8-2013

Novel methods based on regression techniques to analyze multistate models and high-dimensional omics data.

Sutirtha Chakraborty
University of Louisville

NOVEL METHODS BASED ON REGRESSION TECHNIQUES TO ANALYZE MULTISTATE MODELS AND HIGH-DIMENSIONAL OMICS DATA

By

Sutirtha Chakraborty
M.STAT, Indian Statistical Institute, India, 2009
B.Sc. Presidency University, 2007

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

Doctor of Philosophy

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY

August 2013

NOVEL METHODS BASED ON REGRESSION TECHNIQUES TO ANALYZE MULTISTATE MODELS AND HIGH-DIMENSIONAL OMICS DATA

By

Sutirtha Chakraborty
M.STAT, Indian Statistical Institute, India, 2009
B.Sc. Presidency University, 2007
A Dissertation Approved on

May 24, 2013

by the following Dissertation Committee:

_____
Dissertation Director
Dr. Somnath Datta

_____
Dissertation Co-Director
Dr. Susmita Datta

_____
Dr. Guy Brock

_____
Dr. Maiying Kong

_____
Dr. Ryan S. Gill

DEDICATION

This dissertation is dedicated to my parents

Mr. Tapan Kr. Chakraborty and Mrs. Debjani Chakraborty

and

my brother

Mr. Rahul Chakraborty

who have given me immense encouragement and endless moral support.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to Dr. Somnath Datta, my mentor and dissertation advisor and my co-advisor Dr. Susmita Datta for their immense academic and moral support that has played the fundamental role in the development of the dissertation. I would also like to thank the other committee members, Dr. Guy Brock, Dr. Maiying Kong and Dr. Ryan Gill for their constructive comments on the dissertation and academic assistance over the past two years. The research projects in this thesis are partially supported by Dr. Susmita Datta's grants from the National Science Foundation and National Institutes of Health and by Dr. Somnath Datta's grant from the National Security Agency. Sincere thanks to the Christopher & Dana Reeve Foundation NeuroRecovery Network for granting access to the Spinal Cord Injury Data and to Dr. Doug Lorentz for providing it. In addition, I would like to thank Ms. Tammi Alvey Thomas and Ms. Lynn Dosker for their enthusiastic support with the administrative issues. Finally, I would also like to express earnest acknowledgements and appreciation for my colleagues Juliet Ndukum, Joe Bible, Jasmit Shah, Hyouyoung Choo-Wosoba, Younathan Abdia, Yubing Wan, Chatura Siriwardhana, Manan Jhaveri, Malhar Jhaveri, who gave me a warm, friendly ambience in the midst of the busy work schedule and made these years full of joy and excitement. Special thanks to Dr. Guy Brock, Mr. Jason Banta of the SPHIS and Mr. Harrison Simrall of the Miller Information Technology center for their support with the computing resources.

ABSTRACT

NOVEL METHODS BASED ON REGRESSION TECHNIQUES TO ANALYZE MULTISTATE MODELS AND HIGH-DIMENSIONAL OMICS DATA

Sutirtha Chakraborty

May 24, 2013

The dissertation is based on four distinct research projects that are loosely interconnected by the common link of a regression framework. Chapter 1 provides an introductory outline of the problems addressed in the projects along with a detailed review of the previous works that have been done on them and a brief discussion on our newly developed methodologies. Chapter 2 describes the first project that is concerned with the identification of hidden subject-specific sources of heterogeneity in gene expression profiling analyses and adjusting for them by a technique based on Partial Least Squares (PLS) regression, in order to ensure a more accurate inference on the expression pattern of the genes over two different varieties of samples. Chapter 3 focuses on the development of an R package based on Project 1 and its performance evaluation with respect to other popular software dealing with differential gene expression analyses. Chapter 4 covers the third project that proposes a non-parametric regression method for the estimation of stage occupation probabilities at different time points in a right-censored multistate model data, using an Inverse Probability of Censoring (IPCW) (Datta and Satten,

2001) based version of the backfitting principle (Hastie and Tibshirani, 1992). Chapter 5 describes the fourth project which deals with the testing for the equality of the residual distributions after adjusting for available covariate information from the right censored waiting times of two groups of subjects, by using an Inverse Probability of Censoring weighted (IPCW) version of the Mann-Whitney U test.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
	1.1 Correcting for hidden sources of heterogeneity in Gene Expression Data	1
	1.2 Multistate Models and estimation of Stage Occupation Probabilities	4
	1.3 Formulation of Mann-Whitney U-tests in Multistate Model Data	7
II.	SURROGATE VARIABLE ANALYSIS USING PARTIAL LEAST SQUARES (SVA-PLS) IN GENE EXPRESSION STUDIES	10
	2.1 Motivation	10
	2.2 Methods	11
	2.3 Simulation Studies	17
	2.3.1 Independent Tests	19
	2.3.2 Cluster Dependent Tests	20
	2.4 Analysis of Leukemia Data	23
	2.5 Discussion	26
	2.6 Tables	29
	2.7 Figures	32
III	*sva-pls*: AN R PACKAGE TO CORRECT FOR HIDDEN FACTORS OF HETEROGENEITY IN GENE EXPRESSION DATA	35
	3.1 Motivation	35
	3.2 Brief Overview of the Package	37
	3.3 Comparative Evaluation with other Available Software	40
	3.4 Application on the Golub Data	41
	3.5 Discussion	42
	3.6 Tables	43
	3.7 Figures	45

