University of Louisville
ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2016

Some contributions to nonparametric and semiparametric inference for clustered and multistate data.

Sandipan Dutta
University of Louisville

Follow this and additional works at: https://ir.library.louisville.edu/etd

Part of the Biostatistics Commons

Recommended Citation
Dutta, Sandipan, "Some contributions to nonparametric and semiparametric inference for clustered and multistate data." (2016). *Electronic Theses and Dissertations.* Paper 2454.
https://doi.org/10.18297/etd/2454

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

SOME CONTRIBUTIONS TO NONPARAMETRIC AND SEMIPARAMETRIC INFERENCE FOR CLUSTERED AND MULTISTATE DATA

By

Sandipan Dutta
B.Sc, University of Calcutta, 2010
M.Sc, Indian Institute of Technology Kanpur, 2012

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in
Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

May 2016

SOME CONTRIBUTIONS TO NONPARAMETRIC AND SEMIPARAMETRIC INFERENCE FOR CLUSTERED AND MULTISTATE DATA

By

Sandipan Dutta
M.Sc, Indian Institute of Technology Kanpur, 2012
B.Sc, University of Calcutta, 2010

A Dissertation Approved on

April 15, 2016

by the following Dissertation Committee:

____
Somnath Datta, PhD
Dissertation Director

____
Susmita Datta, PhD

____
K.B Kulasekera, PhD

____
Riten Mitra, PhD

____
Ryan Gill, PhD

DEDICATION

This dissertation is dedicated to

my wife

Sinjini

without whose constant support and unconditional love this achievement could not have been possible,

and

my parents

Mr. Mihir Kanti Dutta and Mrs. Khuku Dutta

for always staying by my side.

ACKNOWLEDGEMENTS

I want to sincerely thank Dr. Somnath Datta, my dissertation director for all his tireless efforts in helping me develop this dissertation. I would also like to thank my other dissertation committee members Dr. Susmita Datta, Dr. K.B Kulasekera, Dr. Riten Mitra, and Dr. Ryan Gill for all the kind support and direction they have provided.

I thank all the faculty, staff, and students of the Department of Bioinformatics and Biostatistics who have helped me reach this point.

Finally, I take this opportunity to thank Mrs. Kanika Paul and Dr. Rajib Paul for their immense encouragement and invaluable advices that have helped me in moving forward in my academic life.

ABSTRACT

SOME CONTRIBUTIONS TO NONPARAMETRIC AND SEMIPARAMETRIC INFERENCE FOR CLUSTERED AND MULTISTATE DATA

Sandipan Dutta

April 15, 2016

This dissertation is composed of research projects that involve methods which can be broadly classified as either nonparametric or semiparametric. Chapter 1 provides an introduction of the problems addressed in these projects, a brief review of the related works that have done so far, and an outline of the methods developed in this dissertation. Chapter 2 describes in details the first project which aims at developing a rank-sum test for clustered data where an outcome from group in a cluster is associated with the number of observations belonging to that group in that cluster. Chapter 3 proposes the use of pseudo-value regression (Andersen, Klein, and Rosthøj, 2003) in combination with penalized and latent factor regression techniques for prediction of future state occupation in a multistate model based on high dimensional baseline covariates. Chapter 4 describes the development of an R package involving various rank based tests for clustered data which are useful in situations where the number of outcomes in a cluster or in a particular group within a cluster is informative. Chapter 5 explains the fouth project which aims at developing a covariate-adjusted rank-sum test for clustered data through alingned rank transformation.

TABLE OF CONTENTS

PAGE

CHAPTER

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

1.1. **Rank-sum Test for Clustered Dat**a

Rank based tests are very popular nonparametric methods for comparing two groups or populations. They are particularly useful when the underlying distributions are suspected to be non-normal. One such widely used test for comparing two groups is the Wilcoxon rank-sum test (Wilcoxon, 1945). One important assumption for applicability of Wilcoxon rank-sum test is that all the observations under the study are independent. However, this assumption may be violated under certain circumstances. In many practical situations we have clustered data where the observations within the clusters are correlated. An example of such clustered data is the data on attachment loss measurement of different teeth of the same individual. Wilcoxon rank-sum test may not be a good option for this type of clustered data. Rosner, Glynn, and Lee (2003) proposed a rank sum test for clustered data for the cases where all the cluster members are from the same group and the correlation structure within a cluster is common across groups. But this approach would not work when the members from a single cluster do not necessarily belong to the same group. Also, this will not maintain the nominal size when the number of observations in a cluster (cluster size) is associated with the outcome of interest from that cluster in some way. This is a case of informative cluster size, where the informativeness comes from the fact that the number of observations in a given cluster (i.e., the cluster size) may be affected by some latent (cluster-specific) factor that affects the outcome variable in that cluster as well. Datta and Satten (2005) proposed a rank-sum test for clustered data that does not make any assumption on the nature of clustering and performs reasonably well in case of informative cluster sizes. But, even the test by Datta and Satten (2005) does not seem to perform well in situations where the outcome of interest belonging to a group in a given cluster appears to be correlated with the number of observations having the same group membership (i.e., the intra-cluster group size) within that cluster. This scenario, following the idea of informative cluster size, can be thought of as informative intra-cluster group (ICG) sizes. This notion of informative ICG sizes can occur in a dental study when one is interested in comparing the nature of attachment losses of teeth between the upper and lower jaws. This is because, the difference (if any) between the nature of attachment loss of the teeth of upper and lower jaws can be suspected to be associated with the difference between the number of teeth present in the upper and lower jaws. Another interesting situation, where informative ICG sizes come into play, can be found in studies relating to hereditary diseases. In many genetic studies it has been observed that an inherited disease is often diagnosed at an younger age in a later generation than that in an earlier generation. This phenomenon of an earlier onset of a disease in each successive generation of a family, called anticipation, is prevalent in diseases like non-Hodgkins lymphoma, breast and ovarian cancer, Huntington's disease among others. In case of testing for this anticipation phenomenon of a disease, or in general, to test whether the age at onset of a disease differs in two different generations of large pedigrees, an interesting information might be the number of affected individuals, belonging to a certain age interval, that are present in each of the two generations under study. “Affected” individuals include subjects who are currently diseased at the time of the study as well as those who were known to be diseased at some point of time before the study. If we find that there is a large difference in the number of affected individuals (belonging to that certain age group) between the two generations, then one may relate this difference to be associated to the difference in the onset age between the two generations. So, this might be a case of informativeness in the number of subjects (affected individuals in a certain age interval) in a group (generation) within a cluster (a large pedigree). Motivated by these, we develop a rank-sum test for comparing the marginal distribution of outcomes from different groups under the cases of informative ICG sizes.

Extending the idea of within-cluster resampling (Hoffman, Sen, and Weinberg, 2001; Williamson, Datta, and Satten, 2003; Datta and Satten, 2005; Datta and Satten, 2008), we obtain a rank-sum test for clustered data, with observations from both groups being present in every cluster. Our resampling scheme is an extension of the usual within-cluster resampling because instead of resampling one observation at random from each cluster, we first resample one group membership (out of the two possible groups) for a cluster and then resample an outcome from that group belonging to that cluster. We repeat this resampling for each cluster and obtain a rank-sum statistic based on the resampled observations. Then, following the approaches of Datta and Satten (2005), we derive our test statistic by averaging the rank sum statistic over all possible choices of the resampled observations given the data. After constructing our test, we compare it with three other existing tests, including the test by Datta and Satten (2005), under naturally occurring simulation scenarios of informative ICG sizes. We show that our test maintains the correct size under the null hypothesis of marginal symmetry, unlike the test by Datta and Satten (2005). Moreover, our test has better power performances that the three other tests under this simulation study. Besides, we show that our test also has acceptable size and power in simulation settings where we have informative cluster sizes but non-informative ICG sizes and also in simulation scenarios having both non-informative cluster and ICG sizes. Additionally, we extend our test statistic for two group comparison to the cases when some of the clusters may have observations from only one of the two groups (i.e., the intra-cluster group structures are incomplete). We present a simulation study to show that our test still maintains the appropriate size and has reasonable power under this scenario of incomplete ICG structures within a cluster. We also discuss an extension to our test where there are observations from more than two groups in every cluster.

1.2 **Prediction of Future State Occupation in a Multistate Model Based on High Dimensional Baseline Covariates**

Multistate models are typically used to describe the progression of a set of subjects through a succession of stages until they reach a certain endpoint (absorbing state). A simple example of such a model is the setting of a survival analysis where there are only two states, viz., the initial state (Alive) and the final state (Dead). In general, the disease process in human can be represented through a multistate model where the different states in the model represent the different stages of the disease. In disease studies, like cancer, prognosis of patients is of much importance. This includes predicting how complicated the stage of the disease will be for a patient after t (say) months from the point of study, or, whether a patient can really survive till t months after a follow-up study. This requires estimation of state occupation probability, which is the probability that an individual would be occupying a particular stage of the disease process at a given time. As a special case, for survival (two state) models the survival probability at a given time can be interpreted as one of the two state occupation probabilities. Estimation of these state occupation probabilities become difficult in the presence of censored data. In a later section, we give an overview of how to estimate the state occupation probability at given time in presence of independent censoring using the Aalen-Johansen estimator (Aalen and Johansen, 1978). However, often we have some additional information on the patients during a disease study. As for example, various -omics data can be collected from the cancer tissues of the patients, and one has to assimilate these additional (covariate) information for better prognosis of the disease pattern in a given individual. This can be done through regression modeling of state occupation probabilities at a given time incorporating the covariate information of the subjects under study and using the resultant model for prediction purposes. Andersen, Klein, and Rosthoj (2003) invented a simple yet effective technique for directly modeling state occupation probability in a multistate process based on a given set of covariates. They proposed the overall marginal estimation of a state occupation probability using Aalen-Johansen estimator, and then using the ‘leave-one-out’ jackknife based ‘pseudo-values’ (Miller, 1974) of the marginal estimate as responses in regression modeling based on covariates. The idea behind this is that the pseudo-value (PV), corresponding to an individual, can be thought of containing information on how the covariates of that individual affect the overall marginal estimator. Under suitable regularity conditions, one can expect that the pseudo-values computed from an asymptotically linear and unbiased estimator will be approximately i.i.d with the same conditional expectation (regression function) that we are trying to estimate (Graw, Gerds, and Schumacher, 2009). Even if censoring is present in the data, the pseudo-values of the censored and the uncensored subjects are calculated in the same way. The usefulness of this pseudo-value based regression is largely due to the fact that the pseudo-values have the correct conditional expectation given the covariates. The pseudo-value based regression technique has since then been applied to other time to event data problems; see, Andersen, Hansen and Klein (2004), Klein and Andersen (2005), Klein, Logan, Harhoff, and Andersen (2007), and Andersen and Klein (2007), among others. Although originally developed for testing the effects of covariates in censored data settings, the pseudo-value based regression technique can also used for prediction of future state occupation. However, most of the existing works based on the pseudo-value technique have been carried out under the generalized linear regression framework. But, in practice, when one faces the task of predicting the state occupation of a patient at a given time based on his/her gene or protein expression profiles, the standard linear or generalized linear models will not be applicable as the covariate dimension (e.g., number of genes in microarrays or next generation sequencing arrays, number of proteins in protein arrays, or mass over charge ratios in mass spectrometry based proteomic profiles) is typically large compared to the number of individuals (sample size) under study. A recent work involving the pseudo-value technique in high dimensional settings was pursued by Mogensen and Gerds (2013) through the random forest approach, but it was limited only to competing risk models. In this article, we try to directly estimate the probability that an individual would be in a certain state of a general multistate (disease) process at a given time based on his or her covariate (gene expression) profile using the pseudo-value based regression approach in combination with a latent factor or a penalized regression technique. We explore the predictive performances of latent factor regressions such as PLS (Wold, 1966; Frank and Friedman, 1993), as well as, penalized regressions such as LASSO (Tibshirani, 1996), all using the pseudo-value approach in cases where the covariate dimension exceeds the sample size. Through extensive simulation settings we find that in majority of the settings, a properly tuned PLS model based on PV yields the best result in terms of predicting covariate-specific future state occupation.

1.3 **R Package for Rank Based Tests in Clustered Data with Informative Cluster Size and Informative Intra-Cluster Group Size**

Clustered data are often encountered in biomedical studies where the whole set of observational units can be classified into distinct "clusters" such that the units within a cluster are correlated while the units between different clusters can be assumed to be independent. Often, the goal of a study involving clustered data is to compare the outcomes from two different groups (e.g., before treatment vs. after treatment, or, presence vs. absence of a factor). In case the responses are non-normal, rank based nonparametric testing procedures are popular for such comparisons. However, the widely used Wilcoxon rank-sum and Wilcoxon signed-rank tests are applicable only if all the underlying observations independent. Such an assumption is not valid for clustered data. Rosner *et al*. (2003) proposed a rank-sum test for clustered data for comparing outcomes from two groups under the assumption that the observations from the same cluster necessarily belong to the same group. Rosner *et al*. (2006) also proposed a signed-rank test for paired comparison in clustered data under the assumption of a common intracluster correlation. In a clustered data, sometimes the cluster size, i.e. the number of units in a cluster, become a quantity of importance. This happens if the outcome from a typical cluster appears to be associated to the number of units in that cluster. For instance, in dental studies, the number of teeth present in an individual may be indicative of the individual's oral health status. Hence, measurement of tooth attachment loss (outcome) in such an individual (cluster) can be assumed to be correlated with its tooth count (cluster size). This scenario is called a case of informative cluster size. Datta and Satten (2005) proposed a rank-sum test for clustered data that performs well in case of informative cluster size. This idea was extended by Datta and Satten (2008) in developing a signed-rank test for clustered data that handles informative cluster size. During comparison of outcomes from two groups in a clustered data, there can be instances where the outcome from a group in typical cluster depends on the number of observation from that group in that cluster. For example, a wide difference in the number of upper and lower teeth of an individual may reflect some potential difference between the teeth decay pattern of his upper and lower jaws. This scenario, where an outcome is associated directly with its intra-cluster group size instead of the cluster size, is called an informative intra-cluster group size scenario. Dutta and Datta (2015) proposed a rank-sum test for clustered data that addresses this case of informative intra-cluster group (ICG) size scenario. Despite the fact that there are a lot of practical situations, like dental studies, where clustered data with informative cluster or ICG size are encountered, there does not exist any readily available software package that can carry out rank based tests for clustered data addressing the special features of informative cluster size or ICG size. Motivated by this need, we develop an R software package that implements the testing procedures developed in Datta and Satten (2005), Datta and Satten (2008), and Dutta and Datta (2015), so that researchers can readily use such tests in clustered data taking into account the potential informativeness in cluster size and intra-cluster group size.

1.4 **AN ALIGNED RANK-SUM TEST FOR CLUSTERED DATA WHEN THE INTRA-CLUSTER GROUP SIZE IS INFORMATIVE**

Rank based tests are popular for comparing distributions of outcomes from two groups when the underlying distributions are non-normal. However, the Wilcoxon rank-sum test, arguably the most popular rank based test, is not valid if the underlying observations are not independent and identically distributed (i.i.d). One such case of violation of i.i.d setup is a clustered data where the observations within a cluster are correlated. There have been a number of attempts in the past to develop rank-sum tests for clustered data. These include tests developed by Rosner, Glynn, and Lee (2003), Datta and Satten (2005), Rosner, Glynn, and Lee (2006), among others. Among these the rank-sum test developed by Datta and Satten (2005) addresses the issue of informative cluster size where the outcome from a cluster is associated with the number of observations (cluster size) in that cluster. In case of comparison of the outcomes from two groups in a clustered data, there may be a situation where the outcome from a particular group in a given cluster turns out to be associated with the number of observations belonging to that group in that cluster. Such a scenario is termed as an informative intra-cluster group (ICG) size within a clustered data. This is common in dental studies when one tries to compare the decay pattern in the upper and lower sets of teeth in individuals. Here the subjects under study form clusters and teeth present in a subject form units within a cluster. Then the decay pattern in the upper (lower) set of teeth may be reflected through the number of upper (lower) teeth present in that individual. Recently, Dutta and Datta (2015) has developed a rank-sum test for clustered data that addresses the issue of informative ICG sizes. Their test also seem to perform well in case the ICG sizes are non-informative but the cluster sizes are informative, and even in case when neither of the two is informative. However, in most studies one has much more auxiliary information than just the outcome values and grouping information. In such cases it may happen that these auxiliary covariates, often known as 'confounders', affect the distribution of the outcome in such a way that ignoring their effects may lead to a biased inference. As for example, suppose we are interested in comparing the dental attachment loss in two different sites of a tooth. In that case the it may be possible that smoking status of the subject plays an important role in this study, as the attachment loss pattern in the two sites may be different between smokers and non-smokers. In that case leaving out the information on the subject's smoking status may lead to incorrect conclusion. This calls for a method that can adjust for the confounder (auxiliary covariate) effect and carry out a test on the covariate-adjusted outcomes. In this work we address this issue by developing an 'aligned rank test' approach (Sen, 1968; Hájek, Šidák, and Sen, 1999, Section 10.1.2) for a rank-sum test in a clustered data with informative ICG sizes. In this approach we first estimate the confounder effects through rank based estimating functions that are appropriate in a clustered data with informative ICG sizes. Then, we obtain the 'aligned residuals' (confounder-adjusted outcomes) by plugging in the estimates, and finally, carry out a rank-sum test for clustered data with informative ICG sizes based on these aligned residuals. Through extensive simulation studies involving clustered data with informative ICG sizes we show that our method has two fold advantages: (i) it accurately estimates the regression effects of the confounders and their interactions in the informative ICG size setting, and (ii) rank-sum test based on the aligned residuals, obtained by plugging in the estimates, has the correct size and high power performance in clustered data with informative ICG size.

CHAPTER 2

A RANK-SUM TEST FOR CLUSTERED DATA WHEN THE NUMBER OF SUBJECTS IN A GROUP WITHIN A CLUSTER IS INFORMATIVE

2.1. Notations, Formulation of the Problem and Proposed Method

Let M denote the number of clusters and let X_{ik} denote the k^{th} observation in the i^{th} cluster, $1 \leq k \leq N_i,\ 1 \leq i \leq M,$ where N_i denotes the number of observations in the i^{th}cluster. Let G_{ik} be the indicator denoting the binary group membership (0 or 1) of the k^{th} observation in the i^{th} cluster. Thus the entire data set consists of $\{\mathbb{V}_i : 1 \leq i \leq M\}$, with $\mathbb{V}_i = \left\{N_i,\, X_{ik},\, G_{ik},\, 1 \leq k \leq N_i\right\}$ corresponding to the i^{th} cluster. Also, let N_{i1} and N_{i0} be the numbers of observations in the i^{th} cluster belonging to group 1 and group 0, respectively. Thus, we have $N_{i1} + N_{i0} = N_i$. We consider the possibility that the cluster size N_i as well as the group memberships G_{ik} are random (and thus, so are the $N_{id},\ d = 0,\ 1$). The members in a cluster could have an arbitrary dependence structure; however, members in different clusters are statistically independent and hence the entire $\mathbb{V}_i$ and $\mathbb{V}_{i'}$ are independent. For mathematical convenience, we further assume that $\mathbb{V}_i$, $1 \leq i \leq M,$ are independent and identically distributed (iid).

The null hypothesis we consider is that the observations from the two groups follow the same marginal distribution. Mathematically, it is written as

$$H_0 :\ P\left(X_{ik} \leq x \,|\, G_{ik} = 0\right) = P\left(X_{ik} \leq x \,|\, G_{ik} = 1\right)\ \left(= \mathcal{F}(x), \text{ say}\right), \text{ for all } x.$$

A penultimate version of this work can be found in Dutta and Datta (2015)

However, the empirical analogue of the above “group specific” (e.g., conditional) marginal distributions can be constructed in three possible ways resulting in three different statistical comparisons:

$$\text{(i) } \widehat{\mathcal{F}}_1(x|d) = \frac{\sum_{i=1}^{M}\sum_{k=1}^{N_i} I(X_{ik} \leq x,\, G_{ik} = d)}{\sum_{i=1}^{M}\sum_{k=1}^{N_i} I(G_{ik} = d)},\ d = 0,\ 1;$$

$$\text{(ii) } \widehat{\mathcal{F}}_2(x|d) = \frac{\sum_{i=1}^{M}\frac{1}{N_i}\sum_{k=1}^{N_i} I(X_{ik} \leq x,\, G_{ik} = d)}{\sum_{i=1}^{M}\frac{1}{N_i}\sum_{k=1}^{N_i} I(G_{ik} = d)},\ d = 0,\ 1;$$

$$\text{(iii) } \widehat{\mathcal{F}}_3(x|d) = \frac{\sum_{i=1}^{M}\frac{1}{2N_{id}}\sum_{k=1}^{N_i} I(X_{ik} \leq x,\, G_{ik} = d)}{\sum_{i=1}^{M}\frac{1}{2N_{id}}\sum_{k=1}^{N_i} I(G_{ik} = d)},\ d = 0,\ 1.$$

Note that (i) represents the (empirical) distribution of group d $(d = 0, 1)$ data values in the entire sample irrespective of their cluster membership. Calculation (ii) is based on sampling a single paired (e.g., (X, G)) observation from each cluster. In other words, $\widehat{\mathcal{F}}_2(\,\cdot\,|d)$ represents the conditional distribution of a typical outcome value X_{iJ_i} for a typical cluster i, given the corresponding group membership G_{iJ_i} equals d. Here, J_i is a discrete uniform on $\{1, \cdots, N_i\}$. Calculation (iii) is based on computing the proportion of outcomes belonging to group d in a typical cluster i which are less than or equal to x and then taking the average of these proportions over all the clusters. Each of quantities in the right hand sides of (i), (ii), and (iii), can be written as an estimate of $P(X_{ik} \leq x,\, G_{ik} = d)/P(G_{ik} = d)$, but the difference lies in construction of the estimates of the probabilities. In (i) the probabilities are estimated by pooling all the observations together irrespective of their cluster membership, while in (ii), and (iii), the estimates are constructed by conditioning on N_i and N_{id} respectively. Every outcome, belonging to group d and having value less than x, contributes equally in the construction of $\widehat{\mathcal{F}}_1$, but in constructions of $\widehat{\mathcal{F}}_2$ and $\widehat{\mathcal{F}}_3$ we have different contributions from the different outcomes depending on their cluster memberships.

Let $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ be the distribution functions which are estimated by $\widehat{\mathcal{F}}_1$, $\widehat{\mathcal{F}}_2$, $\widehat{\mathcal{F}}_3$ respectively. When the cluster sizes as well as the ICG sizes formed by the two groups within each cluster are not suspected to be associated to the outcome variable in any way, then hypotheses involving $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ become equivalent and one can test any one of these three hypotheses. If there is some association between the cluster size and the outcome variable in that cluster, one can think of testing hypothesis involving $\mathcal{F}_2$ for appropriate comparison. This is a situation of informative cluster sizes. Again, if the ICG sizes formed by the two groups in a cluster appears to be correlated (even after conditioning on the overall cluster size) with the outcomes from the respective groups in that cluster, one may think of testing hypothesis comprising of distribution $\mathcal{F}_3$ instead of $\mathcal{F}_1$ and $\mathcal{F}_2$ to get more meaningful results. We can refer to this as a case of informative ICG sizes. In the absence of this informativeness in the ICG sizes, one can test the null hypotheses of equality of marginal distributions involving any one of the marginal distributions $\mathcal{F}_2$ and $\mathcal{F}_3$, possibly leading to similar conclusion in each case.

In this paper we are interested in comparing $\mathcal{F}_3$ in the two groups when the ICG sizes are potentially informative. Currently, no rank based tests are available for testing group differences for clustered data that takes into account the informativeness of the ICG sizes formed by the groups under study. We denote the common marginal distribution under the null hypothesis as $\mathcal{F}(\,\cdot\,)$. It is perhaps worth pointing out that the estimation of the marginal regression parameters via weighted estimating equations in presence of informative ICG size has been considered by Huang and Leroux (2011).

2.1.1 *Development of the Test Statistic*

For the sake of simplicity, let us relabel the observations according to their group membership within each cluster in the following way. In the i^{th} cluster, let $\{X_{i1}^{(1)}, X_{i2}^{(1)}, \cdots, X_{iN_{i1}}^{(1)}\}$ represent the set of observations belonging to the group indexed by 1, while $\{X_{i1}^{(0)}, X_{i2}^{(0)}, \cdots, X_{iN_{i0}}^{(0)}\}$ represents the set of observations belonging to the group indexed by 0. We denote these sets as $\underline{\mathrm{X}}_i^{(1)}$and $\underline{\mathrm{X}}_i^{(0)}$ respectively. Thus, $\underline{\mathrm{X}}_i^{(1)}$ and $\underline{\mathrm{X}}_i^{(0)}$ form a partition of $\underline{X}_i$={X_{i1}, $\cdots$, X_{iN_i}}, the set of all observations in cluster i. The number of observations belonging to the set $\underline{\mathrm{X}}_i^{(d)}$ $(d = 0, 1)$ is the intra-cluster group size of group d in the i^{th} cluster. Till the end of the Section 2.1.1, we would assume that at least one observation from each group is present in every cluster. In this Section, this assumption means that N_{i1} > 0 and N_{i0} > 0 with probability one for every cluster i. A relaxation of this condition is discussed in Section 2.1.2.

Our test statistic, for testing the hypothesis involving marginal distributions $\mathcal{F}_3$ as estimated in (iii), can be generated from a resampling scheme which is an extension of the within-cluster resampling (WCR). An outline of the resampling scheme is as follows: For each cluster i, let us resample group membership as G_i^*, where G_i^* takes value 0 or 1 with equal probability $\frac{1}{2}$. If G_i^* = 1, we resample one observation for the i^{th} cluster from the set of observations $\underline{\mathrm{X}}_i^{(1)}$ and name it X_i^*. If G_i^* = 0, resample X_i^* from the set $\underline{\mathrm{X}}_i^{(0)}$.

The fact that the outcomes are resampled from the subsets formed by the two groups in a cluster and not from the whole cluster makes this resampling scheme different from the usual WCR technique. Now, this resampling gives us M pairs of independent observations (X_i^*, G_i^*). If R_i^* be the rank of X_i^* among the set $\{X_j^*, 1 \leq j \leq M\}$, i.e., R_i^*= 1+0.5$\Big\{\sum_{j\neq i} I(X_j^* \leq X_i^*) + \sum_{j\neq i} I(X_j^* < X_i^*)\Big\}$, then the Wilcoxon rank sum statistic based on these M pairs of resampled observations (X_i^*, G_i^*) would be of the form : $S^* = \sum\limits_{i=1}^{M} G_i^* R_i^*$. One can use S^* as a valid test statistic and carry out the test based on S^*. But that test would be inefficient as the test statistic would depend too much on one particular observation chosen from each cluster. So to get rid of the imposed randomization due to resampling we propose a test statistic based on earlier approaches of Williamson et. al (2003), Datta and Satten (2005), and Datta and Satten (2008), that corresponds to averaging S* over all possible choices of $\left(X_i^*,\ G_i^*\right)$ values given the data.

Thus, our test statistic is $T = E(S^* |\, X, G)$, where $X = \{X_{ik} : 1 \leq k \leq N_i;$ $1 \leq i \leq M\}$ and $G = \{G_{ik} : 1 \leq k \leq N_i\, ; 1 \leq i \leq M\}$. We can calculate the theoretical expression of T. After some necessary steps, a convenient expression of T (see Technical Details for the detailed steps) turns out to be

$$T = \sum_{i=1}^{M}\left(\sum_{k=1}^{N_{i1}}\frac{1}{2N_{i1}}\left[1 + \frac{1}{2}\sum_{j\neq i}\left\{F_j\left(X_{ik}^{(1)}\right) + F_j\left(X_{ik}^{(1)}-\right)\right\}\right]\right),$$

where $F_j(x) = \frac{1}{2N_{j1}}\sum_{h=1}^{N_{j1}} I\left(X_{jh}^{(1)} \leq x\right) + \frac{1}{2N_{j0}}\sum_{h'=1}^{N_{j0}} I\left(X_{jh'}^{(0)} \leq x\right)$.

