0.8333
3	(12,26]	(28,29]	31	0	1						3	0.7917
4	(6,21]	$(\infty, \infty]$	35	1	1	0					2	0.6250
5	(13,24]	(30,32]	45	0	1	1	0				2	0.6333
6	(29,37]	$(\infty, \infty]$	47	0	0	1	1	0			2	0.7375
7	(21,52]	$(\infty, \infty]$	57	0	0	1	1	1	1		0	0.5090
8	(17,25]	(40,52]	63	0	1	0	0	1	1	0	0	0.5632

Table 29: The estimated survival function $\hat{S}_{IEE}$ up to two gaps with the modified example in Table 27. For comparison, $\hat{S}_{IEE}$ up to one gap with the original example in Table 16 is displayed in the last column.

t	$\hat{S}_{IEE}(t)$ up to two gaps	$\hat{S}_{IEE}(t)$ up to one gap
$0 < t \leq 10$	1	1
$10 < t \leq 20$	0.8333	0.8571
$20 < t \leq 31$	0.6250	0.6857
$31 < t \leq 35$	0.4667	0.5476
$35 < t \leq 45$	0.3625	0.3810
$45 < t \leq 47$	0.2569	0.2626
$47 < t \leq 57$	0.1340	0.1633
$57 < t \leq 63$	0.0704	0.1036
$63 < t$	0	0

for a given censoring time t. To compare the IEE up to one gap, another column is added to display the estimated survival function based on the example introduced in Section 2.3.1. Figure 28 is drawn to compare the IEE methods up to two gaps (solid black line) and up to one gap (dashed red line) with the traditional empirical survival function $\hat{S}_{T_1}(t)$, denoted as S_3 (dotted blue line). As the number of gaps per subject increases, there are more chances to have other subjects falling in the gaps and more chances to underestimate for a given censoring time.