IV. ESTIMATION OF TEMPORAL FUNCTIONS IN RIGHT CENSORED MULTISTATE MODEL DATA USING NON-PARAMETRIC REGRESSION VIA ADDITIVE MODELS 49
4.1 The Proposed Methodology 49
4.1.1 Data Structure and Notations 49
4.1.2 Additive Models 50
4.1.3 Conditional Transition Hazard Rates and State Occupation Probabilities 54
4.1.4 Censoring Hazards and Estimation of the Weights 56
4.2 Simulations 57
4.2.1 The Simulation Design 57
4.2.2 Conditionally Semi-Markov Network 58
4.2.3 Conditionally Markov Network 59
4.2.4 Study of the Censoring Bias 59
4.2.5 Study of the Overall Estimation Bias 62
4.2.6 Bootstrap Confidence Intervals 64
4.2.7 Tests for Regression Effects and a Power Study 66
4.3 Real Data Applications 68
4.3.1 Bone Marrow Transplant Data 68
4.3.2 Spinal Cord Injury Data 71
4.4 Discussion 73
4.5 Tables 75
4.6 Figures 78

V. TESTING THE EQUALITY OF THE WAITING TIME DISTRIBUTIONS BETWEEN TWO GROUPS OF INDIVIDUALS BY AN IPCW-BASED MANN-WHITNEY U-TEST AFTER ADJUSTING FOR AVAILABLE COVARIATES 83
5.1 Data Structure and Notations 83
5.2 The modified Mann-Whitney U-statistic 86
5.3 Simulation Studies 86
5.3.1 An Uncorrelated Model 87
5.3.2 A Correlated Model 88
5.3.3 Bias and Variance Study 88
5.3.4 Testing the Equality of Waiting Time Distributions between two groups of individuals 89
5.4 Application to the Spinal Cord Injury Data 91
5.5 Discussion 93
5.6 Tables 94
5.7 Figures 96
VI EXTENSIONS AND FUTURE RESEARCH 97
References 99
Appendix 107
CURRICULUM VITAE 110

CHAPTER 1: INTRODUCTION

1.1 **Correcting for hidden sources of heterogeneity in a Gene Expression Data**

Differential gene-expression analyses in microarray studies typically overlook the important aspect of subject specfic heterogeneity. Subjects in a microarray study can have certain plausible biological profiles which are not known to be connected with the primary outcome of interest and therefore, the subjects may not be matched with respect those profiles in a case-control study. For example, in an expression profiling study with cancer/non-cancer patients the main objective is to identify the genes that are differentially expressed between these two varieties, which can lead to the discovery of potential biomarkers related to cancer. But, this true picture of dierential expression can be blurred by several hidden biological eects specific to the subjectsrecruited in the study. It may happen that some genes are very highly expressed in the subjects with a certain biological, environmental or demographic profile (say, with high blood pressure, regular smoking habits, persons living in rural environments, persons sharing some hidden racial, familial or cryptic pattern pertaining to some inherent structure in the population, etc). On the other hand, some other genes may be repressed because of a similar reason. These factors distort the true signals of differential expression and introduce spurious effects of expression heterogeneity. Thus, many genes which are truly differentially expressed between the two varieties can get rendered as silent, whereas many others may be falsely detected as positives. To complicate things further, we can have a multitude of such hidden confounders in the study and their effects can also vary over different clusters of potentially correlated genes. Thus, it is

also not possible to get rid of them by simply modifying the arrays of gene-expression measures using a standard normalizing method. These difficulties pose serious problems in analyzing a gene-expression data and can lead to erroneous conclusions along with a substantial reduction in the power of the testing procedure.

Only a limited number of studies are available in this area which specifically address this issue of hidden variation in the context of gene-expression profiling. With the exception of Leek and Storey (2007), Scheid and Spang (2007) and Listgarten et al., (2010), most of the works in this area have considered specific types of confounding factors that can produce spurious signals of heterogeneity in the context of expression quantitative trait locus (eQTL) mapping. Stegle et al., (2008; 2010) have devised methods to improve the power of eQTL studies under the presence of non-genetic confounders (unobserved cell-culture conditions, batch eects, etc.). Yu et al., (2005), Kang et al., (2008a; 2008b), Listgarten et al., (2010) discuss the use of linear mixed effect models to correct for confounders from some unknown experimental effects or some hidden population structure. Price et al., (2006) proposed the use of principle component analysis (PCA) to correct for some hidden stratification in genome wide association studies. Scheid and Spang (2007) proposed a method using filtered permutations of the variety labels, which borrows information across the genes to identify and correct for unknown eects of the hidden confounders. Leek and Storey, (2007) introduced the Surrogate Variable Analysis (SVA) method and discussed its relevance in gene expression profiling analyses. This is treated as a benchmark technique in comparing the performance of our method. The method considers a singular value decomposition (SVD) of the residual matrix obtained after fitting a simple linear regression model to the log-transformed gene expression data. The significant eigenvectors from the SVD are

then used to create a reduced residual matrix (containing statistically significant traces of residual expression heterogeneity). The eigenvectors of the original residual matrix that are maximally correlated with the eigenvectors of this reduced matrix are taken as the surrogate variables. These variables are then used in the original linear model to test for the truly differentially expressed genes. Overall, the method is fairly complex and uses a two step process for the construction of surrogate variables. Moreover, the method in its current form, is also limited in terms of model selection, as it uses a very simple regression framework without considering the eects of each gene and its possible interactions with the surrogate variable (containing eects of the hidden confounders on potentially correlated genes and the two sample varieties). This reduces its applicability to situations where the effect of the hidden confounders can be far more complicated. In essence, all existing techniques in the literature address certain specific patterns of residual expression heterogeneity and discuss relevant modeling techniques to compensate for their effects. In this article we attempt to excavate the hidden sources of expression heterogeneity by the more generalized approach of partial least squares. The proposed method (SVA-PLS), due to its inherent principle, can perform the entire surrogate variable analysis from a more general perspective, by extracting the maximally correlated projections of the residual and original gene expression variables to two different latent factor spaces (connected by a linear relation), thereby ensuring an appropriate estimation of the hidden variables in terms of a set of orthogonal scores in the residual space. Also, our method considers a reasonably wide choice of models which can potentially explain a large variety of confounding effects.