Besides T, we need to know its expected value $E(T)$ and its variance estimate $\widehat{V}(T)$ to properly carry out inference based on T. To get $E(T)$ we note that $E(T) = E(S^*)$. The unconditional expectation of S^*can be calculated easily through conditioning on the vector of group membership indicator $G^* = (G_1^*, G_2^*, ..., G_M^*)$. So we get, $E(T) = E(S^*) = E(E(S^*|G^*)) = E\left(\sum_{i=1}^{M} G_i^*(\frac{M+1}{2})\right) = \left(\frac{M+1}{2}\right)\sum_{i=1}^{M}\frac{1}{2} = \frac{M(M+1)}{4}$.

The next step is to find a variance estimate $\widehat{V}(T)$. To get the variance estimate of T, we employ the jackknife technique. Here the clusters can be thought of as iid units and thus we can use a 'delete-1-cluster' jackknife approach to get the necessary results. Mathematically, this can be formulated as follows. Let T_{-i} be the value of the statistic T calculated after deleting the i^{th} cluster. Let us define, $T_i^* = T - T_{-i}$. Then the estimate of variance of T, which is the jackknife variance estimate, is given by

$$\widehat{V}(T) = \widehat{V}_{\text{JK}} = M.Var(T_i^*) = \frac{M}{M-1}\sum_{i=1}^{M}\left(T_i^* - \overline{T}^*\right)^2.$$

Now that we have the expressions for T, $E(T)$, $\widehat{V}(T)$, we can carry out the testing using the absolute value of the standardized statistic $Z = (T - E(T))/(\{\widehat{V}(T)\}^{1/2})$.

The asymptotic distribution of Z is established through the following theorem. An outline of its proof is given in the Technical Details.

THEOREM 1 (Asymptotic normality). *Under* H_0*, as* $M \to \infty, Z \xrightarrow{d} N(0,1)$ *under certain regularity conditions of a Lindeberg Central Limit Theorem.*

The p-value for the test is computed as the probability that, under H_0, the absolute value of the Z-statistic exceeds its observed value in magnitude. We would reject the null hypothesis H_0 at a 100α% level of significance if the p-value is less than α.

Till this point we have assumed the existence of only two groups in every cluster. In Technical Details, we have discussed a more general situation where there are m groups in every cluster, such that $m > 2$.

2.1.2 *Extension to Incomplete Intra-cluster Group Structure in One or More Clusters*

In case of binary grouping, (i.e., $G_{ik} = 0$ or 1),we have assumed that there is at least one observation from each group in every cluster. In practice, one may encounter a few clusters (not all) with one group of observations completely missing. In other words, there may be some clusters having outcomes from only one of the two possible groups. We call such a case as incomplete informative intra-cluster group structure within a cluster. The hypothesis of interest remains the same, viz., whether the marginal distributions of outcomes are same for the two groups. We cannot directly apply the test statistic in the form described in Section 2.1.1 to this setting. This is mainly because of the fact that the test statistic developed in Section 2.1.1 is only applicable under the assumption that outcomes from both groups are available within each cluster. We extend the approach described in Section 2.1.1, to get a valid test statistic in this setting.

Here we follow the same notations as described in Section 2.1.1. In cases of incomplete ICG structures within a cluster, the empirical analogue of the "group specific" marginal distributions of our interest can be constructed as a modification of $\widehat{\mathcal{F}}_3$ as $\widehat{\mathcal{F}}_4(x|d) = \sum_{i=1}^{M} W_{id} \sum_{k=1}^{N_i} I(X_{ik} \leq x, G_{ik} = d) / \{\sum_{i=1}^{M} W_{id} \sum_{k=1}^{N_i} I(G_{ik} = d)\}$, where $W_{id} = (2N_{id})^{-1}$, or N_{id}^{-1}, or 0, according to whether the i^{th} cluster has observations from both groups, the dth group only, or not.

We extend the idea of within cluster resampling also to this setting to get a valid test statistic. (1) If both $N_{i1} > 0$ and $N_{i0} > 0$, group membership is resampled as G_i^*, where G_i^* takes value 0 or 1, with equal probability $\frac{1}{2}$. If G_i^*=0, resample X_i^* from $\underline{\mathrm{X}}_i^{(0)}$; otherwise, if G_i^*=1, resample X_i^* from $\underline{\mathrm{X}}_i^{(1)}$. (2) If $N_{i1} = 0$ and $N_{i0} = N_i > 0$, we resample X_i^* from $\underline{\mathrm{X}}_i^{(0)}$ and have G_i^*=0. Here $\underline{\mathrm{X}}_i^{(0)}$ is same as $\underline{\mathrm{X}}_i$ as the set $\underline{\mathrm{X}}_i^{(1)}$ is an empty set. (3) If $N_{i0} = 0$ and $N_{i1} = N_i > 0$, we resample X_i^* from $\underline{\mathrm{X}}_i^{(1)}$ and have G_i^*=1. Here $\underline{\mathrm{X}}_i^{(1)}$ is same as $\underline{\mathrm{X}}_i$ as the set $\underline{\mathrm{X}}_i^{(0)}$ is an empty set.

To obtain our test statistic T in this case, we proceed in the same way as in Section 2.1. With R_i^* being the rank of X_i^* among the set $\{X_j^*,\, 1 \leq j \leq M\}$, we obtain $\mathrm{S}^* = \sum_{i=1}^{M} G_i^* R_i^*$, the Wilcoxon rank sum statistic based on the M pairs of resampled observations (X_i^*, G_i^*). Then, our proposed test statistic T is calculated as $T = E(\mathrm{S}^*|X, G)$. After some algebra (see Technical Details) we obtain T as

$$T = \sum_{i=1}^{M} \sum_{k=1}^{N_{i1}} \frac{1}{2N_{i1}} \Big[\big(I(N_{i1} > 0, N_{i0} > 0)\big) + 2I\big(N_{i1} > 0, N_{i0} = 0\big)$$

$$+ \frac{I(N_{i1} > 0, N_{i0} > 0) + 2N_{i1}.I(N_{i0} = 0)}{2} \sum_{j \neq i} \Big\{ F_j^{'}(X_{ik}^{(1)}) + F_j^{'}(X_{ik}^{(1)} -) \Big\} \Big],$$

where

$$F_j^{'}(x) = \Big\{ \frac{1}{2N_{j1}} \sum\nolimits_{h=1}^{N_{j1}} I(X_{jh}^{(1)} \leq x) + \frac{1}{2N_{j0}} \sum\nolimits_{h'=1}^{N_{j0}} I(X_{jh'}^{(0)} \leq x) \Big\} I(N_{j1} > 0, N_{j0} > 0)$$

$$+ \{1 - I(N_{j1} > 0, N_{j0} > 0)\} \Big\{ \frac{1}{N_j} \sum\nolimits_{h=1}^{N_j} I(X_{jh} \leq x) \Big\}.$$

The expected value of the test statistic is estimated to be

$$\widehat{E}(T) = \frac{(M+1)}{4} \sum_{i=1}^{M} \Big\{ I(N_{i1} > 0, N_{i0} > 0) + 2I(N_{i1} > 0, N_{i0} = 0) \Big\}.$$

Now, to find the estimated variance $\widehat{V}$ of $T - \widehat{E}(T)$, we use the same 'delete-1-cluster' jackknife approach described in Section 2.1. Finally, as in Section 2.1.1, we carry out the testing using the standardized Z-statistic $Z = \{T - \widehat{E}(T)\}/\{\widehat{V}\}^{1/2}$, that has asymptotic $N(0,1)$ distribution under H_0.

2.2. Simulation Results

In this Section we present three simulation studies corresponding to the tests discussed in the Sections 2.1.1 and 2.1.2. In the simulation scenario 1, we consider clustered observations such that every cluster has outcomes from both the groups. In each cluster, the number of observations belonging to group 1 and the number of observations belonging to group 0, that is the two ICG sizes, are both influenced by some latent factor, that also influences the outcomes in that cluster. Also, the distributions of the two ICG sizes, within each cluster, differ between themselves. So, there is some association between the ICG sizes and outcomes in a given cluster (even after conditioning on the overall cluster size) and we can think of this as informative ICG sizes. Under this simulation scenario, we compare the performances of four tests, namely, (1) our new rank sum test developed in Section 2.1.1, (2) the test by Datta and Satten (2005), (3) the naive Wilcoxon rank sum test assuming all the observations as iid and ignoring their cluster membership, and (4) the signed rank test taking cluster averages for each group of observations . Further, each test was carried out under three different choices of the number of clusters (M), namely, 30, 50 and 150. In simulation scenario 2, we generate a setting that closely represents the dental setting discussed in Section 1. Basically, the idea is to have a clustered data with informative ICG sizes, where the number of units belonging to each group in a cluster cannot exceed a certain value. Under this setting we compare the four tests (1)-(4) for 50 clusters. In scenario 3, we again consider informativeness in the ICG sizes, but we do not restrict ourselves to the condition that observations from both the groups have to be present in each cluster. In other words, we include the cases of incomplete ICG structures within a cluster for which our test statistic developed in Section 2.1.2 looks appropriate. We investigate the performance of this new test for a simulation model with 30 clusters under scenario 3. Additionally, we consider two more simulation scenarios (Scenario 4 and Scenario 5), where we compare the four tests (1)-(4) under situations such that either the ICG sizes or both the ICG sizes and the cluster sizes are noninformative.

Performances of all the tests are evaluated on the basis of their sizes (nominal $\alpha =$ 0.05) and power values. These are estimated by the proportion of 3,000 Monte Carlo iterates in which null hypothesis is rejected.

2.2.1 Simulation Scenario 1

Let M be the number of clusters (fixed). For a typical cluster i, we define, N_{i1}as the number of observations from group 1 in the i^{th} cluster, N_{i0} as the number of observations from group 0 in the i^{th} cluster, a_i as the random cluster effect due to the i^{th}cluster. In the i^{th} cluster, we generate a_i from Normal$(0, 0.25)$ distribution, N_{i1}^* from Poisson(10+5a_i) distribution where $N_{i1} = N_{i1}^*$+1, N_{i0}^* is generated from Poisson(10+5a_i^2) such that N_{i0}=N_{i0}^*+1. Also, we know that $N_i = N_{i1} + N_{i0}$. Let G_{ij} be the group indicator of the j^{th} observation in the i^{th} cluster. We assign $G_{ij} = 0$ for $1 \leq j \leq N_{i0}$, while $G_{ij} = 1$ for $N_{i0} + 1 \leq j \leq N_i$. We generate Y_{ij}, the j^{th} outcome in the i^{th} cluster, through a random effects model as $Y_{ij} = 0.5 + a_i + e_{ij}$, $1 \leq j \leq N_i$, $1 \leq i \leq M$, such that if $G_{ij} = 0$, then $e_{ij} \sim$ Normal$(0,\ 0.3)$, while if $G_{ij} = 1$, then $e_{ij} \sim$ Normal$(\delta, 0.3)$. Under the null model, δ = 0.

Performances of the four tests (1)-(4) are summarized in Table 2.1 for three choices of M, namely, 30, 50 and 150.

Table 2.1 illustrates a number of points. Our new test closely maintains the nominal size and is sufficiently strong in terms of power even under small effect sizes. The rank sum test proposed by Datta and Satten (2005) and the standard Wilcoxon rank sum test have grossly inflated size and very low power compared to our test for all three choices of the number of clusters. The size of the cluster average signed rank test tends to be close to the nominal size under this simulation scenario. Its power is also close to our test, though a bit less in almost all cases. Although the clustered average signed rank test appears to be a good competitor of our test in this simulation scenario, one can acknowledge the fact that the distribution of the average of independent and identical random variables is not always same as that of the individual variables. Thus, it is expected that the cluster average signed rank test is not a good choice for testing the hypothesis of our interest and this fact might be evident if we have widely different ICG sizes within each cluster.

2.2.2 Simulation Scenario 2

This simulation setting is carried out to mimic the setting of dental study mentioned in Section 1, where the number of units (teeth) within a cluster (mouth of an individual) cannot exceed 32. This can be generalized for any study where the cluster sizes or the ICG sizes are bounded.

This simulation scenario is almost same as that described in Section 2.2.1, the only difference being that both the ICG sizes within each cluster are less than or equal to 16, such that the cluster size cannot exceed 32. Following the same notations for the quantities in 2.2.1, in the i^{th} cluster, we generate a_i from Normal(0,0.25), N_{i1}^* from Poisson(10+5a_i) such that $N_{i1} \leq$ 16, N_{i0}^* from Poisson(10+5a_i^2) such that $N_{i0} \leq$ 16. So, we have $N_i = \ N_{i1} + N_{i0} \ \leq$ 32. Apart from these, G_{ij}, e_{ij}, and the outcome Y_{ij} are generated in the same manner as in simulation scenario 1. Table 2.2 compares the four tests (1)-(4) under this simulation scenario with the number of clusters (M) as 50, and the results are similar to the results obtained from simulation scenario 1. Table 2.2 shows that our new test closely maintains the nominal size and has substantial power under a variety of effect sizes. The rank-sum test proposed by Datta and Satten (2005), as well as the standard Wilcoxon rank sum test, has highly inflated size. The clustered average signed rank test, just like in simulation scenario 1, apparently maintains the nominal size and has substantial power. But, as mentioned before in Section 2.2.1, theoretically it is not a good choice for testing the hypothesis of our interest.

2.2.3 Simulation Scenario 3

This simulation scenario is almost similar to that described in Section 2.2.1, the only difference being that the ICG sizes within each cluster are not restricted to be strictly positive always. Following the same notations for the quantities in 2.2.1, in the i^{th} cluster, we generate a_i from Normal(0,0.25), N_{i1} from Poisson(10+5a_i), N_{i0} from Poisson(10+5a_i^2). We have $N_i = N_{i1} + N_{i0}$. Apart from these, G_{ij} , e_{ij}, and the outcome Y_{ij} are generated in the same manner as in simulation scenario 1. Evidently, observed values of any of the ICG sizes N_{i1} and N_{i0} in the i^{th} cluster can be 0, as long as $N_i > 0$.

In Table 2.3, we evaluate the empirical size and power of our test, developed in Section 2.1.2, with the choice of $M = 30$. Thus, from Table 2.3, we see that our test closely mimics the nominal size and has moderate to high power under different effect sizes.

2.2.4 Simulation Scenario 4

In this simulation scenario, for a typical cluster i, we generate N_i^* from a Poisson(20+5a_i^2) distribution, such that $N_i = N_i^* + 2$. Then, we generate N_{i1} as $[N_i/2]$, where $[x]$ is the largest integer not exceeding x. Also, $N_{i0} = N_i - N_{i1}$. Here a_i, G_{ij}, e_{ij}, Y_{ij} are generated in the same way as in Section 2.2.1. This is a scenario of clustered data with informative cluster sizes, but the ICG sizes in a given cluster are not informative once we condition on the overall size of that cluster. We compare the empirical performances of the four tests (1)-(4) for $M = 50$ in Table 2.4. Table 2.4 reveals that under this simulation of informative cluster sizes, the naive Wilcoxon rank-sum test is very conservative as its empirical size is far below the nominal size of 0.05. But all the other three tests including our new test closely maintain the nominal size. Also, for small effect size, the power of the naive Wilcoxon rank sum test is lower than that of the other three tests. This difference in power slowly decreases with increase in the effect size. Our new test from Section 2.1.1, the test by Datta and Satten, as well as the cluster average signed rank test closely agree in their power performances under different effect sizes.

2.2.5 Simulation Scenario 5

This simulation scenario differs from the previous simulation setting as we do not consider informativeness in any of the cluster sizes or ICG sizes. For a typical cluster i, $N_{i1}^{*} \sim$ Poisson(10), $N_{i1} = N_{i1}^{*} + 1$, $N_{i0}^{*} \sim$ Poisson(10), $N_{i0} = N_{i0}^{*} + 1$, and $N_i = N_{i1} + N_{i0}$. The other random quantities a_i, G_{ij}, e_{ij} as well as the outcome Y_{ij}, are generated in the exact same way as in simulation scenario 1. In Table 2.5, we compare the empirical size and power of the four tests (1)-(4) under this simulation model. Looking at this table, we find that all the four tests closely maintain the nominal size. But our new test and the paired signed rank test with cluster averages have superior power than the other two tests in this simulation setting.

2.3. Application to Dental Data

We consider data from the Piedmont 65+ Dental study by Beck *et al.* (1990). This study examined two older populations, urban whites and urban and rural blacks. The Piedmont Health Study of the Elderly by Blazer and George (2004), which was the parent study for this Piedmont 65 + Dental Study, was a longitudinal study of the health status of a stratified, clustered, random sample of people aged 65 and over in five contiguous North Carolina counties. The Piedmont 65+ Dental Study used the data available from the parent study while collecting additional information. For the Piedmont 65+ Dental Study, we have the gingival recession and pocket depth measures for all teeth present in the mouth, at baseline, 18, 36 and 60 months, respectively. Attachment level scores (attachment losses) were computed from the gingival recession and pocket depth measures. Also, all these clinical measures were computed for two sites, buccal and mesial, for every tooth measured. A number of additional covariates were also available which are ignored for the present marginal analyses. The number of subjects observed varied across the four data points. This may be because, being a study involving elderly population, many subjects who were reported at the beginning of the study failed to come back at later time points of the study. For our illustration, we investigate the baseline and 18 month data cross-sectionally.

Attachment loss is a common problem associated with periodontal diseases in elderly population, often indicating the severity of certain diseases. It has been suggested in some studies that the nature of attachment loss varies across the different surfaces of a tooth. Suspecting one such possibility, it may be interesting to identify whether the distributions of attachment loss scores are same for the buccal and mesial surfaces of teeth. Since the outcomes (attachment level scores) from the units (teeth surfaces) within a cluster (individual) are correlated, while that from the units between different clusters are independent, the data fall into the category of the type of clustered data we are interested in. In addition since the cluster size (number of teeth surfaces an individual has) may indicate the overall oral health, the cluster size might be associated to the outcome of interest (attachment loss score). We apply our new test and the test of Datta and Satten to investigate possible differences in the distributions of attachment loss at the buccal and mesial sites (the two groups under study) to data at baseline involving 697 subjects with at least one tooth. A significant difference was obtained for the novel test (Z=-10.29, p-value=7.92×10^{-25}) and for the Datta and Satten test (Z=-9.40, p-value=5.56×10^{-21}). So, our new test and the test by Datta and Satten lead to the same conclusion but with different p-values. We then consider the same testing problem but with the data for 18 month (with 496 available subjects) where, again, significant difference was obtained using the new test (Z=-11.49, p-value=1.48×10^{-30}) as well as the test by Datta and Satten (Z=-9.94, p-value=2.72×10^{-23}). Overall, we conclude that the distribution of the attachment loss of teeth differs between the mesial and buccal sites. Also, we see that our new test gives consistent result in a situation where the test by Datta and Satten appears to be valid as well. Plots of the empirical cumulative distribution functions $\left(\widehat{\mathcal{F}}_3(.)\right)$ of attachment scores in the two groups (buccal and mesial) are shown for both the baseline data and the 18-month data in Figure 2.1 and Figure 2.2 respectively. Some indications regarding the significant difference in the distributions of attachment scores between buccal and mesial sites can be obtained from these figures. In addition, plots of the empirical mass functions for mesial and buccal attachment loss scores at baseline study are given in Figure 2.3 that explain the substantial differences between the mesial and buccal attachment loss scores at the low score values of 1 and 2. Incidentally, these two scores together constitute more than half of the observed scores for the population under study. To calculate the effect size we use the following approach: if $\underline{Y}^{(0)}$and $\underline{Y}^{(1)}$ denote the sets of mesial and buccal attachment scores, such that the test statistic $T = T(\underline{Y}^{(0)}, \underline{Y}^{(1)})$, and Δ be a real number such that $T_\Delta = T(\underline{Y}^{(0)}, \underline{Y}^{(1)} + \Delta)$, then the effect size is estimated by the absolute value of Δ^*, where $\Delta^* = \sup\ \{\Delta : T_\Delta - E(T) = 0\}$. For both the baseline and 18 months data unstandardized effect size turns out to be approximately 0.5.

Another interesting question, as discussed previously in Section 1, would be whether the distributions of attachment loss scores differ between the teeth of upper and lower jaws. To investigate this fact using the same data, we have considered attachment loss at the mesial site of tooth, although one can also pose the same question with the buccal site. The null hypothesis here is that the distribution of attachment loss at the mesial site of a tooth is the same for the upper and lower jaws. Here the setting for this problem is quite similar to that of the previous problem. The difference is that in this setting the mesial site attachment loss score (outcome) of a tooth (unit) in any particular jaw (group) of an individual (cluster) may be related to the number of teeth present in that jaw of that individual. So, we may have some informativeness in the ICG size (number of teeth present in a jaw of an individual) even after conditioning on the cluster sizes. We consider the 60 month data for this analysis with 292 available subjects at that point. This data falls under the category of clustered data with some clusters having incomplete ICG structures, as described in Section 2.2, because there are a few subjects (clusters) who have teeth (units) in only one of the two jaws (groups). Our new test, developed in Section 2.2, is the only test that can be used to test the hypothesis under this setting and it gives a p-value of 4.06×10^{-5} (Z = 4.10). Thus, we conclude that there is a significant difference between the distributions of the attachment loss at the mesial sites of the upper and lower jaws. The estimated effect size, estimated like before, comes out to be around 3.0 units for this data. Figure 2.4 shows the empirical cumulative distribution functions $\left(\widehat{\mathcal{F}}_4(.)\right)$ for the attachment loss scores of upper and lower sets of teeth.

2.4. Discussions

For clustered data with informative cluster sizes, the ordinary rank-sum test assuming independent observations can be biased as indicated in a simulation study in Section 3. The rank-sum test by Datta and Satten (2005), which compares group-specific marginal distributions $\mathcal{F}_2$, appears to be a valid test under informative cluster sizes. But when an outcome from a group d $(d = 0,\, 1)$ in a typical cluster depends on the number of observations from the group d in that cluster, we have informativeness in the ICG sizes formed by the two groups. As discussed earlier in Section 1 and Section 4, this type of clustered data with informative ICG sizes are common in dental studies. Simulation studies from Section 3 indicate that even the rank-sum test by Datta and Satten (2005) has inflated size under this scenario of informative ICG size. There are no rank based tests in current literature that address this issue of informative ICG sizes. Thus, our main focus was to develop a rank-sum test for clustered data which works under this scenario of informative ICG sizes. This has led us to compare group-specific marginal distribution $\mathcal{F}_3$ that gives equal weights to each cluster (treating cluster as the basic sampling unit), but the weight given to an outcome from group d in a cluster depends on the number of observations from group d in that cluster. This is in contrast with $\mathcal{F}_2$ where the weight given to an outcome from a typical cluster depends on the number of outcomes in that cluster ignoring the information on the group membership of that outcome. Thus, the question of importance is which marginal distribution should be considered in testing hypothesis. It appears that comparing $\mathcal{F}_3$ may be more meaningful under informative ICG sizes and through a number of simulation settings, we have showed that our test maintains the nominal size and has substantial power in clustered data with informative ICG sizes. Even when the ICG sizes are not informative, simulation studies reveal that our test closely maintains the nominal size and has acceptable power when compared to other rank tests based on $\mathcal{F}_2$ or $\mathcal{F}_1$.

As we consider clustered data, we may, in practice, encounter a few clusters which have outcomes from only one of the two groups under study. In that case, there are two possible ways of addressing this issue. One simple way is to ignore the clusters which do not have outcomes from both the groups and carry out the test, developed in Section 2.1, based on the remaining clusters. But, oftentimes, it is suspected that the information on the outcome of interest may be different between clusters with incomplete ICG structures (i.e., clusters with observations from one of the two groups) and clusters having both groups of observations. Keeping this in mind, we extended our test, in Section 2.2, to account for the clusters with incomplete ICG structures, so that we effectively use all the information present in the data. A simulation study showed that our test has the correct size and substantial power for a model accommodating incomplete ICG structures with informative ICG sizes. But, one can expect the power of this test to be low compared to that of the test involving only clusters with complete ICG structures. Therefore, in presence of a few clusters with incomplete ICG structures among a large number of clusters, it might be important to decide beforehand whether to apply the test developed in Section 2.1 ignoring a few clusters or to use the test from Section 2.2 keeping the full data. In case of clustered data where the outcomes within the same cluster belong to the same group, our test statistic reduces to that of Datta and Satten (2005), and, thus, will have superior size and power performance than the rank-sum test by Rosner et al. (2003) when the correlation structure within a cluster depends on the group membership.

Sometimes, when testing for group effect in outcomes from clustered data, one can expect the presence of some additional covariate(s) unrelated to the grouping factor. In such cases these additional covariates (confounders) may act as nuisance factors in comparing the group-specific marginal distributions of the outcomes. For example, suppose we have a linear regression of the form

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \epsilon_{ij},\ 1 \leq i \leq M\ , 1 \leq j \leq N_i.$$

Here Y_{ij} is the outcome of the j^{th} observation in the i^{th} cluster, X_1 is the binary indicator variable taking value 1 or 0 according to the group membership, X_2 and X_3 are the confounders (unrelated to the group membership) and ϵ is the random error following some unknown distribution F_ϵ. To compare the group-specific marginal distributions of the outcomes, one may want to test the null hypothesis $H : \beta_1 = 0$ against the alternative hypothesis $K : \beta_1 \neq 0$. But, if the distributions (unknown) of the confounders are different from that of the random error and also among themselves, then the rank tests based on the outcome Y can be misleading. This is, in general, true for any regression model involving confounders. To overcome this, one, often, uses aligned rank tests (see, e.g., Hájek, Šidák, and Sen, 1999, Section 10.1.2). The basic idea involves estimation of the (nuisance) parameters relating to the confounders through some appropriate rank statistics, formation of aligned observations (residuals) by plugging in the estimates and then developing a rank test based on the aligned observations. In presence of informative ICG size, one can extend the resampling technique discussed in this article to formulate suitable rank based statistics for estimating the nuisance parameters and testing the appropriate (sub)hypothesis under aligned rank tests.

Table 2.1

Size, along with a 95% confidence interval, and power comparisons of four tests (nominal $\alpha = 0.05$) under Simulation Scenario 1.

	M= 30 clusters			
Test	Size (CI)	Power (under effect size δ)		
		δ = 0.05	δ = 0.10	δ = 0.15
New Test	0.060 $(0.052,\ 0.068)$	0.319	0.833	1.000
DS	0.132 $(0.120,\ 0.144)$	0.050	0.203	0.500
W	0.159 $(0.146,\ 0.172)$	0.058	0.263	0.645
CA	0.055 $(0.047,\ 0.063)$	0.296	0.814	0.985

	M = 50 clusters			
Test	Size (CI)	Power (under effect size δ)		
		δ = 0.05	δ = 0.10	δ = 0.15
New Test	0.053 $(0.045,\ 0.061)$	0.500	0.960	1.000
DS	0.199 $(0.185,\ 0.213)$	0.050	0.310	0.730
W	0.215 $(0.200,\ 0.230)$	0.061	0.390	0.830
CA	0.051 $(0.043,\ 0.059)$	0.460	0.950	1.000

	M = 150 clusters			
Test	Size (CI)	Power (under effect size δ)		
		δ = 0.05	δ = 0.10	δ = 0.15
New Test	0.055 $(0.047,\ 0.063)$	0.910	1.000	1.000
DS	0.508 $(0.490,\ 0.526)$	0.052	0.699	0.900
W	0.528 $(0.510,\ 0.546)$	0.073	0.778	0.993
CA	0.050 $(0.042,\ 0.058)$	0.896	1.000	1.000

New Test = Test developed in Section 2.1.1, DS= rank-sum test by Datta and Satten, W=Wilcoxon rank-sum test, CA=signed rank test with cluster averages

Table 2.2

Size, along with a 95% confidence interval, and power comparisons of four tests (nominal $\alpha = 0.05$) under Simulation Scenario 2. The number of clusters, M, equals 50.