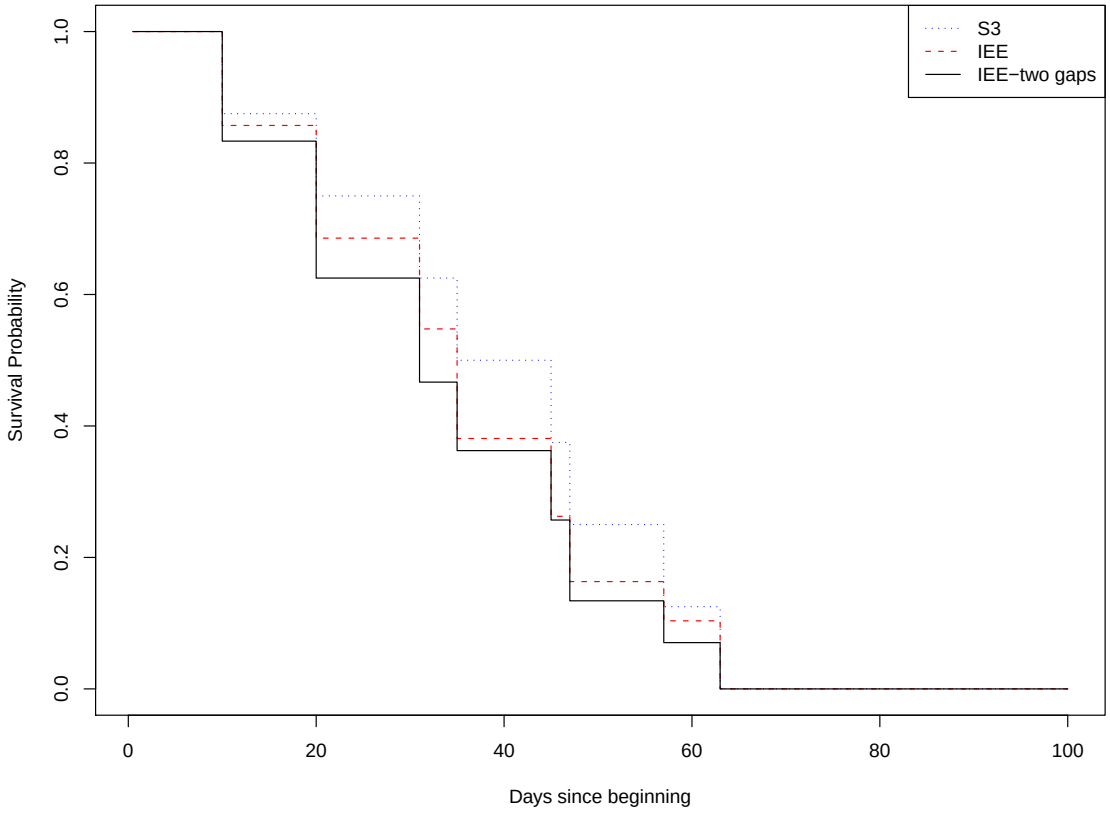


Figure 28: The estimated survival function based on the IEE method with multiple gaps in an example up to two gaps. The dotted blue line represents the traditional empirical survival function $\hat{S}_{T_1}(t)$ based on the observed first event times when ignoring the gaps. The dashed red line represents the estimated survival function based on the IEE method when considering only up to one gap, which is the case using the example introduced in Section 2.3.1. The solid black line represents the estimated survival function for the first true event time W_1 based on the IEE with multiple gaps.

APPENDIX D

EXPECTATION OF THE IEE AND ITS BIAS

It is shown that the IEE estimate is simplified as

$$\hat{S}_{IEE}(t) \;=\; 1 - \sum_{i=1}^{n} \frac{P(i)}{M_i} I(T_{1,(i)} \le t),$$

where $P(i) = 1 - \sum_{j=1}^{i-1} \frac{H_{ji}}{M_j} P(j)$, $M_i = n - \sum_{j=i+1}^{n} H_{ij}$ and $H_{ij} = I(T_{1,(i)} \in G_{(j)})$. This appendix studies its expectation and bias. First, consider the conditional expectation of $\sum_{i=1}^{n} \frac{P(i)}{M_i} I(T_{1,(i)} \le t)$ given $H_{ij} = h_{ij},\ i = 1, \cdots, n$. Since $H_{ij} = h_{ij}$s are given, $P(i)$ and M_i become known values $p(i)$ and m_i, respectively. The conditional expectation satisfies

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} \frac{\hat{p}(i)}{m_i} I(T_{1,(i)} \le t)\right] &= \sum_{i=1}^{n} \frac{\hat{p}(i)}{m_i} P(T_{1,(i)} \le t) \\
&= \sum_{i=1}^{n} \sum_{j=i}^{n} \frac{\hat{p}(i)}{m_i} {}_nC_j \left\{P(T_1 \le t)\right\}^j \left\{1 - P(T_1 \le t)\right\}^{n-j} \\
&\qquad (\because\ P(T_{(i)} \le t) = P(\text{at least } i \text{ of the } n\ T_1\text{'s are } \le t) \\
&= \sum_{i=1}^{n} \frac{\hat{p}(i)}{m_i} {}_nC_n \left\{P(T_1 \le t)\right\}^n \left\{1 - P(T_1 \le t)\right\}^0 \\
&\quad + \sum_{i=1}^{n} \sum_{j=i}^{n-1} \frac{\hat{p}(i)}{m_i} {}_nC_j \left\{P(T_1 \le t)\right\}^j \left\{1 - P(T_1 \le t)\right\}^{n-j} \\
&\le P(T_1 \le t) + \sum_{i=1}^{n} \sum_{j=i}^{n-1} \frac{\hat{p}(i)}{m_i} {}_nC_j \left\{1 - S_{T_1}(t)\right\}^j \left\{S_{T_1}(t)\right\}^{n-j}. \\
&\qquad \left(\because\ P(T_1 \le t) \le 1\ \&\ \sum_{i=1}^{n} \frac{\hat{p}(i)}{m_i} = 1\right)
\end{aligned}$$

Therefore,

$$\mathbb{E}\hat{S}_{IEE}(t) \;=\; \mathbb{E}\mathbb{E}[\hat{S}_{IEE}(t) \mid H_{ij}]$$

$$
\begin{aligned}
&\geq 1 - P(T_1 \leq t) - \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=i}^{n-1}\frac{\hat{P}(i)}{M_i}{}_nC_j\left\{1-S_{T_1}(t)\right\}^j\left\{S_{T_1}(t)\right\}^{n-j}\right] \\
&= S_{T_1}(t) - \sum_{i=1}^{n}\sum_{j=i}^{n-1}\mathbb{E}\left[\frac{\hat{P}(i)}{M_i}\right]{}_nC_j\left\{1-S_{T_1}(t)\right\}^j\left\{S_{T_1}(t)\right\}^{n-j}.
\end{aligned}
$$