1.2. **Multistate Models and Estimation of Stage Occupation Probabilities**

Multistate models (networks) are typically used to characterize the progression of a set of individuals through a succession of stages until they come to a certain endpoint (absorbing state). A simple example of such a model is the survival setup, where patients move from an initial stage (Alive) to an absorbing stage (Dead). Under a more complex scenario, there can be a number of intermediate phases (transient states) between the two terminal stages with a complicated chain of transtions interconnecting them. In such contexts, the fundamental quantities of interest are the transition counts of individuals moving from one stage to another and the number of individuals at risk of transition from a particular stage along the course of time inside the multistate model. Now, the exact evaluation of these two stochastic processes is practically hindered by the presence of several types of censoring. Under the present context censoring occurs when we loose track of the movement of an individual from a certain stage of the model. Under such a problematic situation, the basic objective of a statistical analysis is to estimate appropriately, the transition and at-risk processes, which can in turn lead to the derivation of the stage-occupation probabilities for the different states in the model. These probabilities are important from the aspect that, they give a precise idea about the chance of an individual occupying a certain state in the model, at a specific point of time. As a result, from a biological perspective, these probabilities can let us ascertain the extent to which an individual can be prone to the risk of transition from one stage of a disease to another, at a certain time point. In this way, we can visualize a clear pattern through which the disease spreads in the body, along with the progression of time, thereby enabling the development of appropriate medical interventions in order to resist it. As can be figured out already, the intended problem is quite complicated and it gets even worse, as the severity of censoring increases.

Over a fairly long period of time, several works involving parametric approaches have been done in this area by Lagakos, (1976); Beck, (1979); Kay, (1982); Sacks and Chiang, (1977); Wu, (1982); Klein et. al., (1984); Andersen and Keiding, (2002); Plevritis et. al., (2007) and others. Earlier theoretical works in this area can be traced back to Aalen and Johansen (1998), who developed a method to estimate the stage occupation probabilities for the different stages in the model, from the Nelson-Aalen estimators of their transition hazards, in case the data are subject to independent censoring (i.e, the when censoring mechanism operates independently of the state-to state transitions in the model). Datta and Satten (2001; 2002) proved the consistency of these estimators under a Non-Markovian structure and further extended their work for the situation when the data is subject to right-censoring (A form of censoring, where the individuals in the network are only followed upto a certain time point). The estimator of the integrated transition hazard proposed by them have a Nelson-Aalen form, where the estimated stochastic processes for state-to-state transitions and at risk of transition from a particular state, are represented by an Inverse Probability of Censoring Weighted (IPCW) version of the corresponding original unobserved counting processes.

Lin et. al., (1999) considered a nonparametric estimation of the waiting time (sojourn times) distributions for the different states in a progressive multistage model with no branching. Methods for non-parametric estimation of waiting time distributions have also been developed by Satten and Datta, (2002), under the situation of dependent censoring, by using the Aalen's linear hazard model (Aalen, 1980). While the estimation of the marginal distributions for these stochastic processes has been cultivated in the literature for quite a long period of time (Satten and Datta, 2002), methods for estimating their corresponding conditional

distributions have recently come under the spectrum. This mode of estimation, in fact, enables the use of different forms of available covariate information on the individuals in the model (which can potentially affect their movements from one stage to another), through a wide variety of elegant regression techniques. Previous works dealing with this approach, have mostly considered regression frameworks by modeling the state-to-state transition hazards with a set of available covariates, by using the Cox's proportional hazards model (Cox, 1972) in a Markovian structure (Andersen et. al., 1993). Application of the additive linear hazards model (Aalen 1989; Lin and Yang, 1994) in Multistate systems under a Markovian framework, can be found in the works of Shu and Klein, (2005). Incorporation of Hazard models like the Aalen's additive hazard model, enable the covariates to cast a highly complicated effect on the estimates of the state occupation probabilities in a Multistate model. A brief overview of the semi parametric approaches in this area, can be found in Andeson and Keiding, (2002). In reality, a time-to-event data setting can either be a simple survival setup with just two states or a more complex multistage model with a large number of states. In such contexts parametric or semi-parametric methods of estimation are mostly based on certain structural assumptions that may not always fit the actual data generating mechanism to a reasonable extent. Moreover, for a complex multistage network, with a large number of transition paths inter-connecting them, formulation of an applicable parametric/semi-parametric methodology becomes intensely difficult, thereby making the development of efficient non-parametric estimation techniques almost indispensable. Doksum and Yandell, (1982) illustrated this point by comparing semi-parametric vs non-parametric estimators using the widely popular Stanford Heart Transplant Data (Crowley and Hu, 1977). Only a limited number of works are available in the field of non-parametric regression modeling of Multistate time-to-event data. Most of them have considered the simple survival setting. Beran (1981) discussed the use of a

conditional Kaplan-Meier estimator using weights obtained either from a nearest neighbor approach or a Kernel method. Doksum and Yandell, (1982) considered a non-parametric alternative to this problem. Extensive studies on the theoretical properties of these estimators have been pursued further in Dabrowska (1987, 1989), Li and Doss, (1995), McKeague and Utikal, (1990), Li and Datta, (2001) and others. Andersen and Klein, (2006) illustrated the use of covariates in a multistate model, by using a combination of non-parametric and semi-parametric techniques. But, their method may not provide estimates of the marginal quantities of interest, in the model. Smoothing methods provide powerful non-parametric alternatives to such estimation problems, although the selection of their underlying tuning parameter (characterizing the extent to which it is designed to fit the data), is a common drawback, with no full-proof solution developed so far.