Test	Size (CI)	Power (under effect size δ)		
		δ = 0.05	δ = 0.10	δ = 0.15
New Test	0.054 $(0.046,\ 0.062)$	0.465	0.960	1.000
DS	0.146 $(0.133,\ 0.159)$	0.071	0.442	0.845
W	0.136 $(0.124,\ 0.148)$	0.073	0.501	0.916
CA	0.047 $(0.039,\ 0.055)$	0.445	0.960	1.000

New Test = Test developed in Section 2.1.1, DS= rank-sum test by Datta and Satten, W=Wilcoxon rank-sum test, CA=signed rank test with cluster averages

Table 2.3

Size, along with a 95% confidence interval, and power calculations (nominal $\alpha = 0.05$) of the new test developed in Section 2.1.2 under Simulation Scenario 3 . Note that the CA test statistic is not computable in this situation. The number of clusters, M, equals 30.

Size (CI)	Power (under effect size δ)		
	δ = 0.05	δ = 0.10	δ = 0.15
0.053 $(0.045,\ 0.061)$	0.275	0.743	0.964

Table 2.4

Size, along with a 95% confidence interval, and power comparisons of four tests (nominal $\alpha = 0.05$) under Simulation Scenario 4.

	M= 50 clusters			
Test	Size (CI)	Power (under effect size δ)		
		δ = 0.05	δ = 0.10	δ = 0.15
New Test	0.057 $(0.049,\ 0.065)$	0.741	0.999	1.000
DS	0.049 $(0.041,\ 0.057)$	0.693	0.999	1.000
W	0.018 $(0.013,\ 0.023)$	0.539	0.998	1.000
CA	0.051 $(0.043,\ 0.059)$	0.746	0.999	1.000

New Test = Test dveloped in Section 2.1.1, DS= rank-sum test by Datta and Satten, W=Wilcoxon rank-sum test, CA=signed rank test with cluster averages

Table 2.5

Size, along with a 95% confidence interval, and power comparisons of four tests (nominal $\alpha = 0.05$) under Simulation Scenario 5.

	M= 50 clusters			
Test	Size (CI)	Power (under effect size δ)		
		δ = 0.05	δ = 0.10	δ = 0.15
New Test	0.065 $(0.056,\ 0.074)$	0.699	0.998	1.000
DS	0.053 $(0.045,\ 0.061)$	0.508	0.975	1.000
W	0.051 $(0.043,\ 0.059)$	0.539	0.988	1.000
CA	0.048 $(0.040,\ 0.056)$	0.688	0.998	1.000

New Test = Test dveloped in Section 2.1.1, DS= rank-sum test by Datta and Satten, W=Wilcoxon rank-sum test, CA=signed rank test with cluster averages

Figure 2.1. Plot of empirical cumulative distribution functions $\left(\widehat{\mathcal{F}}_3(.)\right)$ of attachment scores in buccal and mesial sites at baseline study.

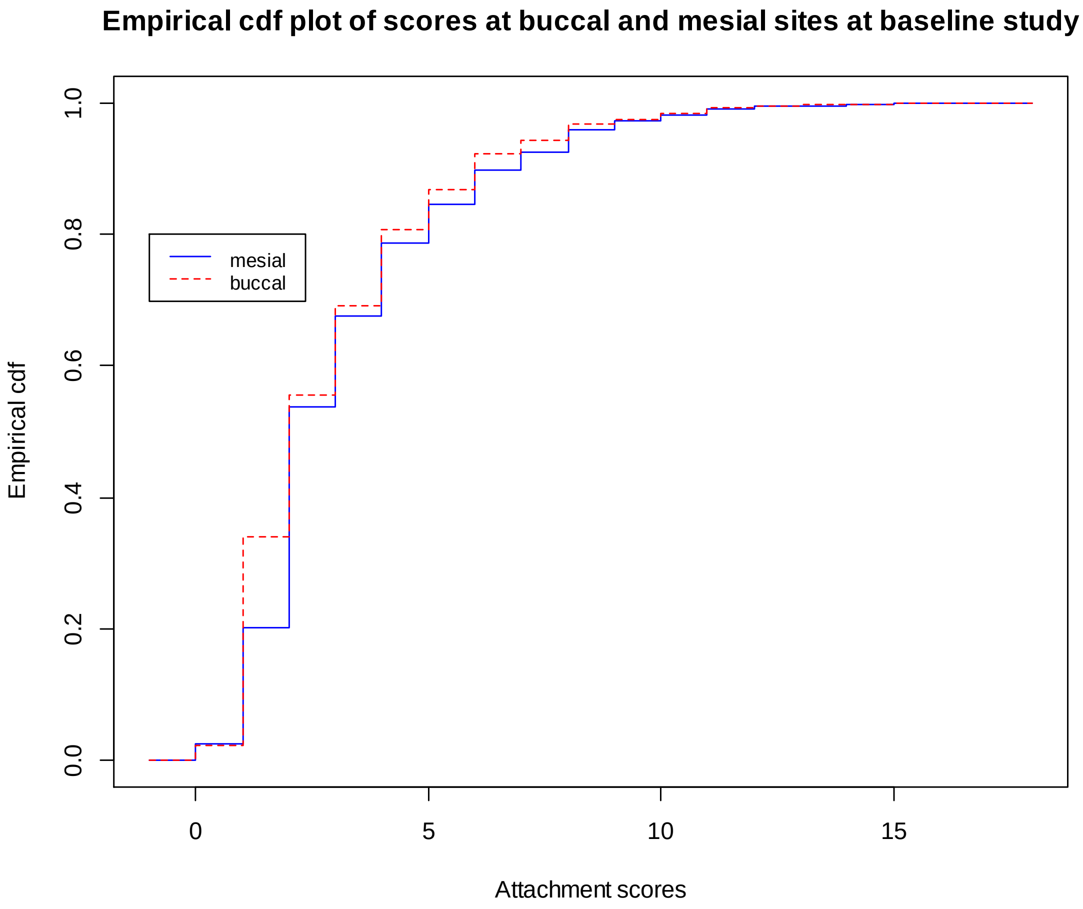


Figure 2.2. Plot of empirical cumulative distribution functions $\left(\widehat{\mathcal{F}}_3(.)\right)$ of attachment scores in buccal and mesial sites at 18 months.

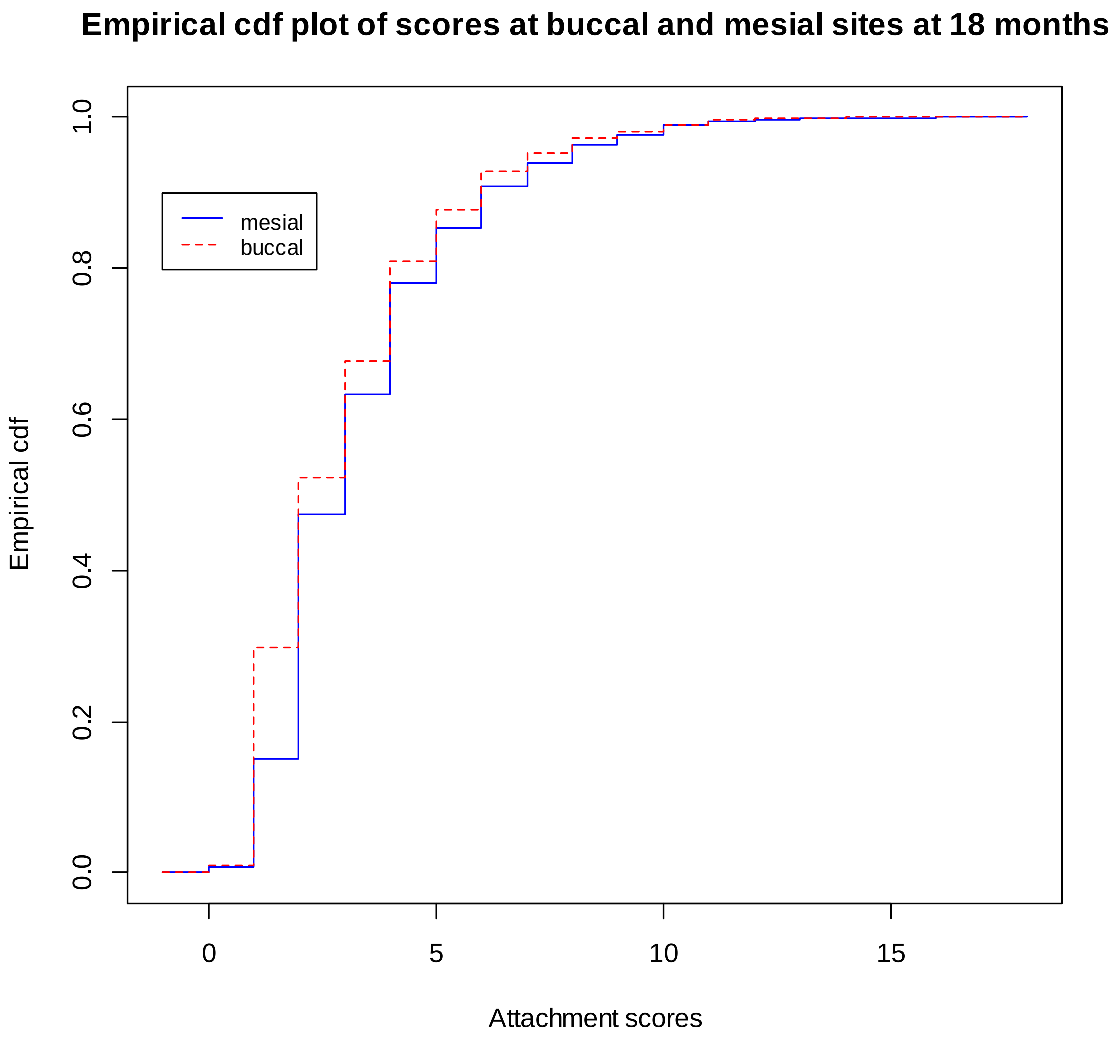


Figure 2.3. Plots of empirical mass functions for mesial and buccal attachment loss scores at baseline study.

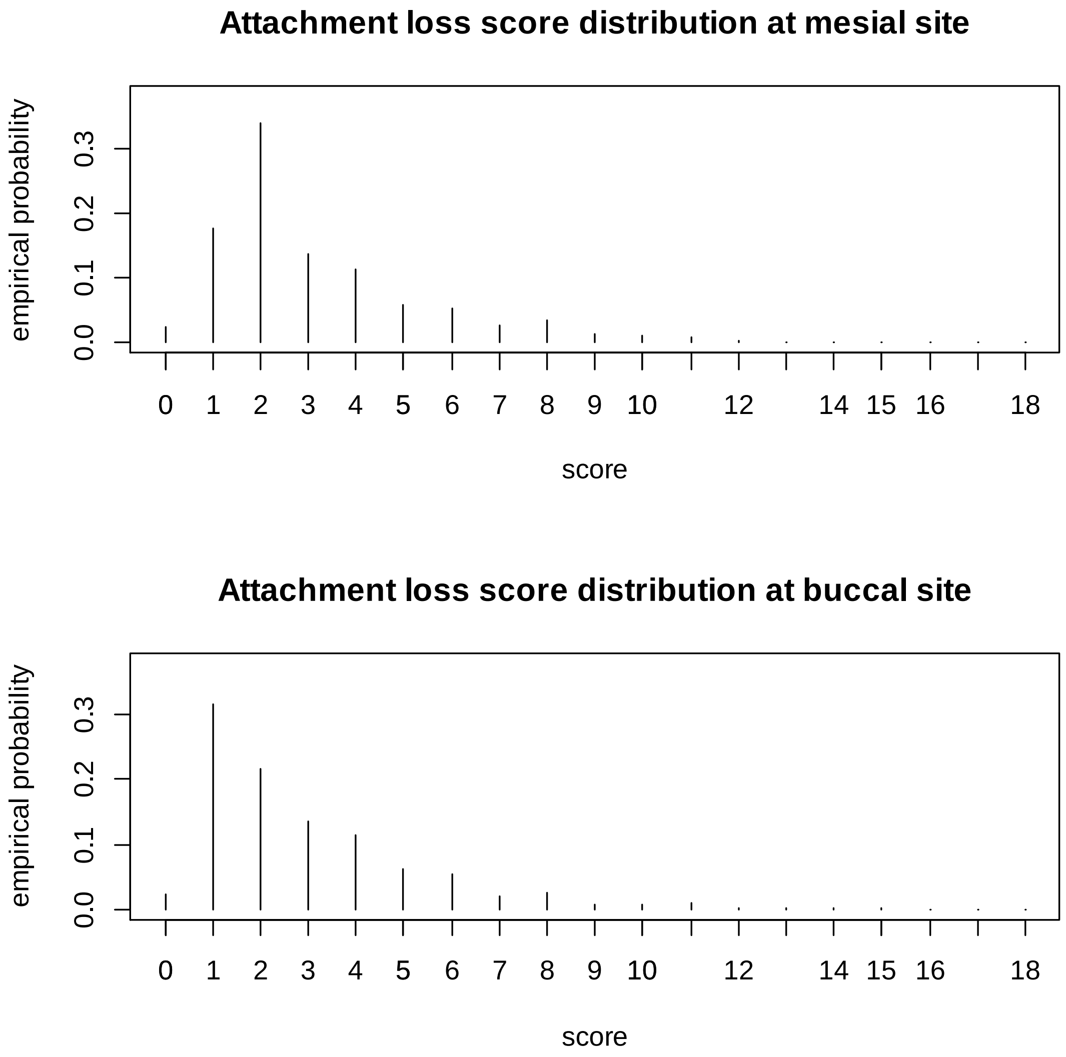
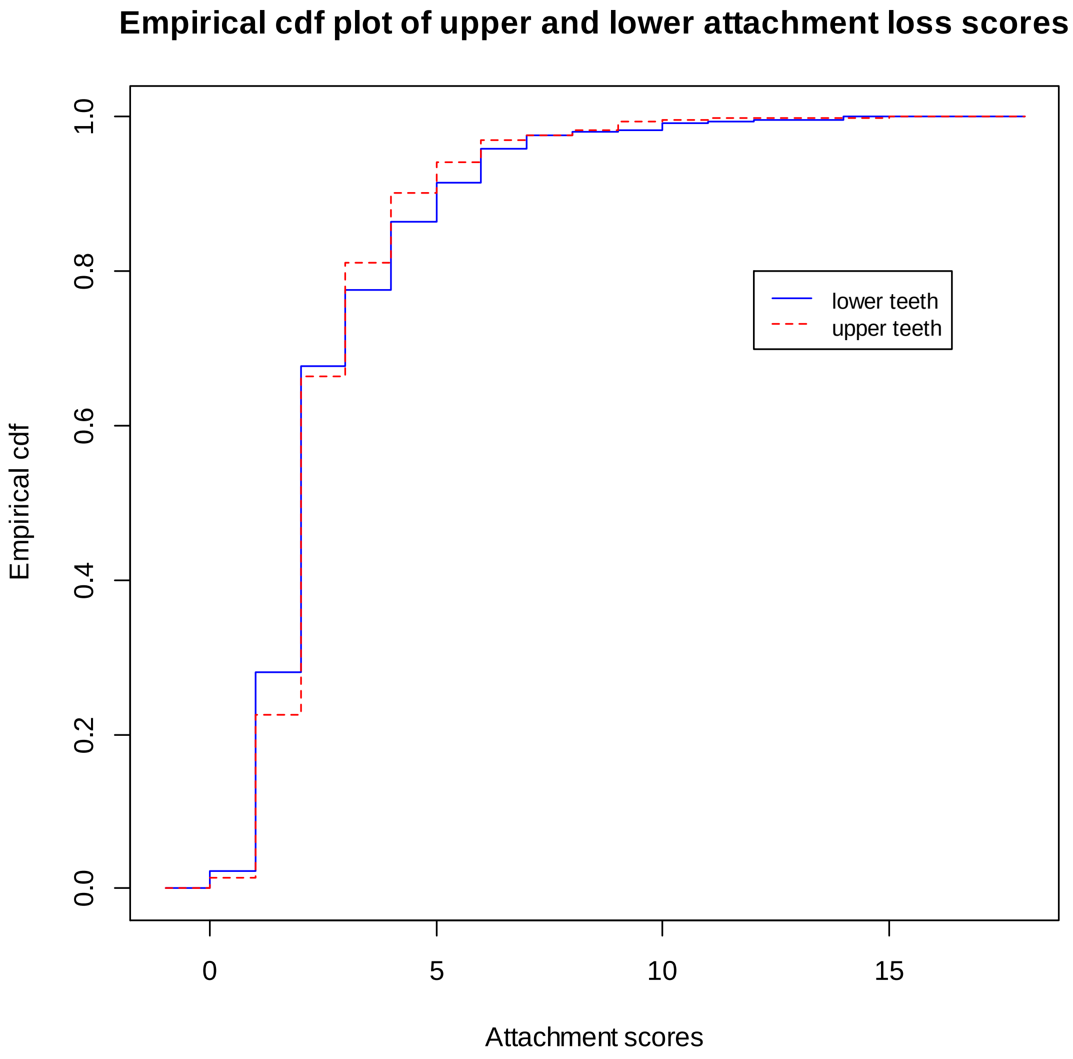


Figure 2.4. Plot of empirical cumulative distribution functions for lower and upper teeth attachment loss scores.



2.5 Technical Details

Calculations for deriving the expression of T from Section 2.1.1

To handle ties in the data, ranks are defined as

$$R_i^* = 1 + \frac{1}{2}\Big\{\sum_{j\neq i} I(X_j^* \leq X_i^*) + \sum_{j\neq i} I(X_j^* < X_i^*)\Big\}.$$

The expression for T can be found through the following steps :

$$E(G_i^* R_i^* \,|\, X, G) = E\left\{ \frac{G_i^*}{2}\Big\{\sum_{j\neq i} I(X_j^* \leq X_i^*) + \sum_{j\neq i} I(X_j^* < X_i^*)\Big\} + G_i^* \,\middle|\, X, G \right\}$$

$$= \frac{1}{2}\sum_{j\neq i} E\Big\{G_i^* I(X_j^* \leq X_i^*) \mid X,\, G\Big\} + \frac{1}{2}\sum_{j\neq i} E\Big\{G_i^* I(X_j^* < X_i^*) \mid X,\, G\Big\} + E\{G_i^* \,|\, X, G\}.$$

Now, $E\Big\{G_i^* I(X_j^* \leq X_i^*) \mid X,\, G\Big\} = E\Big[E\Big\{G_i^* I(X_j^* \leq X_i^*) \,|\, G_i^*, X_i^*\Big\} \mid X,\, G\Big]$

$$= E\Big\{G_i^* F_j(X_i^*) \mid X,\, G\Big\},$$

where $\quad F_j(x) = \frac{1}{2N_{j1}}\sum_{h=1}^{N_{j1}} I(X_{jh}^{(1)} \leq x) + \frac{1}{2N_{j0}}\sum_{h'=1}^{N_{j0}} I(X_{jh'}^{(0)} \leq x).$

Thus, $E(G_i^* R_i^* \,|\, X, G)$

$$= \tfrac{1}{2}\textstyle\sum_{j\neq i} E\big(G_i^* F_j(X_i^*) \mid X,\, G\big) + \tfrac{1}{2}\textstyle\sum_{j\neq i} E\big(G_i^* F_j(X_i^* -) \mid X,\, G\big) + \tfrac{1}{2}$$

$$= \frac{1}{2}\sum_{j\neq i}\left(\sum_{k=1}^{N_{i1}} \frac{F_j(X_{ik}^{(1)}) + F_j(\ X_{ik}^{(1)} -)}{2N_{i1}}\right) + \frac{1}{2}.$$

Finally,

$$T = E(\mathrm{S}^* | X, G) = E\left(\sum_{i=1}^{M} G_i^* R_i^* \,|\, X, G\right) = \sum_{i=1}^{M}\sum_{k=1}^{N_{i1}} \frac{1}{2N_{i1}}\Big[1 + \frac{1}{2}\sum_{j\neq i}\{F_j(X_{ik}^{(1)}) + F_j(\ X_{ik}^{(1)} -)\}\Big].$$

An outline of the proof of Theorem 1 from Section 2.1.1

Recall that

$$T=\sum_{i=1}^{M}\sum_{k=1}^{N_{i1}}\frac{1}{2N_{i1}}\left[1+\frac{1}{2}\sum_{j\neq i}\{F_j(X_{ik}^{(1)})+F_j(\ X_{ik}^{(1)}-)\}\right].$$

The summands of the Hájek projection of T under H_0 is given by $T_i=E(T|\mathbb{V}_i)$.

To obtain an expression for T_i under H_0, we note that

For $i\neq j$,

$$E\left[\frac{1}{N_{_{i1}}}\{F_j(X_{ik}^{(1)})+F_j(X_{ik}^{(1)}-)\}|\,\mathbb{V}_i\right]=\frac{1}{N_{_{i1}}}\left[\mathcal{F}(X_{ik}^{(1)})+\mathcal{F}(X_{ik}^{(1)}-)\right].$$

For $i\neq j$,

$$E\left[\frac{1}{4N_{_{j1}}}\sum\nolimits_{k=1}^{N_{j1}}\{F_i(X_{jk}^{(1)})+F_i(X_{jk}^{(1)}-)\}\middle|\mathbb{V}_i\right]$$

$$=E\left[\frac{1}{2N_{i1}}\sum\nolimits_{h=1}^{N_{i1}}\frac{1}{4N_{j1}}\sum\nolimits_{k=1}^{N_{j1}}\{I(X_{ih}^{(1)}\leq X_{jk}^{(1)})+I(X_{ih}^{(1)}<X_{jk}^{(1)})\}\right.$$

$$\left.+\frac{1}{2N_{i0}}\sum\nolimits_{h'=1}^{N_{i0}}\frac{1}{4N_{j1}}\sum\nolimits_{k=1}^{N_{j1}}\{I(X_{ih'}^{(0)}\leq X_{jk}^{(1)})+I(X_{ih'}^{(0)}<X_{jk}^{(1)})\}\middle|\mathbb{V}_i\right]$$

$$=\frac{1}{2N_{i1}}\sum\nolimits_{h=1}^{N_{i1}}H(X_{ih}^{(1)})+\frac{1}{2N_{i0}}\sum\nolimits_{h'=1}^{N_{i0}}H(X_{ih'}^{(0)}),$$

where $H(a)=E\left[\sum_{k=1}^{N_{j1}}\frac{I\left(a\leq X_{jk}^{(1)}\right)+I\left(a<X_{jk}^{(1)}\right)}{4N_{j1}}\right]$.

Using the above expressions it can be shown that T_i= W_i+c_i, where c_i contains terms independent of $i,$ and

$$W_i=\ \frac{1}{2}+\frac{1}{2}.\frac{1}{2N_{i1}}.(M-1)\sum_{k=1}^{N_{i1}}\left(\mathcal{F}(X_{ik}^{(1)})+\mathcal{F}(X_{ik}^{(1)}-)\right)$$

$$+ (M-1)\left[\frac{1}{2N_{i1}}\sum_{h=1}^{N_{i1}} H(X_{ih}^{(1)}) + \frac{1}{2n_{i0}}\sum_{h'=1}^{N_{i1}} H(X_{ih'}^{(0)})\right].$$

Now that W_i and hence T_i are independent random variables , we have $\sum_{i=1}^{M} T_i$ as a sum of independent random variables. But, T is very close to a U-statistic and through linearization we have obtained the projection of T as $\sum_{i=1}^{M} T_i$, the sum of M independent random variables. Thus, we can apply the Lindeberg CLT (assuming that the sufficient conditions hold) to establish the asymptotic normality of T under H_0.

Extension to m ($>$ 2) Groups in Each of the M clusters

Suppose, instead of two groups, we have m groups in each of the M clusters. An observation within any cluster may belong to any of the m possible groups. A null hypothesis of interest would be that the marginal distributions ($\mathcal{F}_3$) are same in all the groups. Here $G_{ik} = \ell$ represents that the group membership of the k^{th} individual of the i^{th} cluster is ℓ. Also, $\{X_{i1}^{(\ell)}, X_{i2}^{(\ell)}, \cdots, X_{iN_{i\ell}}^{(\ell)}\}$ represents the set of observations for individuals with group membership l in the i^{th} cluster, where $1 \leq \ell \leq m$.

For $m = 2$, the test statistic from Section 2.1.1 is given by $T = \sum_{i=1}^{M}\sum_{k=1}^{N_{i2}}(2N_{i2})^{-1}[1 + \frac{1}{2}\sum_{j \neq i}\{F_j(X_{ik}^{(2)}) + F_j(X_{ik}^{(2)}-)\}] = T^{12}$, say. When $m > 2$, for comparing the marginal distribution of the outcome from the ℓ^{th}group with that from the ℓ'^{th}group, where $1 \leq \ell^{'} = \ell - 1 < \ell \leq m$, a valid statistic is of the form $T^{\ell^{'}\ell} = \sum_{i=1}^{M}\sum_{k=1}^{N_{i\ell}}(2N_{i\ell})^{-1}[1 + \frac{1}{2}\sum_{j \neq i}\{F_j(X_{ik}^{(\ell)}) + F_j(X_{ik}^{(\ell)}-)\}]$. In this way, we can construct an $(m-1)$ dimensional vector statistic $\Delta = (T^{12}, T^{23}, \cdots, T^{(m-1)m})^T$. Finally, one can reject the null hypothesis of equality of m marginal distributions for large values of the test statistic

$$U_m = \{\Delta - E_{H_0}(\Delta)\}^T \widehat{\Sigma}^{-1}\{\Delta - E_{H_0}(\Delta)\}.$$

Here $\widehat{\Sigma}$ is the estimated variance-covariance matrix of order $(m-1)$ comprising of the jackknife estimates of the variances and covariance entries $\widehat{V}(T^{\ell^{'}\ell}, T^{r^{'}r})$, $1 \leq \ell, r \leq m$. The jackknife estimates of the variances and covariances are computed through the same approach mentioned in the variance estimation of Section 2.1.1. Therefore, we have

$$\widehat{V}(T^{\ell'\ell}, T^{r'r}) = \frac{M}{(M-1)}\sum_{i=1}^{M}(T_i^{*\ell'\ell} - \overline{T}^{*\ell'\ell})(T_i^{*r'r} - \overline{T}^{*r'r}),$$

where $1 \leq \ell' = \ell - 1 < \ell \leq m$, $1 \leq r' = r - 1 < r \leq m$, $\ell \leq r$, $T_i^{*\ell'\ell} = T^{\ell'\ell} - T_{-i}^{\ell'\ell}$, $T_i^{*r'r} = T^{r'r} - T_{-i}^{r'r}$.

Under the null hypothesis, $U_m \sim \chi^2_{m-1}$ asymptotically. The p-value for the test is computed as the probability that, under the null hypothesis, the statistic U_m exceeds its observed value. We would reject the null hypothesis at a 100α% level of significance if the p-value is less than α.

Calculations for deriving the expression of T from Section 2.1.2

Following are the detailed steps to derive the final expression for the statistic T from Section 2.1.2.