The bias of $\hat{S}_{IEE}(t)$ is calculated as

$$
\begin{aligned}
\mathbb{B}ias[\hat{S}_{IEE}(t)] &= \mathbb{E}\hat{S}_{IEE}(t) - S_{W_1}(t) \\
&\geq S_{T_1}(t) - \sum_{i=1}^{n}\sum_{j=i}^{n-1}\mathbb{E}\left[\frac{\hat{P}(i)}{M_i}\right]{}_nC_j\left\{1-S_{T_1}(t)\right\}^j\left\{S_{T_1}(t)\right\}^{n-j} - S_{W_1}(t) \\
&= P(T_1 > t,\ B < W_1 \leq \min(E,t)) \\
&\quad - \sum_{i=1}^{n}\sum_{j=i}^{n-1}\mathbb{E}\left[\frac{\hat{P}(i)}{M_i}\right]{}_nC_j\left\{1-S_{T_1}(t)\right\}^j\left\{S_{T_1}(t)\right\}^{n-j}. \qquad (\because\ (8))
\end{aligned}
$$

Note that $\mathbb{B}ias[\hat{S}_{NPEG}(t)] = P(T_1 > t,\ B < W_1 \leq \min(E,t)) - \mathbb{E}_{(B,E)}[P_{T_1}(T_{1,1} > t,\ T_{1,2} \in (B, \min(E,t)])]$ is shown in Section 2.3.2.

APPENDIX E

RELATIONSHIP BETWEEN $P(T_1 > T)$ AND $P(W_1 > T)$ FROM GREEN'S APPROACH

This appendix considers a general form of the bias of a gap estimate based on the GLF ([14, 15]). Since a gap data only up to one gap has been studied in Chapter 2, Equation (6) is considered. With this relationship, the difference between the survival function of the observed first event time and the true survival function for the observed first event time can be calculated as below. Note that the survival function $P(T_1 > t)$ for a gap data is the expectation of $\mathbb{E}[I(T_1 > t \mid (B, E))]$. Therefore, consider $\mathbb{E}[I(T_1 > t \mid (b, e))]$ first:

$$\begin{aligned}
& \mathbb{E}[I(T_1 > t \mid (b,e))] \\
= \; & P(T_1 > t \mid (b,e)) \\
= \; & \int_t^\infty f_{T_1|(b,e)}(x)dx \\
= \; & \begin{cases} \int_t^\infty f_{W_1}(x)dx & , \; 0 < t < b \\ \int_t^\infty f_{W_1}(x)dx + \int_t^\infty \int_b^e f_{W_2}(x-y) f_{W_1}(y) dy dx & , \; e < t < \infty \end{cases} .
\end{aligned}$$

Therefore, $P(T > t)$ is calculated as

$$\begin{aligned}
& \mathbb{E}[P(T_1 > t \mid (B,E))] \\
= \; & \int_{(B,E)} \int_t^\infty f_{T_1|(b,e)}(x)dx f_{(B,E)}(b,e)d(b,e) \\
= \; & \begin{cases} \int_t^\infty f_{W_1}(x)dx = P(W_1 > t) & , \; 0 < t < b \\ P(W_1 > t) & \\ + \int_{(B,E)} \int_t^\infty \int_b^e f_{W_2}(x-y) f_{W_1}(y) f_{(B,E)}(b,e) dy dx d(b,e) & , \; e < t < \infty \end{cases} .
\end{aligned}$$

For $0 < t < b$, the difference becomes

$$P(T_1 > t) - P(W_1 > t) \;\; = \;\; 0.$$

This is obvious, because $T_1 = W_1$. On the other hand, the difference for $e < t < \infty$ is calculated as follows:

$$\begin{aligned}
& P(T_1 > t) - P(W_1 > t) \\
= & \int_{(B,E)} \int_t^\infty \int_b^e f_{W_2}(x-y) f_{W_1}(y) f_{(B,E)}(b,e) dy dx d(b,e) \\
= & \int_{(B,E)} \int_b^e \int_t^\infty f_{W_2}(x-y) f_{W_1}(y) f_{(B,E)}(b,e) dx dy d(b,e) \\
= & \int_{(B,E)} \int_b^e \int_{t-y}^\infty f_{W_2}(x) f_{W_1}(y) f_{(B,E)}(b,e) dx dy d(b,e).
\end{aligned}$$

Note that $\int_{t-y}^\infty f_{W_2}(x) > 0$ and the difference $P(T_1 > t) - P(W_1 > t)$ is positive, unless t is large enough to have $\int_{t-y}^\infty f_{W_2}(x) = 0$. Since the distributions of $W_1,\ W_2$ and (B, E) are unknown in practice and cannot be estimated completely, one cannot reduce the amount of the difference between the observed ones and the unobservable true ones. If all the distributions of $W_1,\ W_2$ and (B, E) are known, the difference and evaluate the bias of the parametric method, GLF, can be calculated.