Recently, Mostajabi and Datta, (2012) used a covariate based kernel smoothing method that estimates the state-to-state transition counts and at-risk number of individuals in a progressive multistate model with right censoring, using the Inverse probability of Censoring (IPCW) principle (Datta and Satten, 2001). In the present work we attempt to conditionally estimate the state occupation probabilities in a multistate model by using additional covariate information on the individuals with a IPCW version of the backfitting regression technique (Hastie and Tibshirani, 1990).

1.3 **Formulation of modified Mann-Whitney U-statistics in Right Censored Multistate Model Data**

The Mann-Whitney U test, which is technically equivalent to the Wilcoxon rank sum test (Mann and Whitney, 1947; Wilcoxon 1945), is a widely popular non-parametric method for comparing two distributions based on independent samples.

The test developed initially by Wilcoxon for equal sample sizes (Wilcoxon, 1945), was further extended by Mann and Whitney, (1947) for the case of unequal samples. It is very useful when the underlying assumption of normality required for the use of a standard two sample parametric t test is not justified. In addition, this test can also be applied to situations when the data is measured in an ordinal scale. Although some extension of the traditional rank tests for situations when the data is subject to censoring have been proposed their use to compare waiting times lead to incorrect size due to induced dependent censoring. Recently Fan and Datta (2013) developed a Inverse Probability of Censoring (IPCW) weighted version of the Mann-Whitney type U statistics for solving the problem of testing the equality of waiting time distributions for two groups of individuals in a multistate model, when the state-to-state transitions of the individuals are subject to right censoring (i.e, the individuals are only followed upto a certain time point till they are censored). Their formulation conceptually provides a marginal comparison of the waiting time distributions in the two groups. Now, in several real life multistate models (like competing risk models, disease progression models, etc.), we have additional information on different covariates/predictors for the individuals. This information, if used effectively under the proposed construct, can provide a clear idea on the disparity between the state waiting time distributions for individuals from the two groups, who share a common range of variation with respect to their corresponding covariate values. For example, in a multistate disease progression model, we may be interested to know whether the distributions of the waiting time (sojourn time) between two specific stages of the disease differ between the individuals coming from the two separate groups, but falling in a particular common age bracket. Here age is a covariate and the idea is to perform a conditional comparison of the waiting time distributions in the two groups, given the information on ages of the individuals. A conditional analysis provides the scope of applying elegant regression

techniques based on these covariates and substantially improves the quality of inference. In the third project we propose a methodology which performs two separate regressions of the state waiting times for the two groups, on their respective individual covariate values, by fitting an Accelerated Failure Time (AFT) model and uses the corresponding two sets of residuals to build an IPCW version of the Mann-Whitney type U statistic. This statistic is then used to test the equality of waiting time distributions in the two groups,after adjusting for the available covariates.

CHAPTER 2: SURROGATE VARIABLE ANALYSIS USING PARTIAL LEAST SQUARES IN GENE EXPRESSION STUDIES

2.1 **Motivation**

In a typical gene expression profiling study, our prime objective is to identify the genes that are differentially expressed between the samples from two different tissue types. Commonly, standard ANOVA/regression is implemented to identify the relative effects of these genes over the two types of samples from their respective arrays of expression levels. But, this technique becomes fundamentally flawed when there are unaccounted sources of variability in these arrays (latent variables attributable to different biological, environmental or other factors relevant in the context). These factors distort the true picture of differential gene expression between the two tissue types and introduce spurious signals of expression heterogeneity. As a result many genes which are actually differentially expressed are not detected, whereas many others are falsely identified as positives. Moreover, these distortions can be different for different genes. Thus, it is also not possible to get rid of these variations by simple array normalizations. This both-way error can lead to a serious loss in sensitivity and specificity, thereby causing a severe inefficiency in the underlying multiple testing problem. In this work, we attempt to identify the hidden effects of the underlying latent factors in a gene-expression profiling study by Partial Least Squares (PLS) and apply ANCOVA technique with the PLS-identified signatures of these hidden effects as covariates, in order to identify the genes that are truly differentially expressed between the two concerned tissue types.

2.2 **Methods**

We consider a gene-expression profiling analysis with g genes and n subjects, distributed over two tissue types/varieties (like, normal and cancer cell lines or two different biological conditions). Let the first n_1 subjects be under variety 1 and the rest n_2 be under variety 2. We start by applying the standard ANOVA technique on the log-transformed gene expression matrix Y (Kerr et. al., 2000, Kerr and Churchill, 2001, Wolfinger et. al., 2001 and Kerr et. al., 2002) and compute the fitted model residuals. Let Y_{ijk} denote the log-transformed gene expression value for the gene i in subject k under variety j, $i = 1, 2 \cdots g$, $j = 1, 2$ and $k = 1, 2 \cdots n_1$ for $j = 1$ and $k = n_1 + 1, n_1 + 2 \cdots n_2$ for $j = 2$. We fit the following ANOVA model to the data and get the residuals.

$$Y_{ijk} = \mu + G_i + V_j + (GV)_{ij} + \epsilon_{ijk}$$

where μ denotes the general mean effect in the model, G_i, V_j respectively stand for the main effects of gene i and variety j and $(GV)_{ij}$ defines their mutual interaction (characterizing the expression effect of gene i on the subjects under variety j). ϵ_{ijk} denotes the random error term which is assumed to follow a $N(0, \sigma^2)$ distribution.