Recall

$$R_i^{*}= 1+ \frac{1}{2}\{\sum_{j\neq i} I(X_j^{*} \leq X_i^{*}) + \sum_{j\neq i} I(X_j^{*} < X_i^{*})\}.$$

Now $E(G_i^{*}R_i^{*} | X, G) =$

$$E\left[\frac{G_i^{*}}{2}\left\{\sum_{j\neq i} I(X_j^{*} \leq X_i^{*}) + \sum_{j\neq i} I(X_j^{*} < X_i^{*})\right\} + G_i^{*} \mid X,\, G\right]$$

$$= \frac{1}{2}\sum_{j\neq i} E\big(G_i^{*}I(X_j^{*} \leq X_i^{*}) \mid X,\, G\big) + \frac{1}{2}\sum_{j\neq i} E\big(G_i^{*}I(X_j^{*} < X_i^{*}) \mid X,\, G\big) + E(G_i^{*}| X, G)$$

$$= \frac{1}{2}\sum_{j\neq i} E\big(G_i^{*}I(X_j^{*} \leq X_i^{*}) \mid X,\, G\big) + \frac{1}{2}\sum_{j\neq i} E\big(G_i^{*}I(X_j^{*} < X_i^{*}) \mid X,\, G\big)$$

$$+ \left\{\frac{I(N_{i1} > 0, N_{i0} > 0)}{2} + I(N_{i0} = 0)\right\}.$$

Next,

$$E\big(G_i^{*}I(X_j^{*} \leq X_i^{*}) \mid X,\, G\big) = E\big(E\big(G_i^{*}I(X_j^{*} \leq X_i^{*}) | G_i^{*}, X_i^{*}\big) \mid X,\, G\big)$$

$$= E\big(G_i^{*}F_j^{'}(X_i^{*}) \mid X,\, G\big).$$

Here, $F_j^{'}(x) =$

$$\left[\frac{1}{2N_{j1}}\sum\nolimits_{h=1}^{N_{j1}} I(X_{jh}^{(1)} \leq x) + \frac{1}{2N_{j0}}\sum\nolimits_{h'=1}^{N_{j0}} I(X_{jh'}^{(0)} \leq x)\right]\big(I(N_{j1} > 0, N_{j0} > 0)\big)$$

$$+\Big(1 - I(N_{j1} > 0, N_{j0} > 0)\Big)\left[\frac{1}{N_j}.\sum_{h=1}^{N_j} I(X_{jh} \leq x)\right] \;.$$

Thus, $E(G_i^* R_i^* \,|\, X, G)$

$$= \frac{1}{2}\sum_{j \neq i} E\big(G_i^* F_j^{'}(X_i^*) \mid X,\, G\big) + \frac{1}{2}\sum_{j \neq i} E\big(G_i^* F_j^{'}(X_i^* -) \mid X,\, G\big)$$

$$+\left\{\frac{I(N_{i1} > 0, N_{i0} > 0)}{2} + I(N_{i0} = 0)\right\}.$$

$$= \frac{1}{2}\sum_{j \neq i}\left(\sum_{k=1}^{N_{i1}} \frac{F_j^{'}(X_{ik}^{(1)}) + F_j^{'}(\;X_{ik}^{(1)} -\,)}{2N_{i1}} (I(N_{i1} > 0, N_{i0} > 0) + 2N_{i1}.I(N_{i0} = 0))\right)$$

$$+\left\{\frac{I(N_{i1} > 0, N_{i0} > 0)}{2} + I(N_{i0} = 0)\right\}.$$

Finally, $T = E(\mathrm{S}^* | X, G) = E\left(\sum\limits_{i=1}^{M} G_i^* R_i^* \,|\, X, G\right)$

$$= \;\sum_{i=1}^{M}\;\sum_{k=1}^{N_{i1}} \frac{1}{2N_{i1}}\Bigg[(I(N_{i1} > 0, N_{i0} > 0) + 2I(N_{i0} = 0))$$

$$+ \frac{(I(N_{i1} > 0, N_{i0} > 0) + 2N_{i1}.I(N_{i0} = 0))}{2}\sum_{j \neq i}\{F_j^{'}(X_{ik}^{(1)}) + F_j^{'}(\;X_{ik}^{(1)} -\,)\}\Bigg].$$

R Code

```
##############################################

R Code for test statistic developed in Section 2.1

##############################################

data= cbind(Cluster,X, grp)

rn<-function(dv){

ik=dv[1]

x=dv[2]

ds1=data[data[,3]==1,]
vs1=(kh==2)*(ds1[,2]<x)+(kh==1)*(ds1[,2]<=x)
sl1=aggregate(vs1,list(ds1[,1]),mean)[,2]

ds2=data[data[,3]==0,]
vs2=(kh==2)*(ds2[,2]<x)+(kh==1)*(ds2[,2]<=x)
sl2=aggregate(vs2,list(ds2[,1]),mean)[,2]

fg=(sl1+sl2)/2
fg[ik]=0
return(fg)
}

rst<-function(il){
        ly=sum(mat[-which(dw[,1]==il),-il])
#ly=apply(mat[-which(dw[,1]==il),-il],1,sum)
        return(ly)
}
m=length(unique(data[,1]))
dw=data[(data[,3]==1),]
ns=(dw[,1])
nv=as.vector(table(ns)[match(ns,names(table(ns)))])

kh=1
mat=t(apply(cbind(dw[,1:2]),1,rn))/nv
 vf1=apply(cbind(seq(1,m)),1,rst)
sFs1=sum(mat)
```

```
kh=2

mat=t(apply(cbind(dw[,1:2]),1,rn))/nv

vf2=apply(cbind(seq(1,m)),1,rst)

sFs2=sum(mat)


v1=((sFs1+sFs2)/4)+(m/2)

vd= ((vf1+vf2)/4)+(m-1)/2

    h=1

T<- v1

E.T<- 0.25*m*(m+1)

test=(m/m^h)*v1-((m-1)/(m-1)^h)*vd

v.test=var(test)

v_hat=(((m^h)^2)/(m-1))*v.test

v.hat=ifelse(v_hat==0,0.00000001,v_hat)

Z<- (T-E.T)/sqrt(v.hat)

p.value<- 2*pnorm(abs(Z), lower.tail=FALSE)
```

CHAPTER 3

TEMPORAL PREDICTION OF FUTURE STATE OCCUPATION IN A MULTISTATE MODEL FROM HIGH DIMENSIONAL BASELINE COVARIATES VIA PSEUDO-VALUE REGRESSION

3.1. Background of the Methods

3.1.1 *Pseudo-values and their Application in Regression Modeling*

The pseudo-value approach was first obtained for the ‘leave-one-out’ jackknife resampling technique, with the initial purpose being studying the bias and standard error of an estimator. The idea behind the construction of pseudo-values is easily comprehensible when the estimator is linear. If $\widehat{\theta}$ is an estimator of a parameter of interest θ based on a random sample of size n, and if $\widehat{\theta}^{(-i)}$ is the estimate of θ obtained by deleting the i^{th} observation from the original sample, then the i^{th} pseudo-value is defined as $\eta_i := n\ \widehat{\theta} - (n-1)\ \widehat{\theta}^{(-i)}$, where $i \in \{1, 2, \cdots, n\}$. Andersen et al. (2003) proposed the use of these pseudo-values in the context of regression modeling via generalized linear models. Suppose data consists of n pairs of independent and identically distributed pairs (X_i, Z_i), $1 \leq i \leq n$, of response X and covariates Z, and we are interested in estimating $\theta(Z) = E(f(X)|Z)$, for some known function f. Starting with an asymptotically linear and unbiased estimator $\widehat{\theta}$ of the corresponding marginal parameter $\theta = E(f(X))$, Andersen et al. (2003) proposed that one can regress the corresponding pseudo-values η_i on Z_i to obtain an estimator of the regression function $\theta(Z)$. In this article, we let f be the indicator function $I(S(t) = h)$ which denotes whether an individual is at state h at time t. With this choice, $P_h(t)$, the occupation probability of a certain state h at a given time $t,$ becomes the parameter θ of interest.

3.1.2 *Estimation and Regression of State Occupation Probability in Multistate Models*

Aalen and Johansen (1978) proposed a non-parametric estimator of the state occupation probability in multistate models with censored outcomes. They showed that, under independent censoring, $\widehat{p}_h(t)$, the Aalen-Johansen estimate of occupation probability of state h at time t, is consistent for estimating the true occupation probability $P_h(t)$ if the underlying multistate process is Markov. Later on, Datta and Satten (2001) showed that even if the underlying process is non-Markov, the Aalen-Johansen estimator of state occupation probability remains consistent. The Aalen-Johansen estimator can be thought of as a generalization of the Kaplan-Meier estimator of survival probability in a two-state survival framework. A detailed description of the Aalen-Johansen estimator is provided in the Technical Details.

If one wants to predict the occupation probability of a typical state h at some future time t through the pseudo-value based regression approach as discussed earlier then the pseudo-values of state occupation probability of state h at time t can be generated as

$$\widehat{p}_{h;i}(t) = n\widehat{p}_h(t) - (n-1)\widehat{p}_h^{(-i)}(t),\, i = 1, 2, \cdots, n,$$

where $\widehat{p}_h^{(-i)}(t)$ is the Aalen-Johansen estimate of occupation probability of state h at time t calculated after removing the individual i from the data. Now, we can regress these pseudo-values on available covariates through a linear model or a generalized linear model as discussed before. But in case the number of covariates (p) available for each individual exceed the total number of individuals (n) under study, i.e. $n < \, < p$, the standard linear or generalized linear models fail and we have to resort to one of the high dimensional regression techniques. Of the different high dimensional regression techniques, we have considered latent factor regression such as Partial Least Squares, Sparse Partial Least Squares, and penalized regression methods such as LASSO, Elastic Net, and Adaptive LASSO in conjunction with pseudo-values in this article. Details on these high dimensional methods, including their computational steps and important features, can be found in the Technical Details section.

3.2. Simulation Studies

We now describe a number of simulation settings. In the first setting, we have a multistate model framework where we compare the performances of different high dimensional regression methods such as PLS, SPLS, LASSO, Adaptive LASSO (AdLasso), and Elastic Net (ENET), based on pseudo-values. In the second setting, we have a survival (two-state) model where we compare the performance of the Cox model based LASSO regression with that of the PV based high dimensional regression techniques when the underlying true model is non-Cox type.

3.2.1 *Simulation designs for a multistate model*

We generate an irreversible three-stage illness-death model with censored outcomes. The three states in this illness-death model are the 'disease-free' state, 'ill' or 'disease' state, and the 'death' state, which are indexed as states 1, 2, and 3, respectively. It is termed 'irreversible' because in this model once an individual leaves a particular state at some point of time, it cannot come back to that state at any later time point. Under this model transition is possible from state 1 to both the states 2 and 3, while once in state 2, transition is possible only to state 3. State 3 (death state) is an absorbing state, i.e., once an individual enters state 3, no more transitions are possible. Moreover, we assume that all the individuals in the study start from the initial 'disease-free' state. In our simulation study, we have generated all the state-to-state transition times of a typical individual from accelerated failure time (AFT) models based on the available covariate information of that individual. In general for a typical individual i, the transition time from the state h to the state k, T_{ihk} (say), is such that the $log(T_{ihk})$ is generated from a linear model based on the available covariates $Z_{i1}, Z_{i2}, \cdots, Z_{iq}$. For our illness death model, we generate the transition times as follows:

$$log(T_{i12}) = \sum_{j=1}^{q} \beta_j Z_{ij} + \epsilon_{i12}\,, \;\; log(T_{i13}) = \sum_{j=1}^{q} \gamma_j Z_{ij} + \epsilon_{i13}\,,$$

and

$$log(T_{i23}) = \sum_{j=1}^{q} \gamma_j Z_{ij} + \epsilon_{i13}\,, \text{ provided } T_{23} \geq T_{12},$$

where ϵ_{12} and ϵ_{13} are the error components, and q is the number of available covariates, and $i = 1, 2, \cdots, n.$ The transition times are generated in such a way that the resulting multistate illness-death process falls under the category of a Markov process. The different entities involved in the above simulation model are chosen in the following way:

(i) Sample size $= n = 200.$

(ii) Number of covariates (covariate dimension) $= q = 10, 000.$

(iii) Regression coeffcient parameters: The parameters concerning the regression

coeffcients in the above mentioned regression models are chosen as one of the two following combinations :

(a) $\beta_j = \begin{cases} 1, \text{ if } 1 \leq j \leq 50 \\ 0, \text{ otherwise} \end{cases}$ $, \gamma_j = \frac{1}{j}, \quad 1 \leq j \leq q.$

(b) $\beta_j = 1, \gamma_j = \frac{1}{j},$ $\qquad 1 \leq j \leq q.$

Case (a) corresponds to the situation where only a few (0.5%) of the total covariates actually contribute to the time of transition of an individual from state 1 to state 2. So this can be thought of as a sparse regression model for transition into the disease state. In case (b), all the available covariates contribute to the transition time from state 1 to state 2, which implies a non-sparse (dense) regression model for transition to the disease state. In both the cases, the number of covariates contributing to the transition to state 3 is neither too large nor too small. Also, we denote the regression coefficient vectors $(\beta_1, \beta_2, \cdots, \beta_q)^T$ and $(\gamma_1, \gamma_2, \cdots, \gamma_q)^T$ as β and γ respectively.

(iv) Design Matrix: Let Z be the design matrix such that covariate vector for the i^{th} individual, namely $Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{iq})^T$, defines the i^{th} row of Z. Then the rows of the matrix Z are generated from a multivariate normal distribution with zero mean vector and variance covariance matrix Σ_Z. The choice of Σ_Z is taken to be a diagonal matrix with all diagonal elements as 1, and all off-diagonal elements as 0.

(v) Errors : We generate both the errors ϵ_{12} and ϵ_{13} from a normal distribution with mean 0 and variance $r\sigma^2$. Here $\sigma^2 = max(\underline{\beta}^T \Sigma \underline{\beta}, \underline{\gamma}^T \Sigma \underline{\gamma})$, where $\underline{\beta}$ and $\underline{\gamma}$ are normalized versions of regression coefficient vectors β and γ respectively, and r is a constant factor controlling the noise-to-signal ratio (NSR) of the simulated regression model.

(vi) Censoring: We consider right censoring in the simulated irreversible illness death model. The censoring time for the i^{th} individual ($C_i,$ say) at each of the states 1 and 2 is generated from a lognormal distribution, such that $log(C_i) \sim N(c_0, \sigma^2)$ independently for $i = 1, 2, \cdots, n$. Here c_0 is determined by the overall censoring rate. We consider three different choices for the censoring rate, namely, 0% (no censoring), 35% (moderate censoring), and 80% (heavy censoring).

In this simulation study our main aim is to estimate the occupation probability of a given state (out of the three states of the illness-death model) at a given time point by fitting a regression model based on the huge covariate set. So, to directly predict the future state occupation based on the huge number of available covariates, we start with the Aalen-Johansen estimator as the marginal estimator for state occupation probability as discussed in Section 2.2. If $\widehat{p}_h(t)$ is the Aalen-Johansen estimate of occupation probability of state h at time t calculated from the full data, then the pseudo-value corresponding to the i^{th} individual $\widehat{p}_{h;i}(t)$ is given by

$$\widehat{p}_{h;i}(t) = n\widehat{p}_h(t) - (n-1)\widehat{p}_h^{(-i)}(t),\ i = 1, 2, \cdots, n.$$

Here $\widehat{p}_h^{(-i)}(t)$ is the Aalen-Johansen estimate of the same state occupation probability calculated after leaving the i^{th} individual from the data. In case of irreversible illness-death model the disease state (state 2) is often of primary importance. As such, in our simulation study we have directly modeled the state occupation probability of state 2 at specific time points with the n pseudo-values $\widehat{p}_{h;i}(t)$ as responses where h is 2. Now just like in microarray-based studies, the covariate dimension (q) is very large compared to the sample size (n) in our simulation settings. So, to carry out proper prediction we use different high dimensional regression methods discussed in Section 2.3 such as PLS, SPLS, LASSO, Adaptive LASSO, and Elastic net. To get a complete picture on the performances of these high dimensional regression methods we vary the number of PLS or SPLS terms (latent factors) as well as the number of LASSO, AdLasso, and ENET steps.

Performance measure: To evaluate the predictive performances of the PV based high dimensional regression methods, we derive the theoretical (true) state occupation probability at a given time conditional on the covariates $Z_1, Z_2, \cdots, Z_q$, for the irreversible illness-death model described in our simulation settings. If $P_h(t;Z_i)$ denote the true occupation probability of the i^{th} individual (conditional on covariate vector Z_i) at state h at time t, and $\Phi(x)$ is standard normal distribution function at x, then it can be shown that $P_2(t;Z_i) = \left(1 - \Phi\left(\frac{log(t)-Z_i\gamma}{\sigma}\right)\right).\left(\Phi\left(\frac{log(t)-Z_i\beta}{\sigma}\right)\right)$ for the Markov setting of our simulated illness-death model. Detailed steps for deriving $P_2(t;Z_i)$ can be found in the Technical Details. If $\widehat{P}_{2;i}(t;Z_i)$ denote the estimated value of the state 2 occupation probability at time t for the i^{th} individual using a PV based regression method, then a measure of the predictive power of that PV based regression method can be given by the mean relative error of estimation

$$\mathit{MREE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\widehat{P}_{2;i}(t;Z_i) - P_2(t;Z_i)\right|}{P_2(t;Z_i)},\ i = 1, 2, \cdots, n.$$

Here lower values of *MREE* indicate better prediction power of the corresponding regression method. In addition, to compare the performances of high dimensional regression methods with that of a 'no-covariate' model, i.e., a model based on the marginal probabilities without any covariate information, we calculate the above measures for a no covariate model, where $\widehat{P}_{2;i}(t;Z_i)$ is replaced by the marginal Aalen-Johansen estimate $\widehat{p}_2(t)$ (ignoring the covariate information Z_i) for state 2 for all values of $i = 1, 2, \cdots, n$. For most parts of our simulation study, we choose the time point t as the median of all the first transition times obtained from the complete data on n individuals. We calculated all the *MREE* values of the pseudo-value based PLS, SPLS, LASSO, AdLasso, and ENET regression methods as well as that of the 'no-covariate' model by averaging over 50 independent Monte-Carlo runs of the previously described data set.

Results: Among the variety of simulation settings we have considered so far, the most important factor turns out to be the choice of the regression coefficients. As we are interested in the occupation probabilities of state 2 of the illness-death model, we have considered, as mentioned before, two different choices for the regression coefficient vector (β) while generating the transition times from state 1 to state 2. So here we present the simulation results separately for each choice of β, and then compare the results between the two choice of β.

First we consider the case (a) where $\beta_j = 1$ if $1 \leq j \leq 50$, otherwise $\beta_j = 0$ if $50 \leq j \leq 10,000$, and $\gamma_j = \frac{1}{j}$ for $1 \leq j \leq 10,000$. In such a sparse regression scenario where only 0.5% of the total available covariates contribute to the transition to state 2, we compute the MREE values under different high dimensional regression methods based on pseudo-values. Figure 3.1 shows the predictive performances $($*MREE* values$)$ of different PV based regression methods for a wide range of regression steps. We compute the *MREE* values assuming different number of latent factors in PLS and SPLS regression, where the threshold tuning parameter (see Technical Details) for a fixed number of latent components in SPLS regression is obtained through crossvalidation. Similarly for LASSO, ENET, and AdLasso, we compute the *MREE* values for the full solution path of the complexity parameter corresponding to the L_1 penalty (refer Technical Details) which corresponds to the different steps in LASSO$/$ENET$/$AdLasso regression. In addition, for ENET regression we consider four choices for the elastic net mixing parameter α (described in Technical Details), namely 0.2, 0.4, 0.6, and 0.8, while for AdLasso we take the ridge regression estimate of the regression coefficient corresponding to minimum cross-validated error as the initial consistent estimator along with the three choices of the weight tuning parameter γ, namely $0.5,\ 1,$ and 2 (as mentioned in Technical Details). Figure 3.5 compares the *MREE* values of different regression methods based on pseudo-values for varying rates of censoring present in the data. Table 3.1 presents the optimal *MREE* values for the different regression methods based on pseudo-values as well as that of a 'no-covariate' model under different censoring rates of 0%, 30%, and 80%. From Table 3.1 we find that the optimal (minimum) *MREE* value for each of PLS, SPLS, LASSO, AdLasso, and ENET is substantially less than the *MREE* value of the 'no-covariate' model. This implies that indeed the PV based high dimensional regression methods are effective as compared to a marginal model in predicting state occupation. This becomes even more clear from Figure 3.5 which shows that even the suboptimal *MREE* values for the PV based regression techniques are lower than that of the no-covariate model in Table 3.1. Also from Figure 3.5 and Table 3.1, we see that among the penalized regression methods, namely, LASSO, AdLasso, and ENET, the LASSO performs the best for all types of censoring rates considered. But even then, the optimal values of PLS and SPLS regression is less than that of LASSO, with PLS regression emerging out to be the best in terms of having minimum *MREE* values overall. Also, we see that with the increase in the censoring rate in the data, the *MREE* values tend to increase, albeit not by a large margin. All the results mentioned so far were carried out under a noise-to-signal ratio (NSR) of 0.01. It may be interesting to see how do these PV based regression methods perform when the NSR is increased by a large extent. For this we obtained the *MREE* values of PLS, SPLS, LASSO, and ENET regression under 80% censoring rate with NSR as high as 1.0. Figure 3.1 compares the performances of PLS, SPLS, LASSO, and ENET under the two different NSR values of 0.01 and 1.0, while Table 3.4 shows the optimal *MREE* values of different regression methods based on pseudo-values as well as that of the no-covariate model under the high NSR of 1.0. Interestingly, with the increase in the NSR the performances of the PV based regression methods tend to deteriorate to such an extent that it is difficult to distinguish them from a no-covariate model, although the optimal values of the PV based methods are marginally better (lower) than that of the no-covariate model. Also, we see that the difference between the PLS-type and the LASSO-type regression methods disappears under high NSR, with ENET with mixing parameter 0.2 having the minimum *MREE* value. Since the marginal estimator of state occupation probability is a function of time, one may be interested in observing how the *MREE* for the PV based regression methods behave as a function of time. In Figure 3.6 we plot the optimal *MREE* values of the PLS and LASSO regression based on pseudo-values as a function of the time at which the underlying Aalen-Johansen estimates are calculated. Figure 3.6 shows that given a fixed time interval, the *MREE* for state 2 probability has lower values at some point in the later half of the interval as the number of individuals occupying state 2 is expected to maximum near this time point .

Next we consider the case (b) where $\beta_j = 1, \gamma_j = \frac{1}{j}$ for $1 \leq j \leq 10,000$. This is non-sparse regression scenario for transition to state 2 where all the covariates contribute to the transition time from state 1 to state 2. Like in case (a), we compute the *MREE* values for PLS, SPLS, LASSO, AdLasso, and ENET based on pseudo-values for a wide range of regression steps in case (b) as well. Figure 3.7 displays the *MREE* values for different steps of PLS, SPLS , LASSO, AdLasso, and ENET regression under three different censoring rates of 0%, 30%, and 80% while the NSR is 0.001. Table 3.2 summarizes the optimal *MREE* values of these PV based regression techniques as well as that of a no-covariate model for the three different censoring rates. Table 3.2 shows that all the PV based regression methods perform substantially better than the no-covariate model. Moreover, from Figure 3.7 and Table 3.2, we also find that PLS regression has the minimum overall *MREE* values among all the PV based regression techniques considered here, which is similar to the results found in case (a) where only a handful of covariate contribute to the transition into state 2. But the most striking difference in case (b) is that the difference between optimal *MREE* values of PLS and LASSO increase to such an extent that the minimum *MREE* of PLS is around six times smaller than that of LASSO, whereas in case (a) the minimum *MREE* of PLS was only 1.7 times lower than that of LASSO. The minimum MREE of PLS is also seven times lower in case (b) than that in case (a), while the minimum *MREE* values of LASSO do not differ substantially between cases (a) and (b). This remarkable decrease in *MREE* values of PLS signifies the fact in case of non-sparse (dense) regression scenario where all, or at least the majority of the covariates contribute to the outcome of interest, the PLS method is vastly superior than the other high dimensional regression techniques in predicting state occupation. Next we investigate the predictive performances of the PV based methods in this simulation setting when we increase the NSR to 0.1. Figure 3.2 and Table 3.5 shows that as the NSR increases the performances of the PV based regression methods do get worse, though the optimal *MREE* values are still better than the *MREE* of the no-covariate model. But the wide difference between the optimal *MREE* values of PLS and LASSO appears to vanish in case of high NSR.

From the simulation results discussed so far we see that the pseudo-value based high dimensional regression methods are indeed good options for directly estimating future state occupation probability in presence of censored data with huge covariate dimension and small sample size. Overall the PLS regression performs better than the other high dimensional regression methods when we use pseudo-value based models in predicting state occupation at a given time. This superiority of PLS is more prominent in case of non-sparse regression scenario where most of available covariates contribute to the outcome of interest, whereas LASSO regression can be thought of as a potential alternative to PLS in case there is a large number of noise variables, i.e., only a very small minority of the total available covariates are actually contributing to the outcome of interest.

3.1.2 *Simulation designs for a survival model*

As indicated before, the widely used survival model is a two-state model $($state $1 =$ 'alive' and state 2 = 'dead'$)$ which can thought of as a special case of a general multistate model. Note that $S(t)$, the survival probability of an individual at time t, can be interpreted as $S(t) = P_1(t) = 1 - P_2(t)$, where $P_2(t)$ is the probability that individual is occupying state 2 at that time t. So the PV based regression method can be applied to directly model the survival probabilities based on large number of covariates and small sample sizes in presence of potential censoring. There exists a LASSO regression method based on Cox regression model in survival framework (Tibshirani, 1997). Note that, PLS regression can also be carried out under the assumption of an underlying Cox model, but the estimation of the model parameters failed due to the non-convergence of the algorithm for the simulation settings we considered. As such, it might be interesting to see how the PV based regression methods fare compared to the Cox model based LASSO method in survival prediction when the underlying survival model is not a Cox type model (i.e., the hazards are not proportional). For this, we engage PLS as well as LASSO, based on pseudo-values, as these two high dimensional methods performed near the top in our extensive simulation studies in Section 3.1. We choose the Kaplan-Meier estimator of survival probability as the marginal estimator and compute the pseudo-values based on this estimator in the same way as we do it for the Aalen-Johansen estimator of state occupation probability.

For carrying out this comparison we simulate a survival data with right censoring. The true event (transition from state 1 to state 2) times are generated via an AFT model based on the available covariates $Z_1, Z_2, \cdots, Z_q$. This results in a model that is not Cox-type. If T_i denote the true event time for the i^{th} individual, then we have $log(T_i) = \sum_{i=1}^{q} \beta_j Z_{ij} + \epsilon_i$, where ϵ_i is the error component, q is the covariate dimension, and $i = 1, 2, \cdots, n$. The entities involved in the model are chosen as follows:

(i) Sample size $= n = 100$.

(ii) Number of covariates (covariate dimension) $= q = 10,000$.

(iii) Regression coefficient parameters: The regression coefficient vector β, where $\beta = (\beta_1, \beta_2, \cdots, \beta_q)^T$, in the above mentioned regression model is chosen in one of the two following combinations :

(a) $\beta_j = \begin{cases} j \bmod 5, \text{ if } 1 \leq j \leq 100 \\ 0, \text{ otherwise} \end{cases} \quad 1 \leq j \leq q.$

(b) $\beta_j = 1$, $\gamma_j = \frac{1}{j}$, $\quad 1 \leq j \leq q.$

Case (a) corresponds to a highly sparse regression scenario where only a few (0.8%) of the total covariates actually contribute to the true event time. In case (b) the number of covariates contributing to the true event time is neither too large nor too small.

(iv) Design Matrix: Let Z be the design matrix such that covariate vector for the i^{th} individual, namely $Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{iq})^T$, defines the i^{th} row of Z. Then the rows of the matrix Z are generated from a multivariate normal distribution with zero mean vector and variance covariance matrix Σ_Z. The choice of Σ_Z is taken to be a diagonal matrix with all diagonal elements as 1, and all off-diagonal elements as 0.

(v) Errors : We generate the errors ϵ_{12} from a normal distribution with mean 0 and variance $10\sigma^2$, where $\sigma^2 = \underline{\beta}^T \Sigma \underline{\beta}$, where $\underline{\beta}$ is the normalized version of regression coefficient vectors β.