Consider a case in which the distributions for $B,\ E,\ W_1$ and W_2 are given as $B \sim Exp(\theta_1),\ E = B + G_{begin},\ G_{begin} \sim Exp(\theta_2),\ W_1 \sim Exp(\alpha_1)$ and $W_2 = W_1 + W_{elapse},\ W_{elapse} \sim Exp(\alpha_2)$, similar to a set in the simulation studies. Then, the difference is

$$\begin{aligned}
& P(T_1 > t) - P(W_1 > t) \\
= & \int_{(B,E)} \int_b^e \int_{t-y}^\infty f_{W_2}(x) dx \; f_{W_1}(y) dy \; f_{(B,E)}(b,e) d(b,e) \\
= & \int_{(B,E)} \int_b^e \int_{t-y}^\infty \int_0^x f_{W_1}(x-z) f_{W_{elapse}}(z) dz \; dx f_{W_1}(y) dy \; f_{(B,E)}(b,e) d(b,e) \\
= & \int_0^\infty \int_b^\infty \left\{ \int_b^e \int_{t-y}^\infty \int_0^x f_{W_1}(x-z) f_{W_{elapse}}(z) dz \; dx f_{W_1}(y) dy \right\} f_E(e) de \; f_B(b) db \\
= & \int_0^\infty \int_b^\infty \left\{ \int_b^e \int_{t-y}^\infty \int_0^x f_{W_1}(x-z) f_{W_{elapse}}(z) dz \; dx f_{W_1}(y) dy \right\} f_E(e) de \; f_B(b) db,
\end{aligned}$$

where $f_E(e) = \int_0^e f_B(e-z) f_{G_{begin}}(z) dz$.

This integral should be calculated for each of the 4 possible cases: (1) $\alpha_1 = \alpha_2 = \alpha$ and $\theta_1 = \theta_2 = \theta$, (2) $\alpha_1 = \alpha_2 = \alpha$ and $\theta_1 \neq \theta_2$, (3) $\alpha_1 \neq \alpha_2$ and $\theta_1 = \theta_2 = \theta$, and (4) $\alpha_1 \neq \alpha_2$ and $\theta_1 \neq \theta_2$. The details are omitted here. For the parametric approach with a gap data, this difference can be large if the underlying distributions are misspecified.

REFERENCES

[1] Bair, E., Hastie, T., Paul, D., and Tibshirani, R., "Prediction by supervised principal components," *Journal of American Statistical Association*, vol. 101, pp. 119–137, 2006.

[2] Bair, E. and Tibshirani, R., "Semi-supervised methods to predict patient survival from gene expression data," *PLOS Biology*, vol. 2, pp. 511–522, 2004.

[3] Bertsekas, D. P., *Nonlinear Programming*. Belmont, Massachusetts: Athena Scientific, 2nd ed., 1999.

[4] Bertsimas, D. and Tsitsiklis, J. N., *Introduction to Linear Optimization*. Belmont, Massachusetts: Athena Scientific, 1997.

[5] Bondell, H. D. and Reich, B. J., "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar," *Biometrics*, vol. *to appear*, 2007.

[6] Cantor, E. J. and Stevens, E., "QEEG correlates of auditory-visual entrainment treatment efficacy of refractory depression," tech. rep., Psychological Sciences Institute, Duluth, GA, 2007.

[7] Cox, D. R., "Partial likelihood," *Biometrika*, vol. 62, pp. 269–276, 1975.

[8] Efron, B., "The efficiency of cox's likelihood function for censored data," *Journal of American Statistical Association*, vol. 72, pp. 555–565, 1977.

[9] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[10] Fan, J. and Li, R., "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360, 2001.