The fitted residuals from the above model are then given by $e_{ijk} = Y_{ijk} - \overline{Y}_{ij0}$, where , A_j being the set of individuals corresponding to variety j. These residuals may contain the traces of subject-specific expression heterogeneity, which is independent of the primary variable signal from the sample types and can confound the true effect behind many potentially positive genes or can overestimate many silent genes as positives. In order to extract these spurious differential expression signals we employ the Partial Least Squares (PLS) technique (Wold, (1975; 1985), Helland, 1999). We construct two $n \times g$ matrices Y and E,

whose ith column contains respectively, the original log-transformed gene expression levels Y_{ijk} s and the residual gene expression levels e_{ijk} s for all the n individuals corresponding to gene i ($i = 1, 2 \cdots g$). Thus, E=$((E_{rc}))_{n \times g}$ and Y=$((Y_{rc}))_{n \times g}$, $r = 1, 2 \cdots n$ and $c = 1, 2 \cdots g$. Conceptually, these matrices can be characterized as two sets of n observations on two g-dimensional random variables E and Y, where each dimension corresponds to a certain gene.

Our approach is to regress E on Y by partial least squares, in order to extract the hidden sources of gene expression heterogeneity. PLS, by virtue of its dimension reduction and covariance maximizing property extracts the additional signals from those groups of genes, whose expression levels, contained in the original gene-expression matrix Y, are influenced by the hidden subject specific effects, contained in the residual gene-expression matrix E. Let the matrices now stand for their respective mean zero versions, obtained by subtracting the respective column means from their initial versions. We assume that $E^T Y$ is non-null. The statistical regression model for PLS can be written as

$$E = UQ^T + \epsilon_1 \tag{1}$$

$$Y = TP^T + \epsilon_2 \tag{2}$$

where $U = [u_1 : u_2 : \cdots u_m]$ is an $n \times m$ matrix, containing the m latent factors $u_1, u_2 \cdots u_m$ in the space of the response matrix E. Similarly $T = [t_1 : t_2 : \cdots t_m]$ is another $n \times m$ matrix containing the m latent factors $t_1, t_2 \cdots t_m$ in the space of the covariate matrix Y. $Q = [q_1 : q_2 : \cdots q_m]$ is a $g \times m$ matrix consisting of the loadings $q_1, q_2 \cdots q_m$, which measure respectively the importance of the latent factors $u_1, u_2 \cdots u_m$ in the response Es space. Similarly $P = [p_1 : p_2 : \cdots p_m]$ is a $g \times m$ matrix consisting of the loadings $p_1, p_2 \cdots p_m$, which measure respectively the importance of

the latent factors $t_1, t_2 \cdots t_m$ in the response Ys space. Further for each$i = 1, 2 \cdots m$, $u_i = (u_{i1}, u_{i2} \cdots u_{in})^T$, $t_i = (t_{i1}, t_{i2} \cdots t_{in})^T$, $q_i = (q_{i1}, q_{i2} \cdots q_{ig})^T$, $p_i = (p_{i1}, p_{i2} \cdots p_{ig})^T$. Here ϵ_1 and ϵ_2 are the random error matrices characterizing the residual terms in the regression models for E and Y respectively.

Now, the basic idea of partial least squares is to estimate the set of latent factor pairs $(u_1, t_1), (u_2, t_2) \cdots (u_m, t_m)$ one by one, along with the corresponding deflation of the matrices E and Y at each step. This is executed by a process of alternating regression. For each latent factor pair $(u_i, t_i), i = 1, 2 \cdots m$ this procedure finds weight vectors c and w in such a way that the covariance of u_i and t_i is maximized. Specifically, c and w are such that

$$[cov(u_i, t_i)]^2 = [cov(Ec, Yw)]^2 = max_{|r|=|s|=1}[cov(Ec, Yw)]^2 \qquad (3)$$

We initialize $E_1 = E$ and $Y_1 = Y$. Now for $i = 1, 2 \cdots m,$ we successively estimate the i-th latent factor pair $(u_i, t_i), i =$ by the partial least squares (PLS) algorithm presented below (see., e.g., Abdi, 2003; Rosipal and Krämer, 2006). In this algorithm we use $a \propto b$, to mean $a = \frac{b}{|b|}$, for any vector b.

We start by setting $u_i = E_{i,.v},$ where $v = argmax_c \sum_{r=1}^{n} E_{i,rc}^2 \quad , t_{i,old} = (0, 0 \cdots 0)$ and repeat steps (i) to (iv) till convergence (as defined in Step (v)):

(i) Regress Y_i on u_i to obtain $w_i \propto Y_i^T u_i$.

(ii) Compute the updated i-th Y space latent factor $t_i = Y_i w_i$.

(iii) Regress E_i on t_i to obtain $c_i \propto E_i^T t_i$.

(iv) Compute the updated i-th E space latent factor $u_i = E_i c_i$.

(v) If $\sum_{j=1}^{n} |t_{ij} - t_{ij,old}|/t_{ij} < \epsilon$, STOP; otherwise let $t_{i,old} = t_i$ and go back to step (i)

Throughout we have used $\epsilon = 10^{-8}$.

Next deflate the matrices E_i and Y_i to obtain $E_{i+1} = E_i - t_i b_i^T$ and $Y_{i+1} = Y_i - t_i p_i^T$ where, $b_i = E_i^T t_i / t_i^T t_i$ and $p_i = Y_i^T t_i / t_i^T t_i$. E_{i+1} and Y_{i+1} are now used in place of E_i and Y_i to

extract the $i+1$-th latent factor pair (u_{i+1}, t_{i+1}). In this way we find the m latent factors from the E and Y spaces. The use of t in deflating both the response (E) as well as covariate (Y) matrices ensures orthogonality of the extracted latent factors $t_1, t_2 \cdots t_m$ in the Y-space, which in turn ensures their estimability in a linear model. From now onwards, we denote m by p_{max} to define the maximum number of hidden(latent) factors (scores) that are needed to be extracted from the two spaces. The p_{max} Y-space scores extracted by the above method can be characterized as a set of surrogate variables $Z^1, Z^2 \cdots Z^{p_{max}}$ that are optimally associated with the latent factors from the E-space, containing the hidden sources of expression heterogeneity in the original gene expression data. The mutual covariances between the extracted latent factors from the two spaces decrease gradually from the first pair (u_1, Z^1) to the p_{max}-th pair $(u_{p_{max}}, Z^{p_{max}})$. Thus Z^1 contains maximum information on the residual gene-expression heterogeneity compared to the other factors. Now, we define a series of ANCOVA models M_p indexed by $p = 1, 2 \cdots p_{max}$, where p_{max} denotes number the surrogate variables incorporated in the model, which capture effects of the residual gene-expression heterogeneity.