(vi) Censoring: As mentioned before, we consider right censoring in the simulated survival model. If C_i be the right censoring time for the i^{th} individual $log(C_i) \sim N(c_0, \sigma^2)$ independently for $i = 1, 2, \cdots, n.$ Here c_0 is determined by the censoring rate. For this simulation we choose the censoring rate to be around 50%.

With the right censored data generated from the above simulation setting, we estimate the marginal survival probability at time t through the Kaplan-Meier estimator, and then calculate the pseudo-values of the survival probability for each of the n individuals. Now, with these pseudo values as responses we fit a regression model based on either LASSO or PLS regression technique, and predict the survival probabilities of the individuals from the fitted model. In addition, we separately fit a Cox proportional hazard model with LASSO-type (L_1) penalization on the simulated right censored data and again estimate the survival probabilities of all the individuals at time t using the fitted Cox-LASSO regression model with Breslow estimate of baseline survival.

Performance measure: In order to evaluate the predictive performances of the different high dimensional regression methods, we derive the theoretical survival probabilities (assuming no censoring) at a given time conditional on the covariates $Z_1, Z_2, \cdots, Z_q$, for the above mentioned simulation model and take these probabilities as benchmarks for evaluation, similar to the simulation study in Section 3.1. If $S(t;Z_i)$ denote the theoretical survival probability at time t of the i^{th} individual with covariate information Z_i, then for the above mentioned simulation setting, we have $S(t;Z_i) = \left(1 - \Phi\left(\frac{log(t) - Z_i\beta}{\sigma}\right)\right) = S_i(t)$ (say). In addition, if $\widehat{S}_i(t)$ denote the estimated survival probability of the i^{th} individual at time t, either from the pseudo-value based high dimensional regression or a Cox model based LASSO regression, then a measure of the (relative) error of estimation is given by

$$MREE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\widehat{S}_i(t;Z_i) - S_i(t;Z_i)\right|}{S_i(t;Z_i)}.$$

Results: Figure 3.3 displays the *MREE* values according to different components or regression steps of PLS and LASSO regression based on pseudo-values, as well as, the *MREE* values of different steps in Cox model based LASSO regression for the sparse regression scenario described in (a). The *MREE* values of the different pseudo-value based and Cox model based regression techniques for simulation scenario (b) are displayed according to different components or regression steps in Figure 3.8. The optimal (minimum) *MREE* values of these regression methods for both case (a) and case (b) are tabulated in Table 3.6. As evident from Table 3.6, in case of the highly sparse regression scenario (a), the minimum *MREE* value for the pseudo-value based LASSO is lower than that of a Cox model based LASSO. It is the other way around in case (b) when the underlying model is less sparse. In both cases, however, the pseudo-value based PLS regression has the least minimum *MREE* value amongst the three.

3.3. Applications

3.3.1 *Michigan Lung Cancer Data*

We demonstrate the use of pseudo-value based PLS and LASSO regression methods in predicting patient survival using a Michigan lung adenocarcinoma data set which was originally analyzed by Beer et al. (2002). The original data set had 7129 gene expressions for 86 lung tumor samples and 10 normal tissue samples. Genes with extremely low levels of expressions were excluded from the final data set. The remaining 4966 genes were used for dividing the 86 lung cancer patients into three clusters by hierarchical clustering. In the original study, Beer et al. (2002) found that these three clusters showed significant differences based on tumor stage and tumor differentiation. That study intended to investigate the relationships between clusters, cancer stages, gene differentiation and overall survival, details of which can be found in Beer et al. (2002).

In this article we use the 4966 gene expressions obtained from the original study along with the survival times, survival indicator, and other associated information of the 86 lung cancer patients to demonstrate the pseudo-value based prediction of patient survival in presence of a high dimensional covariate. From the full data of 86 lung cancer patients, we estimate the overall survival probability at a given time t (in months) by the Kaplan-Meier estimator. We predict the survival at time t of each of the 86 patients based on his or her gene expression profile, irrespective of the censoring status. For this we use the PV based PLS and LASSO regression where the pseudo-values are based on the Kaplan-Meier estimator as described in details in Section 3. The data under study has a high proportion (around 70%) of censored observations. But, unlike in simulation studies, we do not have the true (theoretical) survival probability at any given time for any of the patients. At a given time t, the only information we have is that whether a patient is alive and under study at that time, or is dead at some time before t, or is censored at some time before t. We use the survival status of the set of individuals who are alive at time t to tune the regression model parameters, while we use the survival status of the set of patients who are known to be dead by time t to check the predictive power of the optimally tuned regression models at time t. Thus, for choosing the optimal number of PLS components or optimal LASSO steps, we calculate the following data-based measure of mean absolute error of prediction

$$\mathit{MAEP} = \frac{1}{n_R(t)}\sum_{i=1}^{n}\delta_i(t)\left|\widehat{S}_i(t;Z_i) - 1\right|,$$

where $\delta_i(t) = 1$ if the i^{th} patient is alive and under study at time t, and $\delta_i(t) = 0$ otherwise, $n_R(t) = \sum_{i=1}^{n}\delta_i(t)$, and $\widehat{S}_i(t;Z_i)$ is the estimated survival probability at time t for the i^{th} patient using PV based regression. Note that *MAEP* measures average absolute error of fit for the state prediction amongst subjects who are still known to be alive at the time point under consideration. The regression model having the minimum *MAEP* value at a given time t is chosen as the optimal model. We have calculated the *MAEP* values for different LASSO steps and PLS components for a wide range of the values of time t. For the choice of t as 30 months, Figure 3.9 shows the *MAEP* values for different LASSO steps and PLS components. Here the minimum value of *MAEP* using LASSO is 0.03154 and it corresponds to 64 LASSO steps, while the minimum value of *MAEP* using PLS is 0.04770 corresponding to a PLS regression with 7 components. Next, we check how well the optimal PV based regression model at a given time point t can indicate the survival status at time t of the individuals who are already dead by the time t. For better interpretation of the estimated survival probability as an indictor of the survival status, we classify a typical individual i as 0 or 1 based on whether the estimated survival probability $\widehat{S}_i(t;Z_i)$ is less than 0.5 or not, respectively. We define $D_i(t) = 0$ if $\widehat{S}_i(t;Z_i) < 0.5$ and $D_i(t) = 1$ if $\widehat{S}_i(t;Z_i) \geq 0.5$, implying that the patient i is predicted to be more likely to be dead than alive at time t if $D_i(t) = 0$. In that case a measure of the misclassification rate for the set of patients already known to be dead at time t can be obtained as

$$\textit{MR} = \frac{1}{n_D(t)} \sum_{i=1}^{n} \alpha_i(t) D_i(t),$$

where $\alpha_i(t) = 1$ if the i^{th} patient is known to be dead by time t and $\alpha_i(t) = 0$ if the i^{th} patient is not known to be dead by time t, $n_D(t) = \sum_{i=1}^{n} \alpha_i(t)$. It is easy to see that $0 \leq$ *MR* ≤ 1, where 0 denotes the case of no misclassification while 1 represents the case of maximum misclassification. Interestingly, for the optimal PV based regression model at a given time t obtained by minimizing *MAEP*, the MR value turns out to be 0 implying perfect classification, and this is true for all the choices of t considered in our analyses.

This indicates that, indeed, the optimal regression model having the minimum *MAEP* value at a given time t, perfectly identifies the individuals who have died before time t.

The Michigan lung cancer data have information on tumor status of the individual under study. To be precise, the individuals have either stage 1 tumor or stage 3 tumor. There are 67 patients with stage 1 tumor whereas there are 19 patients with stage 2 tumor. One interesting question is whether there is any difference in the survival chances based on tumor status. For this we predict the survival at time t for each of the 86 patients through both the PLS and LASSO regression based on pseudo-values where the optimal number of PLS components and the optimal number of LASSO steps are obtained by the procedure described in the last paragraph. Then we classify the patients according to their tumor status, namely stage 1 and stage 3, and then take the average of the estimated probabilities in each of the two groups. We repeat this for different choices of t and plot these average survival probabilities of stage 1 and stage 3 tumors as a function of time as shown in Figure 3.10. It can be seen that at any time point t the average survival probability of a stage 3 tumor patient is much less than that of a stage 1 tumor patient. So tumor status does play a differentiating role in overall patient survival. Also, out of the 86 lung cancer patients 35 are male and 51 are female. Table 3.7 compares the average predicted survival probability of the male population with that of female population at different points of time. From Table 3.7 we see that there appears to be no significant difference between the average survival probability of the male and the female population. Due to the censoring present in the data the main challenge is to get survival information of the patients who have already been censored before the time point of interest. As mentioned before, the pseudo-value technique enables us to estimate the survival probability of any patient at any given time point irrespective whether the patient was censored before the time point of interest. Figure 3.4 shows the temporal estimated survival probability of a lung cancer patient who was censored at 28.3 months. Survival probabilities are estimated using both LASSO and PLS regression methods based on pseudo-values where the number of LASSO steps or PLS components is chosen in the same way of minimizing *MAEP*. We can see that the estimated survival probability of this patient goes on decreasing as time increases which is consistent with the general nature of survival function. A bootstrap based confidence interval (CI) can be obtained for the estimated survival probability of a patient at a given time. Table 3.3 shows bootstrap based 95% confidence intervals (based on 1000 bootstrap resamples) of the estimated survival probabilities at multiple time points for the patient censored at 28.3 months.

3.3.2 *ICGC Lung Adenocarcinoma Data*

We also showcase the use of the pseudo-value based high dimensional regression of state occupation probability using another Lung Adenocarcinoma data supplied by the International Cancer Genome Consortium (ICGC). This dataset contains the clinical information of lung cancer patients including their gender, age at diagnosis, age at the time of follow-up, vital (mortality) status and disease status at the time of follow-up, along with the expression values of 132 proteins for each these patients. Each subject is identified as ‘alive’ or ‘deceased’ at the time of follow up, and depending on whether the subject was cured from the cancer or the cancer was still in the progressive stage in his/her body at the time of follow-up, the disease status is labeled as ‘complete remission’ or ‘progression’, respectively. After removing subjects with missing information on mortality or disease status, we are left with 123 individuals who have been diagnosed with lung cancer at some point of their lives. For our illustration, we construct a multistate model representation of these data where the initial state 0 is the ‘alive’ state and the two absorbing states are: state 1 representing ‘dead while in progression’ and state 2 representing ‘dead while in complete remission’. Individuals who have been reported to be alive at the time of last follow-up, are treated as censored observations, leading to a very high rate of censoring (85%) in the data under study. In this analysis, our aim is to predict the occupation of state 1, ‘death while in progression’, at a future time based on the expression values of the 132 available proteins (covariates). For this purpose, we employ the pseudo-value based LASSO and PLS regression, where the pseudo-values are calculated using the Aalen-Johansen estimator of occupation probability of state 1 at a given time t.

For our demonstration, we have chosen the time t to be 3 years which means that we are interested in assessing the probability that a lung cancer patient, given his/her protein expression profile, would end up dying from cancer within 3 years. The marginal estimate of the state 1 occupation probability at 3 years, using Aalen-Johansen estimator, turns out to be 0.171 from the given full data. As is the case with real data, we do not have information on the true covariate-specific occupation probability of state 1 at time t for any of the lung cancer patients. So, in order to choose the optimal number of PLS components or LASSO steps while regressing state 1 occupation probability based on the protein expression profile using pseudo-values, we take help of the following data-based measure of prediction error: *MAEP* $=\ (n_R(t))^{-1}\sum_{i=1}^{n}\delta_i(t)\left|\widehat{P}_i(t)-1\right|$, where $\delta_i(t)=1$ or 0 depending on whether the i^{th} patient is in state 1 at time t or not, respectively, while $n_R(t)=\sum_{i=1}^{n}\delta_i(t)$ and $\widehat{P}_i(t)$ is the estimated state 1 occupation probability at time t for the i^{th} patient using PV based regression. Note that the target *MAEP* value may not be 0 in this case, as it is expected that there will be some positive occupation probability associated with state 2 at time t even for individuals with $\delta_i(t)=1$. Nevertheless, this *MAEP* can still be used as a benchmark for selecting different PV based regression models. Figure 3.11 shows the *MAEP* values for different number of LASSO steps and different number of LASSO components. Using PV based LASSO, *MAEP* reaches its minimum (0.281) at 36 LASSO steps, while for PV based PLS, *MAEP* is minimized (0.299) with 2 components. Among other things, we find that the average estimated state 1 occupation probability at 3 years does not differ substantially between male and female patients using these optimally tuned LASSO and PLS regressions. We omit the details.

3.4. Discussion

The pseudo-value method allows direct prediction of future state occupation instead of indirect modeling through state-to-state transition hazards even when censoring is present. This is particularly useful when the main objective is to interpret the estimated probability that an individual is in a particular stage of a multistate disease process at a given time in terms of the covariate profile of that individual. When the dimension of the covariate profile (e.g. gene expression profile) exceeds the underlying sample size, one can use latent factor regressions such as PLS regression or penalized regression such as LASSO regression in conjunction with the pseudo-value approach. Through extensive simulation studies, we have seen that, among the various high dimensional regression techniques that we considered, overall PLS works the best with the pseudo-value based responses for predicting future state occupation or survival. In cases of underlying sparsity, where a majority of the available covariates are noise variables not contributing to the state occupation or survival probabilities, the pseudo-value based LASSO regression is a powerful alternative to PLS regression for prediction purposes. Even in case of simple survival (two-state) model with a huge covariate dimension, the pseudo-value based regression methods seem to work better than the Cox model based penalized regression for predicting survival when the proportional hazards assumption is violated.

We have demonstrated the use of pseudo-value based high dimensional regression using a lung cancer data set which had a high proportion of censored samples. We employed pseudo-value regression using PLS as well as LASSO regression in predicting patient survival at a given time. Overall, meaningful and consistent results were obtained on patient survival, e.g., differentiation based on tumor stages. We have also used a lung adenocarcinoma data provided by ICGC to show how the pseudo-value based high dimensional regression methods can be applied in presence of censoring to predict the cancer remission status at death at a given time based on their individual proteomic profiles using a multistate model framework.

This article is mainly motivated by the question of forecasting the state occupation or survival probability of a typical patient at some future time point based on its high dimensional covariate profile. Another interesting task can be finding out which of the available covariates (genes in case of gene expression profiles) are most significant in predicting state occupation or survival. But as state occupation probabilities are functions of time, one may have different optimal pseudo-value based regression models at different time points leading to the possibility of non-uniformity in the list of significant covariates over time. For example, some genes may turn out to be significant at two widely different time points but insignificant in the intermediate time points. Such results may be difficult to interpret from a biological perspective, and we plan to focus on this need of using the pseudo-value based high dimensional regression techniques for variable selection in future studies.

In our current work, as well as most of the past works based on pseudo-value regression, it has been assumed that the censoring mechanism is independent of the covariates under study. This assumption can be relaxed and a recent work in this direction (Binder, Gerds, and Andersen, 2014) suggests using a correctly specified regression model for the censoring time in a competing risk framework where the covariate dimension is smaller than the sample size. However, extension of this approach to high dimensional settings, especially in the presence of huge covariate dimensions consisting of omic expression profiles, is not straightforward as it would be challenging to specify the correct model for the censoring time based on the high dimensional genomic or proteomic covariates. So, the idea of including covariate dependent censoring in the temporal prediction of state occupation based on high dimensional baseline covariates needs separate attention in future studies.

Table 3.1. Minimum *MREE* values for different pseudo-value based regression as well as the *MREE* of a no-covariate model under different censoring rates for the sparse regression scenario (a) from the illness-death model in Section 3.2.1

Type of regression	Miminum *MREE*		
	0% censored	35% censored	80% censored
PLS	0.406965	0.411595	0.508458
LASSO	0.756544	0.776842	0.852360
ENET(0.8)	0.758218	0.777923	0.853899
ENET(0.6)	0.759626	0.778182	0.855966
ENET(0.4)	0.763583	0.782479	0.860280
ENET(0.2)	0.778851	0.796568	0.876685
AdLasso (0.5)	0.928864	0.940945	0.962294
AdLasso (1.0)	1.213042	1.234326	1.298804
AdLasso (2.0)	2.090289	2.113102	2.148114
No-covariate model	21.325611	21.338571	21.948334

Table 3.2. Minimum *MREE* values for different pseudo-value based regression as well as the *MREE* of a no-covariate model under different censoring rates for the non-sparse regression scenario (b) from the illness-death model in Section 3.2.1

Type of regression	Minimum *MREE*		
	0% censored	35% censored	80% censored
PLS	0.004346	0.017815	0.076456
LASSO	0.354781	0.396940	0.463035
ENET(0.8)	0.360327	0.399476	0.466124
ENET(0.6)	0.375991	0.404862	0.471734
ENET(0.4)	0.389865	0.414592	0.482065
ENET(0.2)	0.418732	0.440582	0.508033
AdLasso (0.5)	0.504661	0.532887	0.602173
AdLasso (1.0)	0.709350	0.748991	0.823003
AdLasso (2.0)	1.386904	1.442611	1.540284
No-covariate model	24.33297	24.66349	25.04612

Table 3.3. Bootstrap based 95% confidence intervals for the estimated survival probability (using both LASSO and PLS prediction) at different time points for the patient censored at 28.3 months of the Michigan Lung Cancer study.

Time (t) (in months)	95% CI for the estimated survival probability at t	
	LASSO	PLS
12	$(0.619, 1.000)$	$(0.618, 1.000)$
36	$(0.516, 1.000)$	$(0.445, 1.000)$
48	$(0.502, 1.000)$	$(0.405, 1.000)$

Table 3.4 Optimal (minimum) MREE values for different pseudo-value based regression as well as MREE of a no-covariate model under a high NSR of 1.0 for the sparse regression scenario (a) from the illness-death model in Section 3.2.1

Type of regression	Optimal MREE value under NSR 1.0
PLS	22.93322
LASSO	22.99224
ENET (0.8)	22.98864
ENET (0.6)	22.98392
ENET (0.4)	22.97804
ENET (0.2)	22.96856
No-covariate model	24.05346

Table 3.5. Optimal (minimum) MREE values for different pseudo-value based regression as well as MREE of a no-covariate model under a high NSR of 0.1 for the non-sparse regression scenario (b) from illness-death model in Section 3.2.1

Type of regression	Optimal MREE value under NSR 0.1
PLS	2.4591
LASSO	2.8076
ENET (0.8)	2.8096
ENET (0.6)	2.8138
ENET (0.4)	2.8213
ENET (0.2)	2.8428
No-covariate model	24.1069

Table 3.6 Optimal MREE values of different regression methods in the survival model in Section 3.2.2 for the regression scenario (a) where $\beta_j = j \bmod 5$, if $1 \leq j \leq 100$, otherwise $\beta_j = 0$, and regression scenario (b) where $\beta_j = 1/j$, $j = 1, 2, \cdots, 10^4$.

Regression type	Optimal MREE for regression (a)	Optimal MREE for regression (b)
Cox-LASSO	0.5117074	0.8826327
Pseudo-LASSO	0.4632573	0.9577252
Pseudo-PLS	0.1240558	0.7305356

Table 3.7. Average estimated survival probabilities of male and female patients in Michigan lung cancer study using both LASSO and PLS regression based on pseudo-value technique.

Time (in months)	Average estimated survival probability			
	LASSO regression		PLS regression	
	Male	Female	Male	Female
6	0.94077	0.93660	0.94216	0.93834
12	0.85419	0.84955	0.85675	0.85095
24	0.76925	0.77772	0.77689	0.77589
28	0.76191	0.77370	0.76371	0.77529
30	0.73162	0.77126	0.73580	0.77177
36	0.68527	0.71694	0.69169	0.71892
48	0.63377	0.69492	0.64496	0.69980
90	0.50420	0.67726	0.41111	0.66226

Figure 3.1. Plot of MREE values of different pseudo-value based regression methods for a wide range of components/steps under low and high NSR values for sparse regression scenario (a) of illness-death model from Section 3.2.1

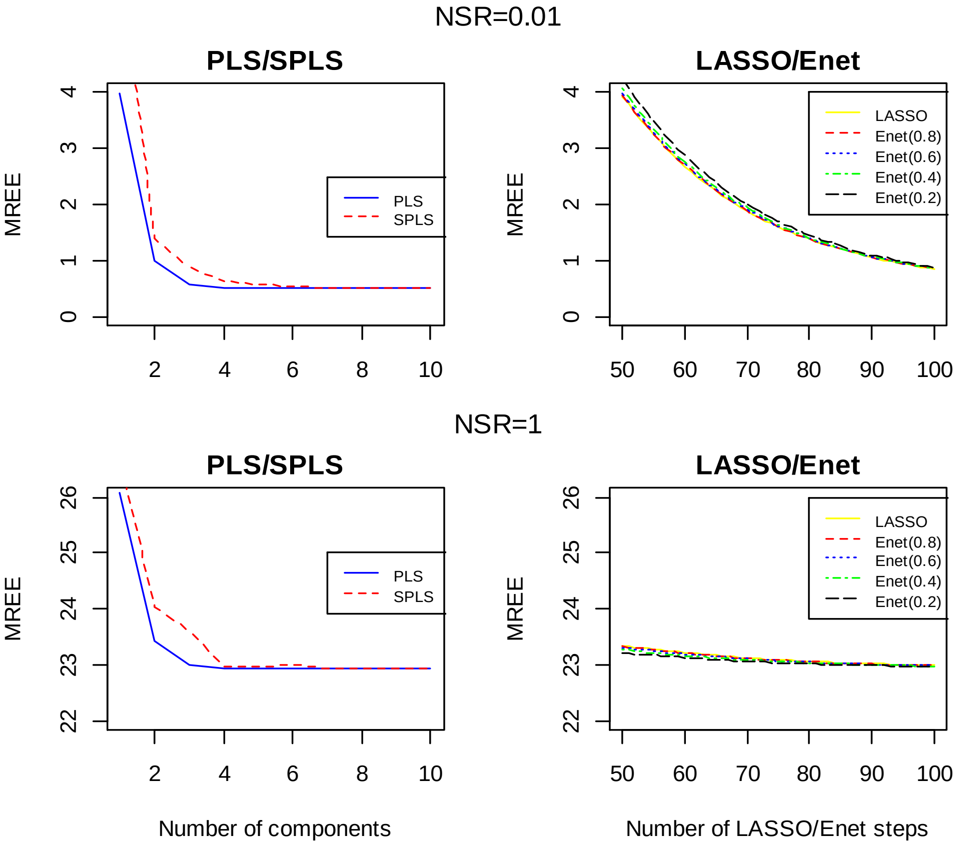


Figure 3.2. Plot of MREE values of different pseudo-value based regression methods for a wide range of components/steps under low and high NSR values for non-sparse regression scenario (b) of illness-death model from Section 3.2.1

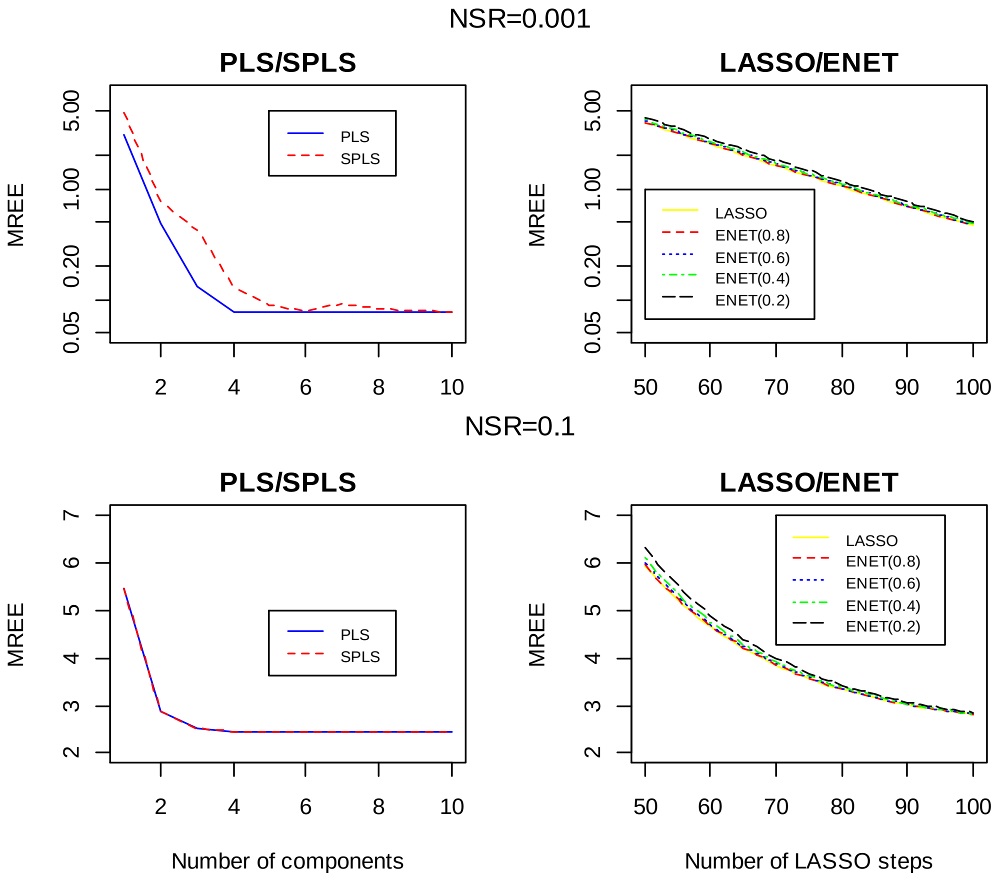


Figure 3.3. Plot of MREE values for Cox-LASSO and Pseudo-value (PV) LASSO for different LASSO steps and MREE values of pseudo-value PLS for different PLS components for survival model (a) from Section 3.2.2 where $\beta_j = j \bmod 5$, if $1 \leq j \leq 100$, otherwise $\beta_j = 0$.

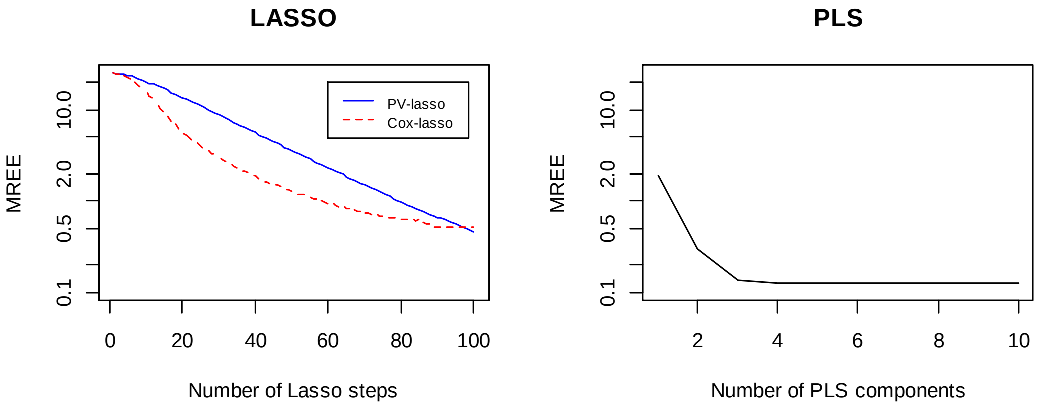


Figure 3.4. Plot of Predicted survival (using both LASSO and PLS methods) of a patient in the Michigan Lung cancer study who was actually censored at 28.3 months of the study.