[11] Fan, J. and Li, R., "Statistical challenges with high dimensionality: feature selection in knowledge discovery," *Proceedings of the International Congress of Mathematicians*, vol. 3, pp. 595–622, 2006.

[12] Fan, J. and Lv, J., "Sure independence screening for ultra-high dimensional feature space," *Journal of the Royal Statistical Society*, vol. 70, pp. 849–911, 2008.

[13] Gentleman, R. and Geyer, C. J., "Maximum likelihood for interval censored data: Consistency and computation," *Biometrika*, vol. 81, pp. 618–623, 1994.

[14] Green, C., *Ph.D. Dissertation*. (Advisor: Jye-Chyi Lu) Department of Statistics, North Carolina State University: Raleigh, NC, 1999.

[15] Green, C. L., Brownie, G., Boos, D. D., Lu, J.-C., and Krucoff, M. W., "Maximum likelihood estimation of time to first event when data gaps and multiple events are possible," tech. rep., Duke Medical School, Duke University, Raleigh, NC, 2000.

[16] Groeneboom, P., "Asymptotics for interval censored observations," tech. rep., Department of Mathematics, University of Amsterdam, Amsterdam, Netherlands, 1987.

[17] Groeneboom, P., "Nonparametric maximum likelihood estimators for interval censoring and deconvolution," tech. rep., Department of Statistics, Stanford University, Palo Alto, California, 1991.

[18] Groeneboom, P. and Wellner, J. A., *Information Bounds and Nonparametric Maximum Likelihood Estimation*. New York: Birkhäuser, 1992.

[19] Hall, P., Marron, J. S., and Neeman, A., "Geometric representation of high dimension, low sample size data," *Journal of the Royal Statistical Society*, vol. 67, pp. 427–444, 2005.

[20] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. New York, New York: Springer, 2001.

[21] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. New York, New York: Springer, 2nd ed., 2008.

[22] Hoerl, A. E. and Kennard, R., "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[23] Jolliffe, I. T., *Principal Component Analysis*. New York, New York: Springer, 2nd ed., 2002.

[24] Kalbfleisch, J. D. and MacKay, R. J., "Censoring and the immutable likelihood," tech. rep., Department of Statistics, University of Waterloo, Ontario, Canada, 1978.

[25] Kalbfleisch, J. D. and Prentice, R. L., *The Statistical Analysis of Failure Time Data*. New York: Wiley, 1980.

[26] Kaplan, E. L. and Meier, P., "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, pp. 457–481, 1958.

[27] Lehmann, E. L., *Theory of Point Estimation*. New York: John Wiley & Sons, 1983.

[28] Madigan, D. and Ridgeway, G., "Discussion of least angle regression," *Annals of Statistics*, vol. 32, pp. 465–469, 2004.

[29] Nocedal, J. and Wright, S. J., *Numerical Optimization*. New York, New York: Springer-Verlag, 1999.

[30] Nocedal, J. and Wright, S. J., *Numerical Optimization*. New York, New York: Springer-Verlag, 2nd ed., 2006.

[31] Peto, R., "Experimental survival curves for interval-censored data," *Applied Statistics*, vol. 22, pp. 86–91, 1973.

[32] Rosenblatt, M., "Remark on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, vol. 27, pp. 832–837, 1956.

[33] Segal, M., Dahlquist, K., and Conklin, B., "Regression approach for microarray data analysis," *Journal of Computational Biology*, vol. 10, pp. 961–980, 2003.

[34] Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[35] Tibshirani, R., Rosset, S., Zhu, J., and Knight, K., "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society*, vol. 67, pp. 94–108, 2005.

[36] Turnbull, B. W., "The empirical distribution function with arbitrarily grouped, censored data," *Journal of the Royal Statistical Society*, vol. 38, pp. 290–295, 1976.

[37] Wu, C. F. J., "Jackknife, bootstrap and other resampling methods in regression analysis," *Annals of Statistics*, vol. 14, pp. 1261–1350, 1986.

[38] Yang, L., *Statistical Inference for Gap Data*. (Advisor: Jye-Chyi Lu) Department of Statistics, North Carolina State University: Raleigh, NC, 2000.

[39] Yuan, M. and Lin, Y., "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society*, vol. 68, no. 1, pp. 49–67, 2006.

[40] Zou, H., "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[41] Zou, H. and Hastie, T., "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, vol. 67, no. 2, pp. 301–320, 2005.