$$M_p : Y_{ijk} = \mu + G_i + V_j + (GV)_{ij} + W_{ijk}^{imp} + \epsilon_{ijk} \quad (4)$$

where μ denotes the general mean effect in the model G_i, V_j and $(GV)_{ij}$ denote respectively, the

main effects of gene i, variety j and their mutual interaction effect.

$W_{ijk}^{imp} = \sum_{l=1}^{p} \beta_l Z_{jk}^l + (GZ^1)_i Z_{jk}^l + (VZ^1)_j Z_{jk}^1$ is incorporated in the model as the PLS-imputed estimate of the hidden residual expression heterogeneity in the data. Here, β_l is the regression coefficient for Z^l in the ANCOVA model (4). $(GZ^1)_i$ and $(VZ^1)_j$ define respectively, the interaction effects of gene i and variety j with the first surrogate variable Z^1. These effects measure respectively, the variation in the impact of the hidden factors (captured by Z^1) over different groups of genes (which may be correlated) and over the two tissue types (which may affect the primary variable signal). As the first surrogate variable Z^1 contains maximum information on the residual expression heterogeneity compared to the other ones we consider only its interactions with the gene and variety effects. The inclusion of these effects in the model ensures accurate estimation of the actual gene-variety interactions, capturing the true expression effects of a gene over the two varieties, if potential hidden variables are embedded in the data structure. ϵ_{ijk} denotes the random error term corresponding to Y_{ijk} in the model, which is assumed to follow a $N(0, \sigma^2)$ distribution. Here p_{max} can be specified by the user, considering the corresponding situation under study and affording a reasonable degree of complexity along with a manageable computational intensity. As for our purpose, we have selected $p_{max} = 3,$ since from several empirical studies (details reported in the supplementary website) we have found that the first three surrogate variables $(Z^1, Z^2$ and $Z^3)$ explain a substantial proportion of the dispersion for the variable

E. Thus, overall we consider three different linear models from which the best is selected by the Akaike's Information Criterion (AIC) (Hirotsugu, (1974; 1980)) and is then used to test for the equality of gene-variety interaction effects for identifying the truly differentially expressed genes. In the concerned multiple testing problem, we use the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to control the false discovery rate. The entire algorithm for our method SVA-PLS is presented below:

Step 1: Fit the standard ANOVA model (1) to the log-transformed gene expression data Y and calculate the fitted residual matrix E.

Step 2: Regress E on Y by partial least squares and extract p_{max} (user-specified) linear combinations (scores) from their respective latent factor spaces.

Step 3: Incorporate, one by one, the p_{max} scores in the Y-space as surrogate variables along with the gene and variety interactions of the first PLS score in model (1) to develop a series of p_{max} new linear models (4) .

Step 4: Compare AIC's for the models to select the best out of them (the model corresponding to the minimum AIC) and denote its corresponding number of surrogate variables by p_{opt}.

Step 5: Fit model $M_{p_{opt}}$ to estimate the actual gene-variety interaction effect $(GV)_{ij}$ for each gene i and variety j $(i = 1, 2 \cdots g$ and $j = 1, 2)$. For each gene i test the null hypothesis of no variety-specific differential expression $H_0 : (GV)_{i1} = (GV)_{i2}$ vs the alternative hypothesis of differential expression $H_1 : (GV)_{i1} \neq (GV)_{i2}$, using the statistic t_i defined below:

$$t_i = \frac{(GV)_{i1} - (GV)_{i2}}{\sqrt{\sigma^2 \widehat{V}((GV)_{i1} - (GV)_{i2})}}$$

which under H_0 follows a central t distribution with $\nu = ng - 3g - p_{opt}$ df and the corresponding p-value is $2(1 - F_\nu(t))$, $F_\nu(t)$ being the distribution function for a central t distribution with ν df.

$(GV)_{ij}$, $j = 1, 2$ is the least squares estimate of $(GV)_{ij}$, $\widehat{V}((GV)_{i1} - (GV)_{i2})$ is the estimated variance of $(GV)_{i1} - (GV)_{i2}$ and $\widehat{\sigma}^2$ is the least squares estimated variance of σ^2, all computed from the model $M_{p_{opt}}$.

Step 6: Perform a multiple testing with these p-values for identifying the truly differentially expressed genes at a prespecified level of the false discovery rate (FDR), using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

2.3 **Simulation Studies**

We envisage a gene expression profiling study with 500 genes and 20 subjects, distributed equally over 2 varieties. The entire simulation study is broadly divided into two settings: (1) assuming the genes to be independent of each other (Independent) and (2) assuming dependence within different groups of genes (Clustered).

The log-transformed gene-expression values (Y) are generated by using a linear model with the gene (G) and variety/tissue type (V) main effects, their interaction (GV) and a hidden variable (W). Thus, Y_{ijk} corresponding to the i-th gene, j-th variety and k-th subject, is obtained as:

$$M_p : Y_{ijk} = \mu + G_i + V_j + (GV)_{ij} + W_{ijk} + \epsilon_{ijk} \tag{5}$$

where $i = 1, 2 \cdots 500$ denote the 500 genes, $j = 1, 2$ denote the two varieties and $k \in A_j$, denote the 20 subjects in the study. Here $A_1 = \{1, 2 \cdots 10\}$ and $A_2 = \{11, 12 \cdots 20\}$ denote respectively, the subsets of individuals corresponding to the two varieties 1 and 2. The error terms ϵ_{ijk} s are assumed to be independently distributed as $N(0, \sigma^2)$, where choice of σ^2 is described next.