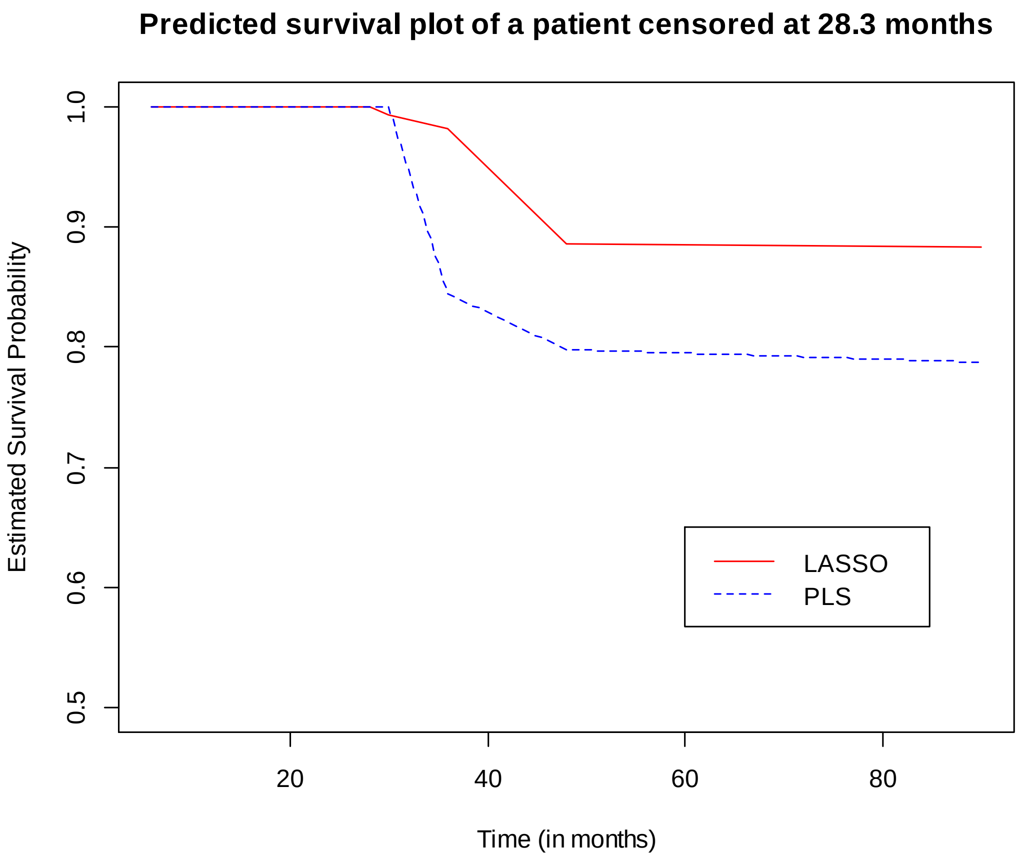


Figure 3.5 Plots showing the error (MREE) rates for different pseudo-value based regression methods under three different censoring rates for sparse regression scenario (a) in illness-death model from Section 3.2.1

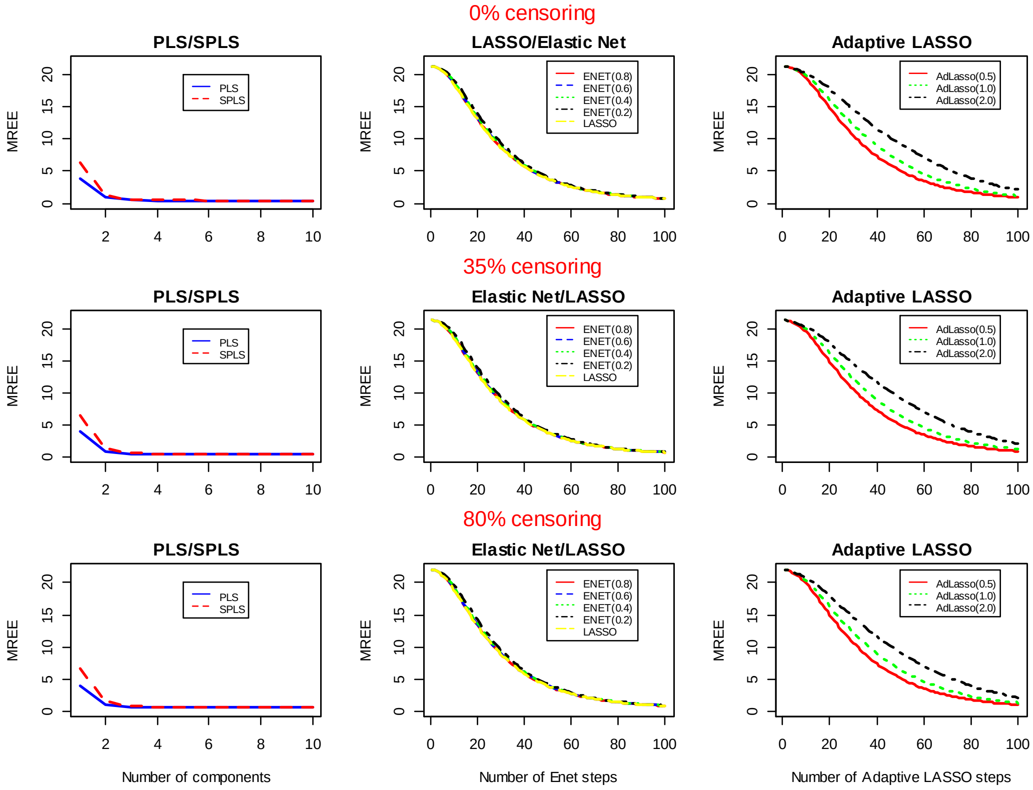


Figure 3.6 Plot showing the temporal plot of MREE values of pseudo-value based LASSO and PLS regression from sparse regression scenario (a) of illness-death model from Section 3.2.1

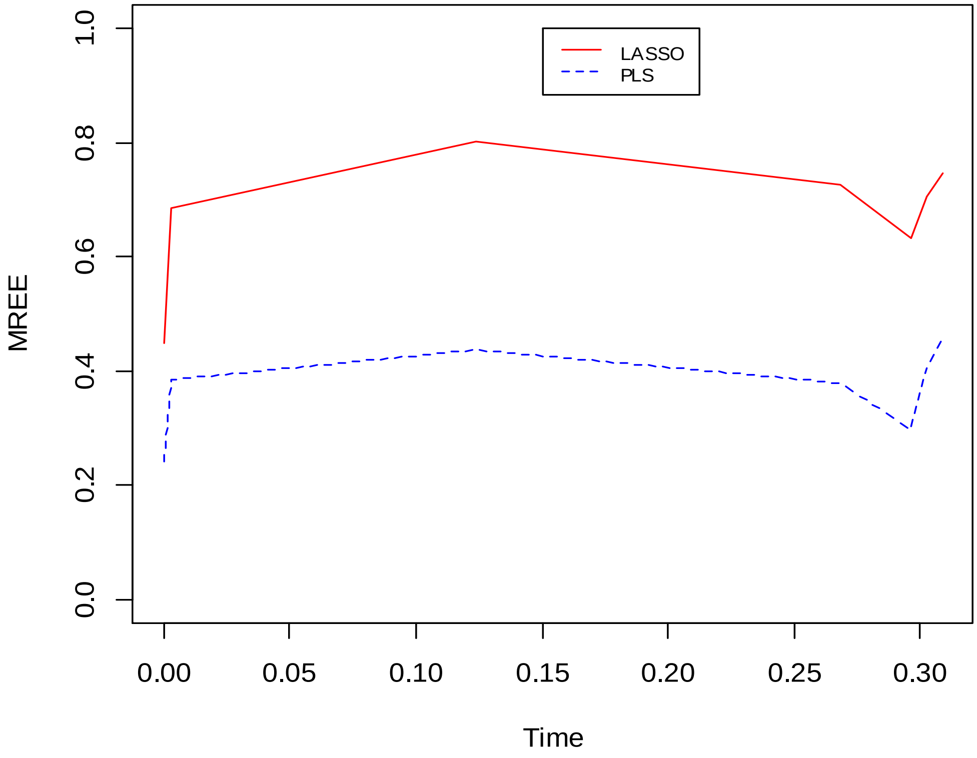


Figure 3.7 Plot showing the error (MREE) rates for different pseudo-value based regression methods under three different censoring rates for non-sparse regression scenario (b) in illness-death model from Section 3.2.1

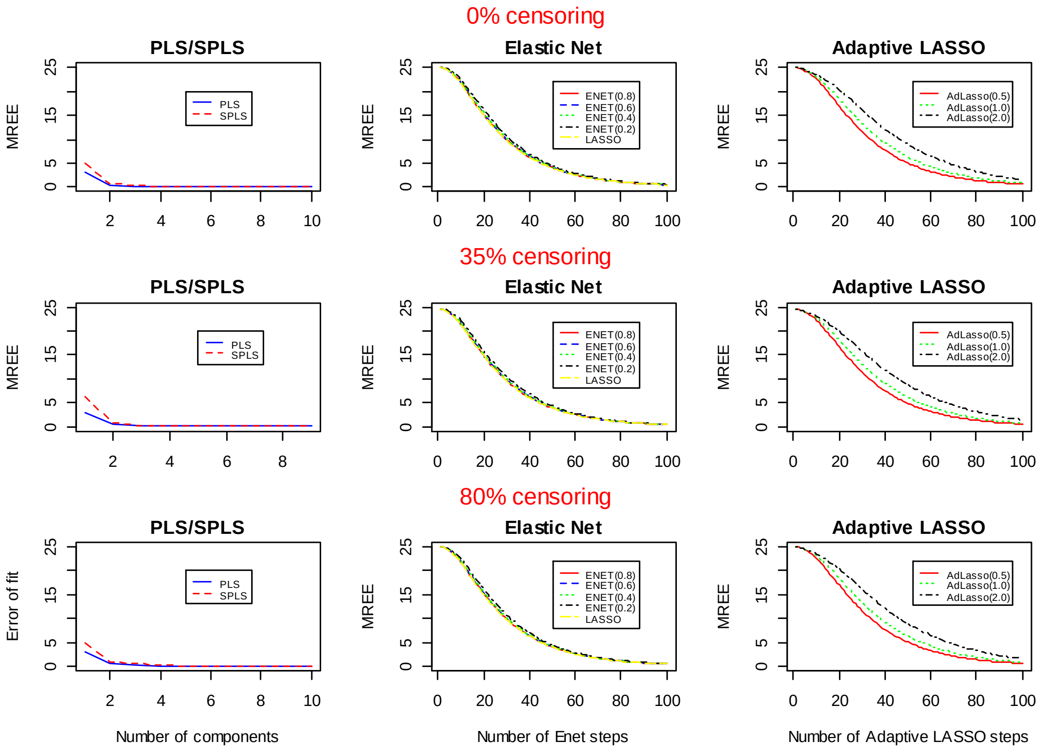


Figure 3.8 Plot showing the MREE values for Cox-LASSO and Pseudo-value LASSO for different LASSO steps and MREE values of pseudo-value PLS for different PLS components for survival model (b) from Section 3.2.2 where $\beta_j = 1/j$, $j = 1, 2, \cdots, 10^4$.

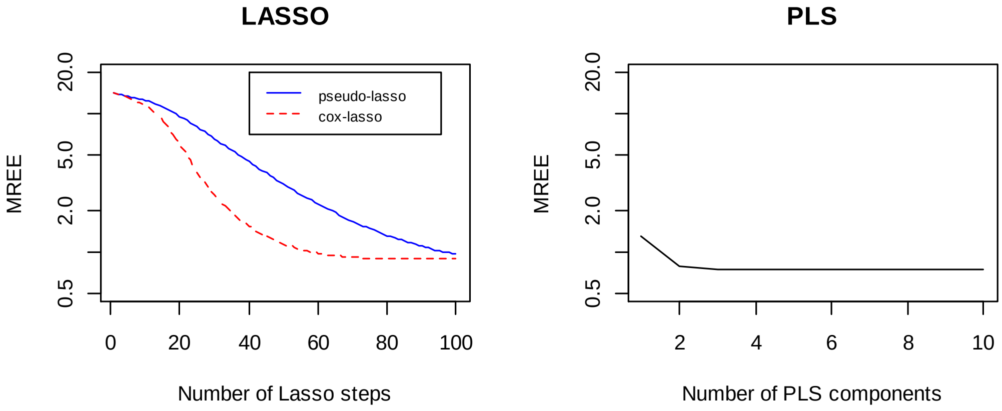


Figure 3.9 Plots of the $MAEP$ values of estimated survival probability of Michigan lung cancer patients at time $t = 30$ months using both LASSO and PLS regression based on pseudo-values.

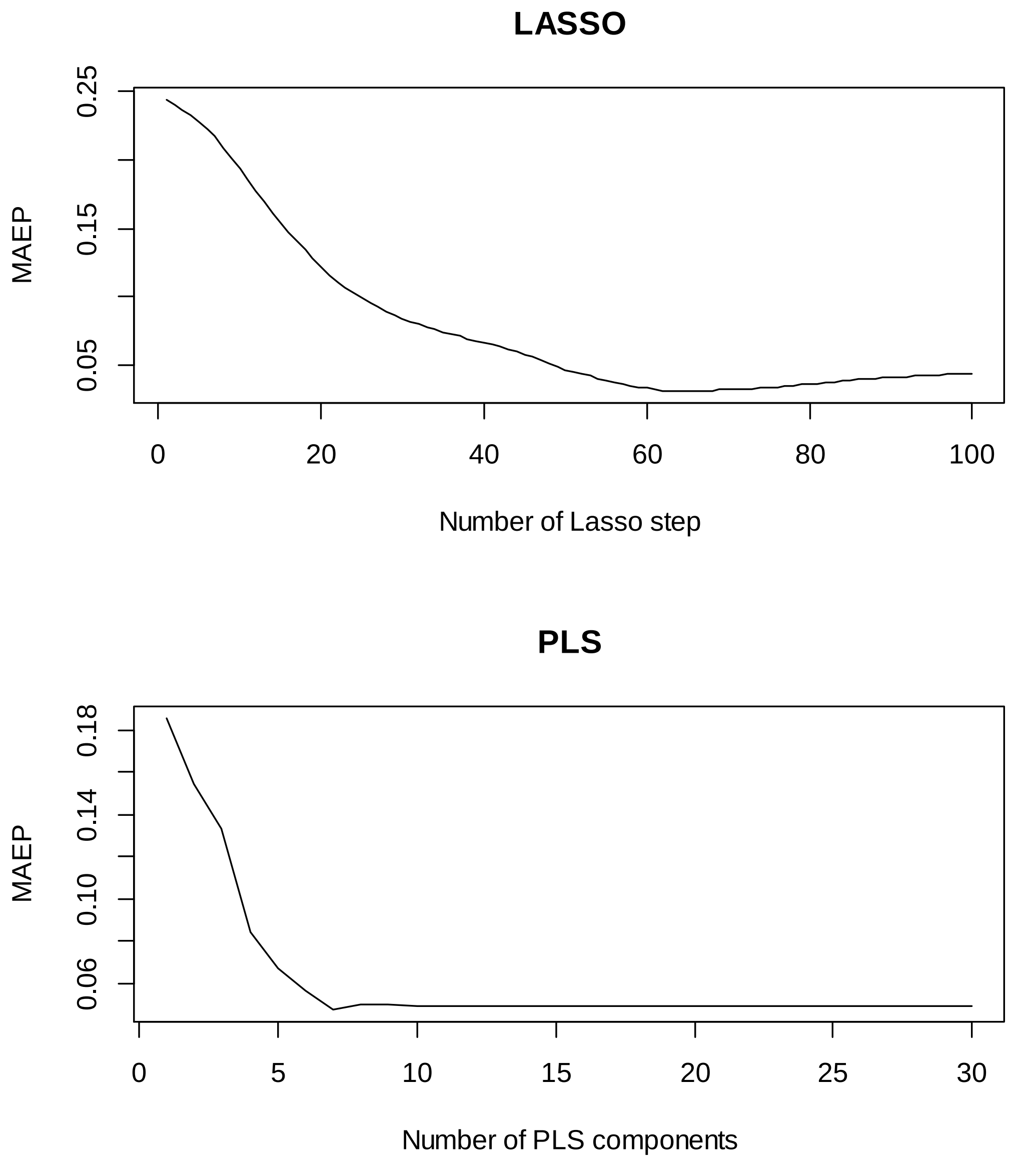


Figure 3.10 Plots of the time varying average predicted survival probabilities of stage-1 tumor and stage-3 tumor patients in Michigan lung cancer study using both LASSO and PLS regression based on pseudo-values.

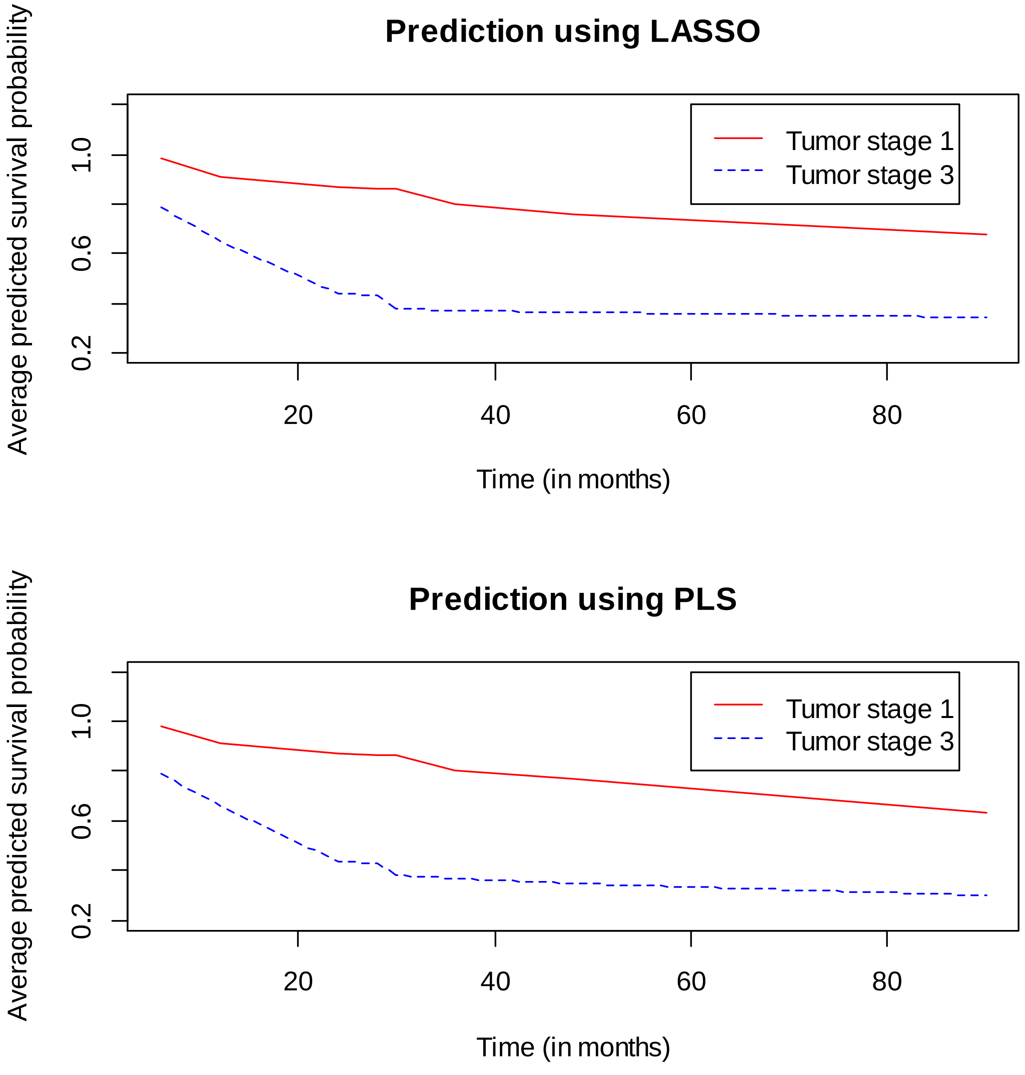
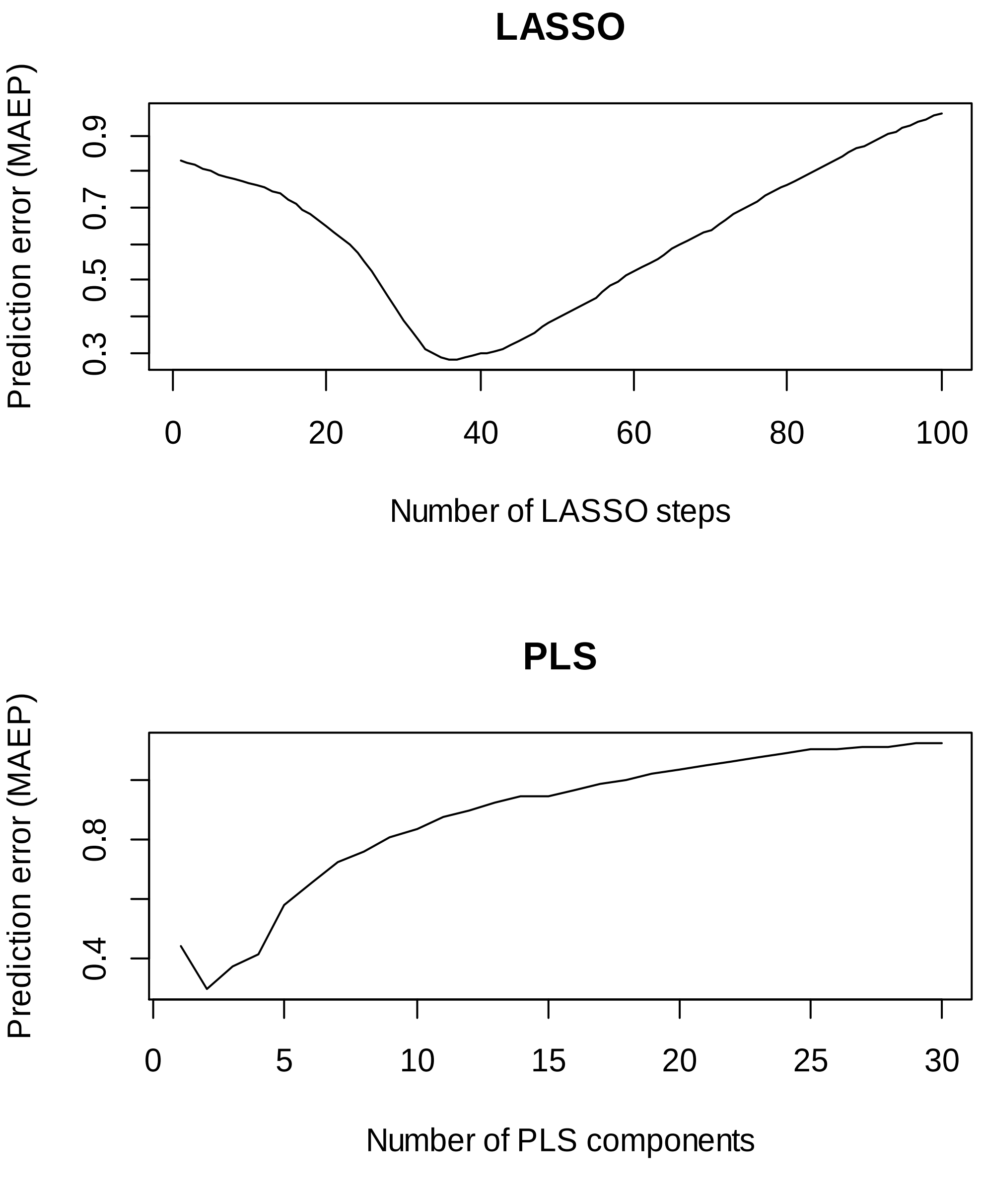


Figure 3.11. Plot of the $MAEP$ values of estimated probability of 'death during progression' within 3 years of diagnosis for ICGC lung adenocarcinoma patients using both LASSO and PLS regression based on pseudo-values.



3.5. Technical Details

Aalen-Johansen Estimator of State Occupation Probability

Suppose we have a multistate Markov process with finite number of states. Let $S = \{1, 2, \cdots, m\}$ denote the state space of the multistate process, and $\alpha_{gh}(t)$ denote the transition hazard, also known as transition intensity, of transition from state g to state h at time t, where g,$h \in S$, and $g \neq h$. The transition intensities describe the instantaneous risk of transition from one state to another. This means that $\alpha_{gh}(t)dt$ denotes the probability that an individual who is in state g just before time t will make a transition to state h within the very small time interval $[t, t + dt)$. Also suppose, for all g, $h \in S$, $P^{tr}_{gh}(s, t)$ denote the probability that an individual who is in state g at time s will be in state h at a later time t, such that $\mathbf{P}(s, t)$ is a matrix of these transition probabilities $P^{tr}_{gh}(s, t)$, where g, $h \in S$. Evidently $\mathbf{P}(s, t)$ is a square matrix of order m. In addition, let $P_h(t)$ denote the probability that an individual will be in state h at a given time t.

Now, suppose we have a sample of n individuals for our study. The individuals are followed up to different time points, such that some of the observations may be right censored. We assume that the censoring mechanism is such that the censoring times carry no information on the risks of transitions between the states, i.e., we have independent right censoring. Let $t_1 < t_2 < \cdots$ be the time points at which transitions were observed between any two states. We denote d_{ghj} as the number of individuals who experience a transition from state g to state h at time t_j and $d_{gj} = \sum_{h \neq g} d_{ghj}$ as the number of transitions at time t_j from state g. Also, let r_{gj} denote the number of individuals in state g just prior to time t_j. Then the Aalen-Johansen estimator of $\mathbf{P}(s, t)$ is given by

$$\widehat{\mathbf{P}}(s, t) = \prod_{s < t_j \leq t} \left(\boldsymbol{I} + \widehat{A}_j \right).$$

Here $\boldsymbol{I}$ is an identity matrix of order $m \times m$, $\widehat{A}_j$ is a matrix of order $m \times m$ where the entry in the (g, h) cell is given by $\widehat{\alpha}_{ghj} = d_{ghj}/r_{gj}$ for $h \neq g$, and the entry in the (g, g)

cell is given by $\widehat{\alpha}_{ggj} = -d_{gj}/r_{gj}$, and the matrix product is taken in the increasing order of t_js. Then the estimator of $P_h(t)$, the occupation probability of state h at time $t,$ as proposed by Aalen and Johansen (1978) is given by

$$\widehat{p}_h(t) = \sum_{g=1}^{m} \frac{r_{g0}}{n} \widehat{\mathrm{P}}_{gh}(0,t)$$

where r_{g0} is the number of individuals in state g at time '0' (initial or starting time point of the process), $\widehat{\mathrm{P}}_{gh}(0,t)$ is the entry of the (g,h) cell of the matrix $\widehat{\mathrm{P}}(0,t)$.

High Dimensional Regression methods

Partial Least Squares (PLS) and Sparse Partial Least Squares (SPLS)

PLS was introduced by Herman Wold in the 1960's (Wold, 1966) and later on it gained immense popularity in the field of chemometrics. The main purpose of the PLS technique is to extract a few underlying or latent variables among a huge set of explanatory variables, such that most of the variation in the data can be accounted by these extracted latent variables. Although a biased regression method, PLS finds its use in predictive modeling in situations where the ordinary least squares (OLS) technique fails, for instance in data where the number of explanatory variables is far more than the sample size. In such cases PLS reduces the dimension of the original explanatory variables by constructing a smaller set of latent variables through linear combinations of the original variables, and then OLS regression is carried out with the new set of variables. An overview of PLS can be found in Tobias (1997). Let Y be a single response variable, and $X_1, X_2, \cdots, X_p$ be the p explanatory or predictor variables. We have a sample of size n and the sample vectors $X_{.j} = (X_{1j}, X_{2j}, \cdots, X_{nj})$, $1 \leq j \leq p$, and $Y = (Y_1, Y_2, \cdots, Y_n)$ are standardized. In the next step a set of orthogonal latent factors $\{t^{(1)},\ t^{(2)}, \cdots, t^{(q)}\}$ is obtained from the full set of p covariates such that $t^{(k)} = (X_1, X_2, \cdots, X_p)c^{(k)},$ for $k = 1, 2, \cdots, q$. Here q is a tuning parameter denoting the number of PLS terms. q has to be less than the sample size n to facilitate OLS regression on the reduced set of latent variables. Besides, q is usually much smaller than p. The variables $t^{(k)}$ are obtained recursively from the covariates $X_1, X_2, \cdots, X_p$, as well as the response Y. After obtaining $t^{(1)}$, $t^{(2)}, \cdots, t^{(k-1)}$, the vector of constants $c^{(k)}$ (of unit length) is obtained in such a way that $t^{(k)}$, which is nothing but the linear combination $(X_1, X_2, \cdots, X_p)c^{(k)}$, is orthogonal to each of the $t^{(i)}$s $(i = 1, 2, \cdots, k-1)$, and has the largest (sample) covariance with Y. Once we obtain the q, Y is regressed on $t^{(1)}$, $t^{(2)}, \cdots, t^{(q)}$ such that $\widehat{Y} = \sum_{k=1}^{q} \widehat{\gamma}_k t^{(k)}$. Let $\widehat{\beta} = C\widehat{\gamma}$, where C is the matrix containing the direction vectors $c^{(1)}$, $c^{(2)}, \cdots, c^{(q)}$, then we have $\widehat{Y} = \sum_{j=1}^{p} \widehat{\beta}_j X_j$. In general to obtain the k^{th} PLS component, when k is large, one may use several available algorithms. In this article, we have used orthogonal scores algorithm (Martens and Naes, 1989) for the purpose of simulations and data analysis. Also, for choosing the optimal value for the tuning parameter q, the number of PLS components in the model, one may take help of the leave-one-out cross-validation technique. For the simulation studies in this article we have used a range of q values for PLS regression.