Let X denote the design matrix corresponding to the above linear model, β denote the corresponding vector of regression coefficients and ϵ denote the vector of the corresponding random error terms. Then we have $Y = X\beta + \epsilon$. Define the Noise to Signal ratio (η) as $\eta = \sigma^2 / \beta^T V(X) \beta$ (σ^2 being the random error variance and $\beta^T V(X) \beta$ being the variance of the signal $X\beta$ generating the actual gene expression levels). This quantity measures the relative intensity of the noise coming from the random error and the confounded primary variable signal depicting the expression effect of the genes over the two varieties. We consider three different value $0.1, 0.5$ and 1 for η to incorporate respectively, the cases of strong, moderate and weak primary signal intensity. From these choices of the η we compute the corresponding values of σ^2 and use them to simulate the values of ϵ_{ijk} s in the model (5).

For data generation, we assume the effects of all terms except the gene-variety interaction (GV) and the hidden confounder (W) to be zero. Overall we consider the first 70 genes to be truly differentially expressed among all the 500 genes. For $1 \leq i \leq 20$ we take $(GV)_{i1} = -3$, $(GV)_{i2} = 3$, for $21 \leq i \leq 70$

$(GV)_{i1} = 3$, $(GV)_{i2} = -3$ and for genes 71 to 500 we assume $(GV)_{i1} = (GV)_{i2} = 0$.

For each gene i in 1 to 500, $j = 1, 2$ and subject $k \in A_j$, we generate a Bernoulli random variable s_{ijk} with success probability 0.4. It is used to generate effects of W over the two varieties, under both the independent as well as clustered settings. Biologically, this accounts for hidden confounding effects from certain specific subjects under each of the two varieties, which is typically expected in a real-life gene-expression analysis. In addition, we consider two separate scenarios, depending on whether the effect of the hidden variable W is same or different over the two varieties.

2.3.1 **Independent Tests**

In this setting, we consider the genes to be independent of one another. We generate their log-transformed expression levels under two scenarios of similar and varying effects of the hidden variable over the two varieties, respectively.

The similarity in the effects of the missing variable over the two varieties is accomplished by simulating the latent variable W_{ijk} from the same normal distribution for $k \in A_1 \cup A_2$ (covering subjects from both the varieties). The effect of W is varied over three different groups of genes by changing the mean parameter of its distribution. That is, we let $W_{ijk} = Z_{ijk} I(s_{ijk} = 1)$, where Z_{ijk} is generated from $N(-3, 0.01)$ or $N(2, 0.01)$ or $N(20, 0.01)$, depending on whether $1 \leq i \leq 20, 21 \leq i \leq 70$ or $i > 70$.

For generating different effects of the hidden variable W_{ijk} over the two varieties we simulate the latent variable for the subjects $k \in A_1$ and $k \in A_2$, from two normal distributions with different means. Once again, the effect of the hidden

variable is varied over the three gene groups. That is, we let $W_{ijk} = Z_{ijk}I(s_{ijk} = 1)$, where for $k \in A_1, Z_{i1k}$ is generated from $N(-3, 0.01), N(2, 0.01)$ or $N(-3, 0.01)$ and for $k \in A_2, Z_{i2k}$ is generated from $N(3, 0.01), N(15, 0.01)$ or $N(3, 0.01)$, depending on whether $1 \leq i \leq 20, 21 \leq i \leq 70$ or $i > 70$.

Next, we consider yet another simulation setting, where the hidden variable results in a complex confounding pattern with the varieties. In this case, for each variety j, we simulate the latent variable W_{ijk} for the subjects $k = 1, 2 \cdots 10j/2$ and $k = 10j/2 + 1, 10j/2 + 2 \cdots 10j$, from two normal distributions with different means. Similar to the previous settings, we vary the effect of W over the three gene groups. Thus, under the variety $j (j = 1, 2)$ we let $W_{ijk} = Z_{ijk}I(s_{ijk} = 1)$, where for $k = 1, 2 \cdots 10j/2$ Z_{ijk} is generated from $N(-3, 0.01), N(2, 0.01)$ or $N(-3, 0.01)$ and for $k = 10j/2 + 1, 10j/2 + 2 \cdots 10j$, Z_{ijk} is generated from $N(3, 0.01)$ or $N(15, 0.01)$ or $N(3, 0.01)$, depending on whether $1 \leq i \leq 20, 21 \leq i \leq 70$ or $i > 70$.

2.3.2 **Cluster Dependent Tests**

Note that the usual statistical model for differential gene analysis by the ANOVA formulation assumes independent error terms. However, it is well known that in reality, certain groups of genes have correlated expressions. In this setting, we consider 3 clusters of correlated genes with the same gene-variety interaction effects as for the case of independently expressed genes, with the hidden variable (W) being generated according to the same set of simulation schemes as in Section 2.3.1.