Although PLS performs dimension reduction by forming linear combinations of the original explanatory variables, it does not necessarily perform variable selection and is sometimes hard to interpret in case of a large number of predictor variables. To introduce this variable selection feature in the formulation of PLS, Chun and Keles (2010) formulated Sparse Partial Least Squares (SPLS). They impose an additional L_1 constraint on the direction vectors in the formulation of the PLS components. As a result there is a new tuning (threshold) parameter (η) involved in addition to the tuning parameter for the number of latent components (q). This user-defined parameter η determines the amount of sparsity introduced in the SPLS modeling. In practice, one can find the optimal values for η and q through cross-validation. For the simulation studies in this article, we have used a range of q values and for each value of q we have chosen the optimal value of the threshold parameter η by cross-validation.

LASSO, Elastic Net, and Adaptive LASSO

LASSO (Tibshirani, 1996) is a very popular shrinkage regression technique that is often useful in situations where the number of regressors are very large compared to the sample size. This method fits a linear model of the form $\widehat{Y} = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j X_j$ by minimizing the error sum of squares $\sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} \widehat{\beta}_j X_j \right)^2$ subject to an L_1 constraint $\sum_{j=1}^{p} \left| \widehat{\beta}_j \right| \leq s$, where s is a user-specified constant. Alternatively the whole minimization problem can be summarized as minimizing the L_1 penalized error sum of squares $\sum_{i=1}^{n} \left\{ \left(Y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} \widehat{\beta}_j X_j \right)^2 + \lambda \sum_{j=1}^{p} \left| \widehat{\beta}_j \right| \right\}$. In this case λ is a tuning (shrinkage) parameter defined by the user. In practice, optimal value for λ can be obtained through crossvalidation. Due to the L_1 penalization, the LASSO regression technique shrinks a number of regression coefficients to zero, thus enabling variable selection besides shrinkage regression. In this article we use the cyclical coordinate descent algorithm Friedman, Hastie and Tibshirani (2010) for fitting and prediction based on LASSO regression model for the entire solution path of λ.

Although LASSO is a very popular shrinkage and variable selection procedure, it lacks oracle property. Besides in presence of a set of highly correlated explanatory variable, LASSO has the tendency to shrink the coefficients of all but one of the correlated regressors to zero. Thus it selects only one out of a group of correlated variable ignoring the rest. To overcome this problem of ignoring all but one of the correlated regressors, Zou and Hastie (2005) proposed the elastic net (ENET) which can be thought of as an extension of LASSO that is robust to the presence of highly correlated variables. The ENET method deals with a mixture of L_1(LASSO) and L_2(ridge regression) penalties by minimizing the penalized sum of squares

$$\sum_{i=1}^{n}\Big\{\Big(Y_i-\widehat{\beta}_0-\sum_{j=1}^{p}\widehat{\beta}_jX_j\Big)^2+\lambda_2\sum_{j=1}^{p}\widehat{\beta}_j^2+\lambda_1\sum_{j=1}^{p}\big|\widehat{\beta}_j\big|\Big\}.$$

In general, elastic net can be viewed as a method for minimizing the following penalized error sum of squares

$$\sum_{i=1}^{n}\Big\{\Big(Y_i-\beta_0-\sum_{j=1}^{p}\beta_jX_j\Big)^2+\lambda\Big(\frac{(1-\alpha)}{2}\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}\big|\beta_j\big|\Big)\Big\}$$

where $\lambda(>0)$ is a complexity parameter and $\alpha\in[0,1]$ is elastic net mixing parameter compromising between LASSO $(\alpha=1)$ and ridge regression $(\alpha=0)$ $(\text{Friedman et al.}, 2010)$. A good introduction on the LASSO and ENET methods can be found in the book by Hastie, Tibshirani, and Friedman (2009).

As mentioned before LASSO suffers from the lack of oracle property. In order to overcome this shortcoming of LASSO, Zou (2006) proposed the adaptive LASSO (AdLasso) technique that minimizes a penalized error sum of squares such that adaptive weights are used for penalizing different coefficients in the L_1 penalty. The AdLasso method minimizes the following penalized error sum of squares $\sum_{i=1}^{n}\Big\{\big(Y_i-\beta_0-\sum_{j=1}^{p}\beta_jX_j\big)^2+\lambda\sum_{j=1}^{p}\widehat{w}_j\big|\beta_j\big|\Big\}$. Here the adaptive data-driven weight $\widehat{w}_j$ $(j=1,2,\cdots,p)$ is calculated as $\widehat{w}_j=\Big(\big|\widehat{\beta}_j^{in}\big|\Big)^{-\gamma}$, where γ (weight parameter) is a positive constant and $\widehat{\beta}^{in}$ is some initial consistent estimator of β. Zou (2006) suggested the use of OLS estimate of β, or the ridge regression estimate of β if multicollinearity is a concern. In practice the optimal choices of λ and γ are chosen from a grid of values by cross-validation, where γ is usually selected from the set $\{0.5, 1, 2\}$. For the simulations in this article we have calculated the full solution path of λ for each value of the γ from the above set. The estimates of the regression coefficients obtained from the AdLasso have been shown to possess the oracle property.

Derivation of theoretical state occupation probability in a Markovian irreversible illness-death model as described in Section 3.1.1

Let $P_h(t;Z)$ denote the occupation probability (conditional on covariate vector Z) at state h at time t, and $P_{h,k}(s,t;Z)$ denote the probability (conditional on Z) that an individual at state h at time s would be at state k at time $t(>s)$. Also, assume $\lambda_{h,k}(t;Z)$ and $\Lambda_{h,k}(t;Z)$ are the transition hazard and the cumulative transition hazard from state h to state k at time t. Then , we have :

$P_1(t;Z) = P_{1,1}(0,t;Z)$,

$P_2(t;Z) = P_{1,2}(0,t;Z) = \int_0^t P_{11}(0,u-;Z).\lambda_{1,2}(u;Z).P_{2,2}(u+,t;Z)du$, and

$P_3(t;Z) = 1 - P_1(t;Z) - P_2(t;Z)$.

For the irreversible illness-death model simulated in Section 3.1,

$P_{1,1}(0,t;Z) = exp(-(\Lambda_{1,2}(t;Z) + \Lambda_{1,2}(t;Z))) = \left(1 - \Phi\left(\frac{log(t) - Z\beta}{\sigma}\right)\right).\left(1 - \Phi\left(\frac{log(t) - Z\gamma}{\sigma}\right)\right)$

$\lambda_{1,2}(t;Z) = \frac{\phi\left(\frac{log(t)-Z\beta}{\sigma}\right)}{1-\Phi\left(\frac{log(t)-Z\beta}{\sigma}\right)}.\frac{1}{\sigma}.\frac{1}{t}$,

$P_{2,2}(s,t;Z) = exp\left(-\int_s^t \lambda_{2,3}(u;Z)du\right) = \frac{exp(-\Lambda_{2,3}(t;Z))}{exp(-\Lambda_{2,3}(s;Z))} = \frac{1-\Phi\left(\frac{log(t)-Z\gamma}{\sigma}\right)}{1-\Phi\left(\frac{log(s)-Z\gamma}{\sigma}\right)}$

Then, after some calculations, we get, $P_2(t;Z) = \left(1 - \Phi\left(\frac{log(t) - Z\gamma}{\sigma}\right)\right).\left(\Phi\left(\frac{log(t) - Z\beta}{\sigma}\right)\right)$.

Here $\Phi(x)$ is the standard normal distribution function at x.

CHAPTER 4

ClusterRankTest: AN R PACKAGE FOR RANK BASED TESTS FOR CLUSTER DATA WITH INFORMATIVE CLUSTER SIZE AND INTRA-CLUSTER GROUP SIZE

4.1. Utility of the Package

This package carries out rank based testing in clustered data through the methods developed in Datta and Satten (2005), Datta and Satten (2008), Dutta and Datta (2015). Among rank-sum tests in clustered data, the test by Datta and Satten (2005) is the most applicable one in case the cluster sizes are informative, while the test by Dutta and Datta (2015) performs best in case of informative ICG sizes. An additional advantage of using these tests is that even if the cluster sizes or the ICG sizes are not informative the Datta-Satten and Dutta-Datta rank sum tests are valid tests with reasonable performances (Dutta and Datta, 2015). For paired comparison in clustered data, the signed-rank test by Datta and Satten (2008) works well in case of informative cluster size. All the three aforementioned tests can be carried out through a single function *clus.rank.sum* of the ClusterRankTest package.

4.2. Tests Involved in the Package

4.2.1. A rank-sum test for clustered data when cluster size is informative

Let M denote the number of clusters and let X_{ik} denote the k^{th} observation in the i^{th} cluster, $1 \leq k \leq N_i,\ 1 \leq i \leq M,$ where N_i denotes the number of observations in the i^{th}cluster. Let G_{ik} be the indicator denoting the binary group membership (0 or 1) of the k^{th} observation in the i^{th} cluster. Thus the entire data set consists of $\{\mathbb{V}_i : 1 \leq i \leq M\}$, with $\mathbb{V}_i = \left\{N_i,\, X_{ik},\, G_{ik},\, 1 \leq k \leq N_i\right\}$ corresponding to the i^{th} cluster. Also, let N_{i1}

and N_{i0} be the numbers of observations in the i^{th} cluster belonging to group 1 and group $0,$ respectively. Thus, we have $N_{i1} + N_{i0} = N_i.$ The null hypothesis we consider is that the observations from the two groups follow the same marginal distribution. Mathematically, $H_0 : P \ (X_{ik} \le x | G_{ik} = 0) = P(X_{ik} \le x | G_{ik} = 1)$, for all $x.$ The empirical analogue of the above "group specific" (e.g., conditional) marginal distributions involved in the hypothesis, can be constructed as

$$\widehat{\mathcal{F}}(x|d) = \frac{\sum\limits_{i=1}^{M} \frac{1}{N_i} \sum\limits_{k=1}^{N_i} I(X_{ik} \le x,\, G_{_{ik}} = d)}{\sum\limits_{i=1}^{M} \frac{1}{N_i} \sum\limits_{k=1}^{N_i} I(G_{_{ik}} = d)},\, d = 0,\, 1.$$

If $\mathcal{F}_1$ is the distribution function which is estimated through $\widehat{\mathcal{F}}$ as defined above, then testing of hypothesis involving $\mathcal{F}_1$ is appropriate if the outcome from a cluster depend on the cluster size. The rank-sum test proposed by Datta and Satten (2005) addresses the case of informative cluster size by testing hypothesis involving $\mathcal{F}_1$. The Datta-Satten rank-sum test statistic is defined as

$$S = \frac{1}{M+1} \sum_{i=1}^{M} \sum_{k=1}^{N_i} \frac{G_{_{ik}}}{N_i} \left[1 + \frac{1}{2} \sum_{j \ne i} \{F_j(X_{ik}) + F_j(X_{ik} -)\}\right],$$

where $F_j(x) = \frac{1}{N_j} \sum\limits_{k=1}^{Nj} I(X_{jk} \le x)$, and $F_j(x -) = \frac{1}{N_j} \sum\limits_{k=1}^{Nj} I(X_{jk} < x).$

For the above statistic S, the expected value $E(S) = \frac{1}{2} \sum\limits_{i=1}^{M} \frac{N_{i1}}{N_i}.$

The variance of S is estimated as $\widehat{V}(S) = \sum\limits_{i=1}^{M} \left(\widehat{W}_i - E(W_i)\right)^2,$

where, $\widehat{W}_i = \frac{1}{2N_i(M+1)} \sum\limits_{k=1}^{N_i} \left((M-1)G_{_{ik}} - \sum\limits_{j \ne i} \frac{N_{j1}}{Nj}\right) \left(\widehat{F}(X_{ik}) + \widehat{F}(X_{ik} -)\right),\ \widehat{F} = \frac{\sum_i N_i F_i}{\sum_i N_i}$

, and $E(W_i) = \frac{M}{2(M+1)} \left(\frac{N_{i1}}{N_i} - \frac{1}{M} \sum\limits_{j=1}^{M} \frac{N_{j1}}{Nj}\right).$

A large sample test can be carried out based on the standardized test statistic.

4.2.2. A rank-sum test for clustered data when the intra-cluster group size is informative

Following from the previous section, the null hypothesis of interest is that the outcome from the two groups follow the same marginal distribution, i.e., $H_0 : P\ (X_{ik} \le x|\ G_{ik} = 0) = P(X_{ik} \le x|\, G_{ik} = 1)$. If $\mathcal{F}$ is a group-specific marginal distribution involved in the aforementioned hypothesis then the empirical analogue of $\mathcal{F}$ can be constructed in a way different from that in section 2.1 as follows,

$$\widehat{\mathcal{F}}_2(x|d) = \frac{\sum\limits_{i=1}^{M} \frac{1}{2N_{id}} \sum\limits_{k=1}^{N_i} I(X_{ik} \le x,\, G_{ik} = d)}{\sum\limits_{i=1}^{M} \frac{1}{2N_{id}} \sum\limits_{k=1}^{N_i} I(G_{ik} = d)},\ d = 0,\ 1.$$

Hypothesis involving $\mathcal{F}_2$, the population counterpart of $\widehat{\mathcal{F}}_2$, is appropriate if the outcome from a certain group in a given cluster is associated with the number of observations belonging to that group in that cluster, i.e., we have an informative intra-cluster group size scenario. The rank-sum test developed by Dutta and Datta (2015) addresses this scenario of informative intra-cluster group size in clustered data.

For the sake of simplicity, we relabel the observations in a typical cluster according to their group membership. In the i^{th} cluster, let $\{X_{i1}^{(1)},\, X_{i2}^{(1)},\, \cdots,\, X_{iN_{i1}}^{(1)}\}$ represent the set of observations belonging to the group indexed by 1, while $\{X_{i1}^{(0)},\, X_{i2}^{(0)},\, \cdots,\, X_{iN_{i0}}^{(0)}\}$ represents the set of observations belonging to the group indexed by 0. We denote these two sets as $\underline{\mathbf{X}}_i^{(1)}$and $\underline{\mathbf{X}}_i^{(0)}$ respectively. Then the Dutta-Datta rank-sum test statistic is given by

$$T = \sum_{i=1}^{M}\left(\sum_{k=1}^{N_{i1}} \frac{1}{2N_{i1}}\left[1 + \frac{1}{2}\sum_{j \neq i}\left\{F_j\big(X_{ik}^{(1)}\big) + F_j\big(X_{ik}^{(1)} -\big)\right\}\right]\right),$$

where $F_j(x) = \frac{1}{2N_{j1}}\sum_{h=1}^{N_{j1}} I\left(X_{jh}^{(1)} \le x\right) + \frac{1}{2N_{j0}}\sum_{h'=1}^{N_{j0}} I\left(X_{jh'}^{(0)} \le x\right)$. The expected value of the test statistic T is given by $E(T) = \frac{M(M+1)}{4}$.

The variance of T can be estimated through a jackknife technique. If $\widehat{V}(T)$ is the estimated variance of T then we have $\widehat{V}(T) = \frac{M}{M-1}\sum\limits_{i=1}^{M}\left(T_i^* - \overline{T}^*\right)^2$, where $T_i^* = T_{full} - T_{-i}$, T_{full} is the test statistic calculated from the full data, T_{-i} is the value of the statistic T calculated after deleting the i^{th} cluster, and $\overline{T}^* = \left(\sum_{i=1}^{M} T_i^*\right) \Big/ M$.

A large sample test is carried out based on the standardized test statistic.

4.2.3. A signed-rank test for clustered data when cluster size is informative

Suppose we have paired outcomes in a clustered data where (U, V) denote the paired outcome variables. If i index clusters and j index pairs within a cluster, then (U_{ij}, V_{ij}) denotes the j^{th} pair in the i^{th} cluster, where $1 \le j \le N_i,\ \ 1 \le i \le M$. Suppose $X_{ij} = U_{ij} - V_{ij}$, i.e. X_{ij} is the pair-specific difference in the outcome measure. If $\mathcal{F}_X$ be the distribution function of pairwise differences of a randomly chosen pair in a randomly chosen cluster, then the empirical analogue of $\mathcal{F}_X$ can be constructed as

$$\widehat{\mathcal{F}}_X(x) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i} I\Big(X_{ij} \le x\Big).$$

The null hypothesis of interest is $H_0 : \mathcal{F}_X$ is symmetric (around 0), which is tested against $H_1 : \mathcal{F}_X$ is not symmetric. The Datta-Satten signed-rank statistic is given by

$$Q_M = \sum_{i=1}^{M}\left(\frac{N_i^+ - N_i^-}{N_i}\right) + \sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i} sign(X_{ij})\widehat{D}_i(|X_{ij}|),$$

where $N_i^+ = \sum_{j=1}^{N_i} I\Big(X_{ij} > 0\Big)$, $N_i^- = \sum_{j=1}^{N_i} I\Big(X_{ij} < 0\Big)$, $\widehat{D}_i(|x|) = \sum_{k \ne i} \widehat{H}_k(x)$, and $\widehat{H}_i(x) = \frac{1}{2N_i}\left(\sum_{j=1}^{N_i} I\Big(|X_{ij}| \le x\Big) + \sum_{j=1}^{N_i} I\Big(|X_{ij}| < x\Big)\right)$.

The variance estimate of the statistic is obtained as $\sum_{i=1}^{M} \widehat{S}_{i,M}^2$ where

$\widehat{S}_{i,M} = \frac{N_i^+ - N_i^-}{N_i} + \left(\frac{M-1}{N_i}\right)\sum_{j=1}^{N_i} sign(X_{ij})\widehat{H}(X_{ij})$ with $\widehat{H}(x) = \frac{\sum_{i=1}^{M} N_i \widehat{H}_i(x)}{\sum_{i=1}^{M} N_i}$

A large sample testing of H_0 against H_1 is carried out through the standardized test statistic $Z_M = Q_M \Big/ \left(\sum_{i=1}^{M} \widehat{S}_{i,M}^2\right)^{\frac{1}{2}}$.

4.3. ClusterRankTest package implementation

The ClusterRankTest package may be applied for comparing the distributions of outcomes from two groups in a clustered data. ClusterRankTest is entirely written in R programming language and it can be installed on all operating systems for which R software is installed. The main function of the package, *clus.rank.sum*, carries out the rank-sum tests for clustered data as outlined in Datta and Satten (2005) and Dutta and Datta (2016) as well as a signed-rank test for clustered data (Datta and Satten, 2008). This function specifically calculates the test statistics, as described in Section 2 and the corresponding p-values. This function has a print method for computed test statistics and p-values. In this section, we give a few examples to show how the package can be used for successful implementation of the aforementioned tests.

For rank-sum test, the user has an option of using either the test outlined in Datta and Satten (2005) or the test developed by Dutta and Datta (2016). In either case there are three arguments for the function: Cluster, X, and grp. Cluster indicates the cluster id in which an observation belongs, X denotes the outcome value corresponding to a given observation, and grp denotes the group membership (binary) indicator taking values 0 or 1. Each of these three inputs should be a numeric vector with matched components. In addition to these, there is another argument 'test' to indicate which test is to be used specifically. For rank-sum test, one can use either test = "DS" to carry out the rank-sum test developed by Datta and Satten (2005) or test = "DD" to carry out the rank-sum test developed by Dutta and Datta (2016). Following is an example to elaborate the usage of clus.rank.sum function in carrying out rank-sum tests for cluster data:

```
R> ## Creating the data ##
R> Cluster= c(1,1,1,1,2,2,2,2,2,2,3,3,3,3)
R> X=c(0.01,0.5,0.4,0.75,0.07,0.33,0.42,-0.1,0.36,0.73,0.38,-0.11,0.24,0.38)
R> group=c(1,0,0,0,1,1,1,1,0,0,1,0,0,0)
```

```
R> ## Dutta-Datta rank-sum test ##
R> clus.rank.sum(Cluster=Cluster,X=X,grp=group,test="DD")
##  pvalue = 0.1742314
##  Test Statistic = 2.625
```

```
R> ## Datta-Satten rank-sum test ##
R> clus.rank.sum(Cluster=Cluster,X=X,grp=group,test="DS")
##  pvalue = 0.3754203
##  Test Statistic = -0.8863661
```

For carrying out the signed-rank test (Datta and Satten, 2008) using the **clus.rank.sum** function, the first three arguments are **Cluster**, **X**, and **Y**. Like before **Cluster** indicates the cluster id in which an observation belongs. Here **X** and **Y** denote the paired outcomes from a given observation. Each of the three inputs, **Cluster**, **X**, and **Y**, should be a numeric vector with matched components. Additionally, the user needs to specify **test = "SDS"** as an argument in the **clus.rank.sum** function to carry out the signed-rank test. An example, showing the proper usage of **clus.rank.sum** function for carrrying out signed-rank test for clustered data, is given below:

```
R> ## Creating the data ##
R> Cluster=c(1,1,2,2,2,2,3,3)
R> X=c(1,4,2,4,6,7,4,7)
R> Y=c(4,8,5,10,7,9,9,8)
R> clus.rank.sum(Cluster=Cluster,X=X,Y=Y,test="SDS")
##  pvalue = 0.08702814
##  Test Statistic = -1.711287
```

CHAPTER 5

AN ALIGNED RANK-SUM TEST FOR CLUSTERED DATA WHEN THE INTRA-CLUSTER GROUP SIZE IS INFORMATIVE

5.1. Rank Based Estimation and Testing

Let M denote the number of clusters and let Y_{ij} denote the j^{th} observation in the i^{th} cluster, $1 \leq j \leq N_i,\ 1 \leq i \leq M,$ where N_i denotes the number of observations in the i^{th}cluster. Let Z_{ij} be the covariate indicating the binary group membership (0 or 1) of the j^{th} observation in the i^{th} cluster. Also, let N_{i1} and N_{i0} be the numbers of observations in the i^{th} cluster belonging to group 1 and group 0, respectively. Thus, we have $N_{i1} + N_{i0} = N_i$. The null hypothesis we consider is that the observations from the two groups follow the same marginal distribution. Mathematically, it is written as

$$H : P\ (Y_{ij} \leq y |\, Z_{ij} = 0) = P(Y_{ij} \leq y |\, Z_{ij} = 1)\ (= \mathcal{F}(y), \text{ say}), \text{ for all } y.$$

In case of informative ICG sizes the empirical versions of the above distribution functions can be constructed as

$$\widehat{\mathcal{F}}(y|d) = \frac{\sum_{i=1}^{M} \frac{1}{2N_{id}} \sum_{j=1}^{N_i} I(Y_{ij} \leq y,\, Z_{ij} = d)}{\sum_{i=1}^{M} \frac{1}{2N_{id}} \sum_{j=1}^{N_i} I(Z_{ij} = d)},\ d = 0,\, 1. \qquad (1)$$

In addition to Z_{ij}, we have an observed confounder covariate vector X_{ij}, that is unrelated to the group factor. Suppose the outcome Y_{ij} can be represented through the marginal linear model

$$Y_{ij} = \alpha + \beta_1 Z_{ij} + \beta_2 X_{ij} + \beta_3 X_{ij} Z_{ij} + \epsilon_{ij}\ ,\ 1 \leq i \leq M,\ 1 \leq j \leq N_i \qquad (2)$$

where ϵ_{ij} are the model errors for the i^{th} cluster having a common cluster-specific distribution.

Here β_1 is the regression coefficient corresponding to the grouping factor, while β_2 and β_3 are the vectors of regression coefficients corresponding to the confounder covariate X, and the interaction effect between Z and X, respectively. The intercept in the model is given by α. In such a case testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$ may seem reasonable for checking whether the outcomes from two groups have the same marginal distribution. But if the distributions of the confounder variable and its interaction term, which are mostly unknown, differ from that of the random model error, then testing the hypothesis H_0 against H_1, may not be the same as testing H against K: not H. So, for proper inference we need to adjust for the effect of the confounder covariates. In case of rank based inference one way to solve this problem is to carry out an aligned rank test (Hájek, Šidák, and Sen, 1999, Section 10.1.2). The idea is to estimate the regression coefficients $(\beta_2$ and $\beta_3)$ through appropriate rank based statistics under H_0 and construct aligned residuals by plugging in these estimates. Suppose, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are the rank based estimators (R-estimators) obtained by minimizing the following weighted score function:

$$S_w(b) = \sum_i \sum_j w_{ij} r_w(e_{ij}(b)) e_{ij}(b) \tag{3}$$

where $b = (b_1, b_2)$, $e_{ij}(b) = Y_{ij} - b_1 X_{ij} - b_2 X_{ij} Z_{ij}$, w_{ij} is an associated weight, and $r_w(e_{ij}(b)) = \sum_k \sum_l w_{kl} I(e_{kl}(b) \leq e_{ij}(b))$.

The estimate of the intercept in then obtained as $\widehat{\alpha} = \inf \; \{t : \overline{F}_{w,\widehat{\beta}}(t) \geq \frac{1}{2}\}$, where $\overline{F}_{w,\widehat{\beta}}(t) = r_w(e_{ij}(\widehat{\beta}))/M$.

Then the aligned residual can be obtained as $Y_{ij}^{'} = Y_{ij} - \widehat{\alpha} - \widehat{\beta}_2 X_{ij} - \widehat{\beta}_3 X_{ij} Z_{ij}$, $1 \leq j \leq N_i, \; 1 \leq i \leq M$. Treating these aligned residuals as responses we can carry out a testing to compare the distributions of outcomes from two groups. In case of the informative ICG sizes, we propose the use of the test by Dutta and Datta (2015) that carries out a large sample test using a standardized version of the test statistic T, where T

is given as

$$T = \sum_{i=1}^{M}\left(\sum_{j=1}^{N_i}\frac{I(Z_{ij}=1)}{2N_{i1}}\left[1+\frac{1}{2}\sum_{k\neq i}\left\{F_k\left(Y_{ij}^{'}\right)+F_k\left(Y_{ij}^{'}-\right)\right\}\right]\right) \tag{4}$$

where $F_k(y) = \frac{1}{2N_{k1}}\sum_{h=1}^{N_k} I(Z_{kh}=1) I\left(Y_{kh}^{'} \leq y\right) + \frac{1}{2N_{k0}}\sum_{h=1}^{N_k} I(Z_{kh}=0) I\left(Y_{kh}^{'} \leq y\right)$

An important question, however, still remains unanswered: what should be the choice of the weights w_{ij} in the construction of the R-estimators $\widehat{\beta}_2$ and $\widehat{\beta}_3$, and hence, in the formation of the aligned residuals $Y_{ij}^{'}$. In case of the informative ICG sizes, where the test by Dutta and Datta (2016) appears to be the most appropriate, we propose the following choice for w_{ij}:

$$w_{ij} = \left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=1)}{N_{i1}}\right) \tag{5}$$

which leads to the following rank-based minimizing function:

$$S^*(b) = \sum_i\sum_j\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=1)}{N_{i1}}\right) r^*(e_{ij}(b))e_{ij}(b) \tag{6}$$

with $r^*(e_{ij}(b)) = \sum_k\sum_l\left(\frac{I(Z_{kl}=0)}{N_{k0}} + \frac{I(Z_{kl}=1)}{N_{k1}}\right) I(e_{kl}(b) \leq e_{ij}(b))$. The weight proposed above basically uses the inverse of the ICG size corresponding to an outcome from a given group in a typical cluster, which seems reasonable in case of informative ICG sizes. Apart from the above choice, there are a few other possible choices of the weights including: $w_{ij} = 1/N_i$, for all $1 \leq j \leq N_i$, where each observation is weighted by the inverse of the cluster size to which it belongs, and $w_{ij} = 1$ for all $1 \leq j \leq N_i$, $1 \leq i \leq M$, where all the observations contributes equally in constructing the R-estimator of the regression coefficients. The above-mentioned choices of the weight w_{ij} lead to different estimating equations and, hence, to different R-estimators of the regression coefficients. In a later section we have compared the performances of the R-estimators obtained from these different choices in estimating the regression parameters as well as their effects on the size and power performance of the Dutta-Datta test through the construction of the corresponding aligned residuals.

5.2. Large Sample Inference of the R-estimators

For the sake of notational convenience, let us denote U_{ij} as a vector such that $U_{ij} = (X_{ij}, X_{ij}Z_{ij})^T$, and $\underline{\beta}$ as a vector of regression coefficients such that $\underline{\beta} = (\beta_2,\ \beta_3)^T$. The marginal distribution of the errors, in case of informative ICG sizes, is given by

$$\mathcal{F}_\epsilon(y) = E\left(\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left\{\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=1)}{N_{i1}}\right)I(\epsilon_{ij} \le y)\right\}\right). \qquad (7)$$

Suppose f and $f^{'}$ are the first and second order derivatives derivatives, respectively, of $\mathcal{F}_\epsilon$. Then the asymptotic distribution of the R-estimator of $\underline{\beta}$ obtained by minimizing (6), namely $\underline{\widehat{\beta}}$, is given by the following theorem (a sketch of its proof is given in the Appendix):

THEOREM 1. *Under* H_0, *as* $M \to \infty, \sqrt{M}\underline{\widehat{\beta}} \xrightarrow{d} N(0, \tau^2\Gamma^{-1}\Sigma\Gamma^{-1})$ *under certain regularity conditions.*

Here $\Sigma = lim_{M\to\infty}M^{-1}\sum_{i=1}^{M}\Sigma_i$ where $\Sigma_i = Var\left[\sum_{j=1}^{N_i}\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=0)}{N_{i0}}\right)F(e_{ij}(\underline{\beta}))U_{ij}\right]$ and $F(e_{ij}(\underline{\beta})) = (r^*(e_{ij}(\underline{\beta})))/M,$ $\Gamma = E\left[M^{-1}\sum_i\sum_j\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=1)}{N_{i1}}\right)U_{ij}U_{ij}^T\right],$ $\tau^{-1} = -E(\mathcal{F}_\epsilon(\epsilon_{ij})S(\epsilon_{ij}))$, such that $S = (\log(f))^{'}$. In that case the asymptotic variance covariance matrix of $\underline{\widehat{\beta}}$ is given by

$$\widehat{V}\left(\underline{\widehat{\beta}}\right) = \widehat{\tau}^2\widehat{\Gamma}^{-1}\widehat{\Sigma}\widehat{\Gamma}^{-1}/M \qquad (8)$$

where $\widehat{\Gamma} = \frac{1}{M}\sum_i\sum_j\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=1)}{N_{i1}}\right)U_{ij}U_{ij}^T,\ \widehat{\Sigma} =$ $M^{-1}\sum_{i=1}^{M}\left(\sum_{j=1}^{N_i}\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=0)}{N_{i0}}\right)F(e_{ij}(\underline{\widehat{\beta}}))U_{ij}\right)\left(\sum_{j=1}^{N_i}\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=0)}{N_{i0}}\right)F(e_{ij}(\underline{\widehat{\beta}}))U_{ij}\right)^T,$ and $\widehat{\tau} = \left[-\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=0)}{N_{i0}}\right)F(e_{ij}(\underline{\widehat{\beta}}))\widehat{S}(e_{ij}(\underline{\widehat{\beta}}))\right]^{-1}$, such that

$$\widehat{S}(e) = \frac{1}{2h} log\left(\frac{\widehat{f}(e+h)}{\widehat{f}(e-h)}\right), \widehat{f}(e) = \frac{1}{Mh}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=0)}{N_{i0}}\right)K\left(\frac{e_{ij}(\underline{\widehat{\beta}})-e}{h}\right).$$

Here h is a bandwidth sequence and K is a kernel density. The selection of optimal bandwidth h is often considered a separate problem that needs a more detailed investigation and may even need some additional assumptions. For our analysis, the asymptotic variance estimate does not appear to be too sensitive to the choice of the bandwidth.

5.3. Simulation Results

5.3.1 *Comparing the accuracy of different R-estimators*

We assume that there are M clusters. For a typical cluster i, the j^{th} outcome is denoted by Y_{ij}, where Y_{ij} is generated through a random effects model given below

$$Y_{ij} = \beta_1 Z_{ij} + \beta_2 X_i + \beta_3 X_i Z_{ij} + a_i + e_{ij},\ 1 \le j \le N_i,\ 1 \le i \le M \qquad (9)$$

Here Z_{ij} is binary group indicator for the j^{th} outcome in the i^{th}cluster taking values 0 or 1, X_i is a a cluster-level covariate (confounder) for the i^{th} cluster unrelated to the binary group indicator covariate Z, while a_i is a random cluster effect due to the i^{th}cluster, and e_{ij} is the random error term.

Simulation Scenario 1

We generate X_i from Normal$(0,\ 1)$, a_i from Normal(0, 0.25), while the cluster size N_i=N_i^*+2 where $N_i^* \sim$ Binomial(20, q_i) such that logit(q_i)=1.8X_i. Given N_i, N_{i1} (the number of observations from group 1 in the i^{th} cluster) is generated as $N_{i1}|N_i \sim max(1,$ Binomial$(N_i - 1, p_i))$, where logit(p_i)=0.5a_i. Then, the number of observations in group 0 i.e. $N_{i0} = N_i - N_{i1}$. We assign $Z_{ij} = 0$ for $1 \le j \le N_{i0}$, while $Z_{ij} = 1$ for $N_{i0} + 1 \le j \le N_i$. Also $e_{ij} \sim$ Normal$(0,\ 0.3)$ for all $1 \le j \le N_i$, and $1 \le i \le M$. This represents a scenario where the cluster size is correlated with the outcome from that cluster through the cluster-level confounder covariate, while given the cluster size the ICG size is also correlated with the outcome through the random cluster effect. So, this is a situation where we have informative cluster size as well as informative ICG size in a clustered data. Then under $H_0 : \beta_1 = 0$, we obtain the R-estimators of the regression parameters β_2 and β_3 by minimizing the score function $S^*(b)$ in (6). Additionally, we also obtain the R-estimators through minimizing the weighted score function $S_w(b)$ in (3) with the choice of w_{ij} as $w_{ij} = 1/N_i$, as well as $w_{ij} = 1$. Let $\left(\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}\right)$ be the R-estimator of (β_2, β_3) obtained by minimizing (6), $\left(\widehat{\beta}_2^{CW}, \widehat{\beta}_3^{CW}\right)$ be the R-estimator of (β_2, β_3) obtained by minimizing (3) with $w_{ij} = 1/N_i$, and $\left(\widehat{\beta}_2^{UW}, \widehat{\beta}_3^{UW}\right)$ be the R-estimator of (β_2, β_3) obtained by minimizing (3) with $w_{ij} = 1$. Then for the number of clusters (M) as 10 and the true value of (β_2, β_3) as $(5, 9)$ we compare the bias and the empirical standard error of each of the above-mentioned R-estimators based on 500 Monte-Carlo simulations in Table 5.1. From Table 5.1 we find that for $\widehat{\beta}^{UW}$, i.e. the unweighted R-estimator, is always the worst in case of estimating the regression coefficient (β_2) the main effect of the confounder as well as the regression coefficient (β_3) of the interaction effect between the confounder and the group indicator covariate. The biases (and standard errors) of $\widehat{\beta}_2^{ICGW}$ and $\widehat{\beta}_2^{CW}$ are similar in estimating β_2, but the bias (and standard error) of the $\widehat{\beta}_3^{ICGW}$ is much smaller than that of $\widehat{\beta}_3^{CW}$ in estimating the regression coefficient (β_3) corresponding to the interaction between the confounder and the group indicator covariate.

Simulation Scenario 2

This simulation scheme is same as the previous except for the fact that the cluster size (N_i) for a typical cluster i does not depend on the cluster level covariate (confounder) X_i. Here, N_i=N_i^*+2 where $N_i^* \sim$ Binomial(20, 0.6). Apart from N_i, all other quantities, namely N_{i1}, N_{i0}, Y_{ij}, X_i, Z_{ij}, a_i, e_{ij}, are generated in the same way as that in model 1. So, in this case the outcome variable Y in a typical cluster is correlated with the ICG size but not with the cluster size. This is a scenario of a clustered data with informative ICG sizes, but the cluster sizes are not informative. This is the major difference from the simulation scenario 1. For 10 clusters and the true value of (β_2, β_3) as $(5, 9)$, Table 5.2 shows the biases and the empirical standard errors of $\left(\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}\right)$ and $\left(\widehat{\beta}_2^{CW}, \widehat{\beta}_3^{CW}\right)$ based on 500 Monte-Carlo simulations. From Table 2 we find that in the estimation of both β_2 and β_3 the ICG-weighted R-estimators $\left(\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}\right)$ have superior performances than the corresponding cluster-weighted R-estimators $\left(\widehat{\beta}_2^{CW}, \widehat{\beta}_3^{CW}\right)$in terms of bias and standard error. This is mainly due to the fact that the cluster sizes are no longer informative in this scenario but the ICG sizes are.

5.3.2 *Size and power performance of rank-sum test based on aligned residuals*

As before, Y_{ij}, the j^{th} outcome in cluster i, is generated through the random effects model $Y_{ij} = \beta_1 Z_{ij} + \beta_2 X_i + \beta_3 X_i Z_{ij} + a_i + e_{ij}$, $1 \leq j \leq N_i$, $1 \leq i \leq M$. We generate X_i from Normal$(0,\ 1)$, a_i from Normal$(0, 0.25)$, N_i=N_i^*+2 where $N_i^* \sim$ Binomial(20, 0.6). Given N_i, N_{i1} is generated as $N_{i1}|N_i \sim max(1, \text{Binomial}(N_i - 1, p_i))$, where logit($p_i$)=0.5$a_i$. $N_{i0} = N_i - N_{i1}$. We assign $Z_{ij} = 0$ for $1 \leq j \leq N_{i0}$, while $Z_{ij} = 1$ for $N_{i0} + 1 \leq j \leq N_i$. Also, e_{ij} is generated from Normal$(0,\ 0.1)$ distribution. This is a scenario of informative ICG sizes in a clustered data. If $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are the R-estimates of β_2 and β_3 respectively, then the aligned residuals are obtained as $Y_{ij}' = Y_{ij} - \widehat{\beta}_2 X_{ij} - \widehat{\beta}_3 X_{ij} Z_{ij}$, $1 \leq j \leq N_i$, $1 \leq i \leq M$. Treating these aligned residuals as modified (covariate-adjusted) outcomes we apply the rank-sum test of Dutta and Datta (2015) as outlined in Section 2. Now, there are three choices for $\left(\widehat{\beta}_2,\ \widehat{\beta}_3\right)$ as discussed before, namely $\left(\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}\right)$, $\left(\widehat{\beta}_2^{CW}, \widehat{\beta}_3^{CW}\right)$, and $\left(\widehat{\beta}_2^{UW}, \widehat{\beta}_3^{UW}\right)$. These choices lead to different sets of aligned residuals. Under $H_0 : \beta_1 = 0$, we calculate the size of the Dutta-Datta rank-sum test based on the aligned residuals using each of the three choices for $\left(\widehat{\beta}_2,\ \widehat{\beta}_3\right)$. We also check the power performances of the test under the alternative hypothesis $H_1 : \beta_1 = 0.1$ using the three different choices for $\left(\widehat{\beta}_2,\ \widehat{\beta}_3\right)$. Table 5.3 illustrates the size and power values of the test for the different choices of $\left(\widehat{\beta}_2,\ \widehat{\beta}_3\right)$

based on 500 Monte-Carlo simulations. From Table 3 we find that the rank-sum test based on $\left(\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}\right)$ is the one that has the empirical size closest to the nominal size of 0.05. Also the test based on $\left(\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}\right)$ has the most superior power performance. This implies that the test based on the $\left(\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}\right)$ is the most appropriate test for a clustered data with informative ICG sizes.

5.3.3 *Comparison of asymptotic variance estimate of* $\underline{\widehat{\beta}}$ *with the empirical variance*

Simulation results from Section 3.1 and Section 3.2 show that the R-estimator $\left(\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}\right)$ performs the best in term of estimating the confounder effects as well as maintaining superior size and power performance in a clustered data with informative ICG sizes. In this section we compare the average asymptotic standard error estimate of $\widehat{\beta}_3^{ICGW}$ (the R-estimator for the coefficient of interaction between the confounder and the grouping factor) with the empirical standard error of $\widehat{\beta}_3^{ICGW}$based on Monte-Carlo simulations. Table 5.4 shows the average asymptotic standard error estimate and the empirical standard error of $\widehat{\beta}_3^{ICGW}$based on 500 Monte-Carlo simulations for two different choices of M (number of clusters). The empirical standard error and the estimated asymptotic standard error appear to be close to each other as seen in Table 5.4.

5.4. Discussion

In this work we have shown that unweighted R-estimators of the confounder effects can lead to large bias in estimation and low power in associated testing problem in clustered data with informative cluster and ICG sizes. Through extensive simulation studies we have shown that the weighted R-estimators with weights involving ICG sizes work in the most efficient way in terms of having low bias in confounder effect estimation and maintaining the appropriate size and high power values in the associated aligned rank testing in clustered data with informative ICG sizes. Similar results can also be obtained in situations where clusters sizes are informative. Additionally, we have discussed the large sample properties of the weighted R-estimators and verified the asymptotic variance estimate through Monte-Carlo simulations.

Although developed for rank-sum test in clustered data with informative ICG sizes, this aligned rank transformation can also be extended to compare paired outcomes for clustered data with informative cluster sizes through the use of the signed-rank test developed in Datta and Satten (2008).

Table 5.1

Biases and the empirical standard errors of different R-estimators in simulation scenario 1

Estimator of β_2 (True value $= 5$)	Bias	Standard Error
$\widehat{\beta}_2^{ICGW}$	0.01109	0.47440
$\widehat{\beta}_2^{CW}$	-0.01165	0.48615
$\widehat{\beta}_2^{UW}$	0.35368	0.72985
Estimator of β_3 (True value $= 9$)	Bias	Standard Error
$\widehat{\beta}_3^{ICGW}$	-0.01306	0.15581
$\widehat{\beta}_3^{CW}$	0.05389	0.22346
$\widehat{\beta}_3^{UW}$	0.14466	0.41129

Table 5.2

Biases and the empirical standard errors of different R-estimators in simulation scenario 2

Estimator of β_2 $(\text{True value} = 5)$	Bias	Standard Error
$\widehat{\beta}_2^{ICGW}$	-0.00282	0.41835
$\widehat{\beta}_2^{CW}$	0.14516	1.68358
Estimator of β_3 $(\text{True value} = 9)$	Bias	Standard Error
$\widehat{\beta}_3^{ICGW}$	0.00341	0.06245
$\widehat{\beta}_3^{CW}$	-0.02169	0.70099

Table 5.3

Size and power values of the rank-sum test (nominal level = 0.05) based on the aligned residuals using the three different choices of the R-estimator.

Type of R-estimator	Size (under $\beta_1 = 0$)	Power (under $\beta_1 = 0.1$)
$\widehat{\beta}_2^{ICGW}, \widehat{\beta}_3^{ICGW}$	0.047	0.890
$\widehat{\beta}_2^{CW}, \widehat{\beta}_3^{CW}$	0.028	0.671
$\widehat{\beta}_2^{UW}, \widehat{\beta}_3^{UW}$	0.033	0.713

Table 5.4

Average asymptotic standard error estimate and the empirical standard error of $\widehat{\beta}_3^{ICGW}$

Number of Clusters (M)	Empirical SE	Estimated Asymptotic SE
10	0.065	0.068
50	0.031	0.036

5.5 Technical Details

A sketch of the proof of Theorem 1

Without loss of generality, let us assume that the true value of $\underline{\beta}$ is 0. From Section 2 we have $\widehat{\underline{\beta}}$ as the solution to the following R-estimating equation:

$$R_M(\underline{\beta}) = \sum_i \sum_j \left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=1)}{N_{i1}} \right) r(e_{ij}(\underline{\beta}))U_{ij} = 0. \qquad (10)$$

Then, following the arguments of Datta, Nevalainen, and Oja (2012), and Datta and Beck (2014), we have

$$M^{-\frac{1}{2}} R_M(0) \xrightarrow{d} N(0, \Sigma) \qquad (11)$$

where

$\Sigma = lim_{M \to \infty} M^{-1} \sum_{i=1}^{M} \Sigma_i$, with $\Sigma_i = Var\left[\sum_{j=1}^{N_i} \left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=0)}{N_{i0}} \right) F(e_{ij}(\underline{\beta}))U_{ij} \right]$,

and $F(e_{ij}(\underline{\beta})) = \frac{r^*(e_{ij}(\beta))}{M}$.

Following the expansions for R-estimators in Hettmansperger and McKean (2011, Chapter 3), and Datta and Beck (2014), we have

$$M^{-\frac{1}{2}} R_M(\underline{\beta})$$
$$= M^{-\frac{1}{2}} R_M(0) - \tau^{-1} \left(M^{-1} \sum_i \sum_j \left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=1)}{N_{i1}} \right) U_{ij} U_{ij}^T \right) \sqrt{M} \underline{\beta} + o_P(1) \qquad (12)$$

in the neighbourhood of 0. Hence, combining (10), (11), and (12) we get

$$\sqrt{M} \widehat{\underline{\beta}} \xrightarrow{d} N\big(0, \tau^2 \Gamma^{-1} \Sigma \Gamma^{-1}\big)$$

where $\Gamma = E\left[M^{-1} \sum_i \sum_j \left(\frac{I(Z_{ij}=0)}{N_{i0}} + \frac{I(Z_{ij}=1)}{N_{i1}} \right) U_{ij} U_{ij}^T \right]$.

R Codes

```
##############################
R code for R-estimation in Section 5.1
###############################
Y=mydata$Y
```

```
Z=mydata$Z
X=mydata$X
Cluster=mydata$Cluster
N <- length(unique(mydata$Cluster) )

n_clus <- NULL
n1_clus <- NULL
n0_clus <- NULL
for(i in 1:N)
{
n_clus[i] <- length(mydata[which(mydata[,4]==i),4])
n1_clus[i] <- length(mydata[which(mydata[,4]==i & mydata[,2]==1),4])
n0_clus[i] <- length(mydata[which(mydata[,4]==i & mydata[,2]==0),4])
}
model <- function(b){
temp2 <- NULL
for(i in 1:N){
c <- n_clus[i]
myclusdata <- mydata[which(Cluster==i),]
temp1 <- NULL
for(j in 1:c){
yij <- myclusdata[j,1]
xij <- myclusdata[j,3]
zij <- myclusdata[j,2]
temp <- NULL
for(k in 1:nrow(mydata)){
   ykl <- mydata[k,1]
```

```
    xkl <- mydata[k,3]
    zkl <- mydata[k,2]
    K <- mydata[k,4]
    Group <- mydata[k,2]
    w <- ifelse(Group==0,n0_clus[K],n1_clus[K])
    temp[k] <- (1/w)*I(ykl-b[1]*xkl-b[2]*xkl*zkl  <= yij-b[1]*xij-b[2]*xij*zij)
    }
reij <- sum(temp)
w <- ifelse(zij==0,n0_clus[i],n1_clus[i])
temp1[j] <- (1/w)*reij*(yij-b[1]*xij-b[2]*xij*zij)
}
temp2[i] <- sum(temp1)
}
final <- sum(temp2)
return(final)
}
 optim(par=c(2,5),fn=model)$par
```

CHAPTER 6

FUTURE WORKS

In this work we have discussed a number of projects related to nonparametric and semiparametric methods in clustered and multistate models. In doing so, we have faced some more questions related to these topics, some of which are being stated next as directions for future research works.

In the first project we have developed a rank-sum test for clustered data with informative ICG sizes. In deciding whether the ICG sizes are informative or not, we have to rely solely on our intuitive knowledge of the data generating process. However, this problem can treated in an objective manner through a hypothesis testing framework to check if the ICG sizes are really infromative.

In the second project involving temporal prediction based on high dimensional pseudo-value regression in multistate models, an important aspect is the selection of an uniform list of most influential covariates over a wide range of future timepoints. This calls for the development of proper variable selection techniques that address the temporal variation of covariate selection.

The new R package discussed in this dissertation can be extended to handle situations where the tests are carried out after adjusting for the effects of some additional covariates that may be present in the data.

The aligned rank test for clustered data has been discussed in this dissertation involves rank-sum test in presence of informative cluster of ICG sizes. This aligned rank transformation method can be easily extended to compare paired outcomes in clustered data in presence of informative cluster size through appropriate signed-rank statistic.

REFERENCES

Aalen, O. O., and Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* **5**, 141-150.

Andersen, P. K., Klein, J. P., and Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90**, 15-27.

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* **10**, 335-350.

Andersen, P. K., and Klein, J. P. (2007). Regression Analysis for Multistate Models Based on a Pseudo-value Approach, with Applications to Bone Marrow Transplantation Studies. *Scandinavian Journal of Statistics* **34**, 3-16.

Beck, J. D., Koch, G. G., Rozier, R. G., and Tudor, G. E. (1990). Prevalence and risk indicators for periodontal attachment loss in a population of older community-dwelling blacks and whites. *Journal of Periodontology* **61**, 521–528.

Beer, D. G., Kardia, S. L., Huang, C. C., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* **8**, 816-824.

Binder, N., Gerds, T. A., and Andersen, P. K. (2014). Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis* **20**, 303-315.

Blazer, D. G. and George, L. K. (2004). *Established Populations for Epidemiologic Studies of the Elderly*, 1996-1997: Piedmont Health Survey of the Elderly, Fourth In-Person Survey [Durham, Warren, Vance, Granville, and Franklin Counties, North Carolina] [Computer file]. ICPSR02744-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], doi:10.3886/ICPSR02744.

Chun, H., and Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society, Series B* **72**, 3-25.

Datta, S., and Satten, G. A. (2001). Validity of the Aalen–Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non-Markov models. *Statistics & Probability Letters* **55**, 403-411.

Datta, S., and Satten, G. A. (2005). Rank-sum tests for clustered data. *Journal of the American Statistical Association* **100**, 908-915.

Datta, S., and Satten, G. A. (2008). A Signed-rank test for clustered data. *Biometrics* **64**, 501-507.

Datta, S., Nevalainen, J., and Oja, H. (2012). A general class of signed-rank tests for clustered data when the cluster size is potentially informative. *Journal of nonparametric statistics* **24**, 797-808.

Datta, S., and Beck, J. D. (2014). Robust estimation of marginal regression parameters in clustered data. *Statistical modelling* **14**, 489-501.

Dutta, S., and Datta, S. (2015). A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics.* doi: 10.1111/biom.12447.

Frank, L. E., and Friedman, J. H., (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-135.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1.

Graw, F., Gerds, T. A., and Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* **15**, 241-255.

Hájek, J., Šidák, Z., and Sen, P. K. (1999). *Theory of Rank Tests*. San Diego, CA: Academic Press.

Hastie T., Tibshirani R., and Friedman, J. (2009). The Elements of Statistical Learning: Prediction, Inference and Data Mining. New York: Springer-Verlag.

Hoffman, E. B., Sen, P. K., and Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika* **88**, 1121-1134.

Huang, Y., and Leroux, B. (2011). Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. *Biometrics* **67**, 843-851.

Klein, J. P., and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* **61**, 223-229.

Klein, J. P., Logan, B., Harhoff, M., and Andersen, P. K. (2007). Analyzing survival curves at a fixed point in time. *Statistics in Medicine* **26**, 4505-4519.

Martens, H. and Naes, T. (1989). *Multivariate Calibration*. New York: Wiley.

Miller, R. G. (1974). The jackknife-a review. *Biometrika* **61**, 1-15.

Mogensen, U. B., and Gerds, T. A. (2013). A random forest approach for competing risks based on pseudo-values. *Statistics in Medicine* **32**, 3102-3114.

Rosner, B., Glynn, R. J., and Ting Lee, M. L. (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. *Biometrics* **59**, 1089-1098.

Rosner, B., Glynn, R. J., and Ting Lee, M. L. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* **62**, 185-192.

Sen, P. K. (1968). Robustness of some nonparametric procedures in linear models. *The Annals of Mathematical Statistics* **39**, 1913-1922.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385-395.

Williamson, J. M., Datta, S., and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59**, 36-42.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1**, 80-83.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, P.R.Krishnaiaah (ed), 391-420. New York: Academic Press.

Zou, H., and Hastie, T., (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 301-320.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

APPENDIX

The research projects presented in this dissertation were supported by NIH grants 1R03DE020839, 1R03DE022538, and NCI/NIH grant CA170091-01A1. I would like to thank Jim Beck and Kevin Moss in the School of Dentistry at the University of North Carolina for providing the data set on periodontal disease from the Piedmont 65+ Dental study, and the International Cancer Genome Consortium (ICGC) for making the Lung Adenocarcinoma data available.

CURRICULUM VITAE

NAME: Sandipan Dutta

ADDRESS: Department of Bionformatics and Biostatistics
University of Louisville
Louisville, KY 40202

EDUCATION: B.Sc, Statistics (Honors), University of Calcutta, India
M.Sc, Statistics, Indian Institute of Technology Kanpur

PROFFESIONAL SOCIETIES: American Statistical Association
Golden Key International Honour Society

AWARDS: Graduate Student Council Travel Award (2015), University of Louisville
SAMSI Travel Award for Bioinformatics: Opening Workshop (2014)
University Fellowship (offered by the School of Interdisciplinary and Graduate studies, University of Louisville)

PUBLICATIONS: Dutta S., Datta S., and Datta S. (2015) - Temporal prediction of future state occupation in a multistate model from high dimensional baseline covariates via pseudo-value regression. Submitted.

Dutta S., and Datta S. (2015) - A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. Biometrics, doi: 10.1111/biom.12447.

Sikdar S., Choo-Wosoba H., Abdia Y., Dutta S., Gill R., Datta S., and Datta S. (2014) - An integrative exploratory analysis of -omics data from the ICGC cancer genomes lung adenocarcinoma study. Systems Biomedicine 2, 56-64.

ORAL PRESENTATIONS: Paper presentation at the Ninth International Triennial Calcutta Symposium on Probability and Statistics, Kolkata, India, December 2015. "A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative".

Contributed paper presentation at the Joint Statistical Meetings (JSM), Seattle, WA, August, 2015. "A novel rank-sum test for clustered data when the number of subjects in a group within a cluster is informative".

POSTER PRESENTATION:	Bioinformatics: Opening Workshop in Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC, September, 2014. "A comprehensive omics study for the ICGC cancer genomes lung adenocarcinoma data".