The underlying dependence among the genes is incorporated by generating the random error term ϵ_{ijk} in the model (5) as a weighted sum of two different errors ϵ^1_{ijk} and ϵ^2_{ijk}, simulated independently of each other, with the values of ϵ^1 being same for all the genes in the same cluster. Let

$C = (1, 2 \cdots 20; 51, 52 \cdots 70; 461, 462 \cdots 500)$ denote the union of the 3 clusters. Then, mathematically the generation of ϵ_{ijk} in the simulation model (5) is expressed as:

$$\epsilon_{ijk} = \begin{cases} \frac{1}{\sqrt{2}}\epsilon^1_{I(i),j,k} + \frac{1}{\sqrt{2}}\epsilon^2_{i,j,k} & if\ i \in C \\ \epsilon^2_{i,j,k} & if\ i \notin C \end{cases}$$

where, $I(i)$ denotes the cluster containing gene i. The random error terms $\epsilon^1_{I(i),j,k}$ and $\epsilon^2_{i,j,k}$ are generated from independent $N(0, \sigma^2)$ distributions (σ^2 being determined from the desired Noise to Signal ratio, as before). From a biological perspective this simulation setting captures the idea that genes in the same cluster act cooperatively, resulting in correlated expression measurements.

The simulation study is concerned with a performance analysis of standard ANOVA, our method SVA-PLS and SVA with respect to four measures: sensitivity, specificity, false discovery rate (FDR) and false non-discovery rate (FNR).

Sensitivity: proportion among differentially expressed genes that were declared significant.

Specificity: proportion among non-differentially expressed genes that were declared non-significant.

False discovery Rate (FDR): proportion among genes declared significant that were not differentially expressed.

False non-discovery rate (FNR): proportion among genes declared non-significant that were differentially expressed.

The entire simulation study is performed 100 times under each scenario in order to compute the average values of the four performance measures. Under the clustered setting the SVA software broke down at several iterations of the simulation study. Hence, for this setting we only report the performance of our method (SVA-PLS) and the standard ANOVA. The detailed results are reported in Tables 2.6.1 to 2.6.6.

From the performance analysis of the three methods on the independent gene expression levels, with similar, varying and complex effects of the hidden variable $(\text{Tables 2.6.1, 2.6.2 and 2.6.3, respectively})$, we see that SVA-PLS achieves the highest sensitivity compared to standard ANOVA and SVA. Interestingly, the margin of sensitivity for our method is very high in the case of complex confounding (Table 2.6.3), followed by the case of varying effects of the hidden variable over the tissue types (Table 2.6.2). This observation demonstrates that our method is most useful for the relatively complicated situations, when the missing variable is in fact a statistical confounder affecting the primary variable signals from the two tissue types. In addition, our method produces a comparatively impressive performance with respect to the other two methods in terms of the high specificity and reasonably small False Discovery and Non-Discovery rates. Under the clustered setting with dependence inside several clusters of genes $(\text{Tables 2.6.4, 2.6.5 and 2.6.6 for the moderate case of } \eta = 0.5)$ SVA-PLS performs really well compared to standard ANOVA by detecting a larger number of truly positive genes with its high margin of sensitivity, at the cost of a slightly increased false discovery rate (FDR), which is an obvious price to pay for achieving a higher performance in terms of detection power. For the other choices of the η too, SVA-PLS shows higher sensitivity compared to standard ANOVA. Under this setting also, our method yields

a reasonably high specificity in comparison to standard ANOVA along with an impressively small FDR and FNR. Specifically, the margin of sensitivity for SVA-PLS under both the simulation settings is the highest in the best case with very strong primary variable signal $(\eta = 0.1)$, closely followed by the moderate $(\eta = 0.5)$ and worst cases $(\eta = 1)$. Thus, overall the results demonstrate that our method, by virtue of its high sensitivity in a wide variety of situations can potentially discover many truly differentially expressed genes that are masked by the effects of hidden factors and can simultaneously maintain acceptably small error rates.

We further illustrate the efficacy of our method by comparing the actual (mean centered) values of the hidden variable W_{ijk} (simulated in the model under the setting of independently expressed genes with serious confounding of the hidden variable W), with the PLS-imputed values W_{ijk}^{imp} incorporated in the ANCOVA model (4). We observe a strongly linear relationship between the two sets of values with a very high positive correlation (0.96) (see Figure 2.7.1). We have noticed a similar effect in the other simulation settings as well (refer to the supplementary website). This demonstrates that our method SVA-PLS is effectively imputing the hidden variable (W) on the actual expression levels of the genes.

2.4. **Analysis of Leukemia Data**

We now explore the performance of our method on a dataset generated from a gene expression study of acute megakaryoblastic leukemia (AMKL), which is a subtype of the disease acute myeloid leukemia (AML). The dataset was featured in Bourquin et. al., (2006). It contains the expression levels of 22283 genes on two types of AMKL patients, 23 with down-syndrome and 38 without down-syndrome.

In general, down-syndrome patients are more prone to AMKL compared to those without it and treatment outcomes are also much more favorable for them.

From an exploratory analysis of the data set (Bourquin et. al., 2006), it was found that the non-DS AMKL patients can be further subdivided into two groups using the expression profiles of the HOX/TALE family members. This latent grouping inside one tissue type can generate hidden confounders, which may in turn perturb the actual signals of variety-specific differential gene expression. Thus it is important to search for the traces of residual gene expression heterogeneity in this dataset for ensuring a more accurate inference on the truly positive genes, which is built into our method. Indeed, we investigated whether the PLS imputed values W^{imp} of the three genes, HOXA9, HOXA10 and MEIS1, belonging to the HOX/TALE family, contain a subgroup signature. Figure 2.7.2 shows a heat map for the normalized values of the estimated PLS contributed part W^{imp}, corresponding to the 38 individuals in the non-DS AMKL group. This W^{imp} is free from the primary signal of variety specific differential expression and is expected to contain the traces of residual expression heterogeneity corresponding to the hidden factors in the data. From Figure 2.7.2, we can observe a sub-group structure amongst these individuals. Clearly, the differential pattern is strongest for the MEIS1 gene, followed by HOXA9 and HOXA10.

The three methods SVA-PLS standard ANOVA and SVA were applied to the log-transformed expression matrix of the 22283 genes in the dataset. Overall, SVA-PLS detected 1585 genes followed by 1407 genes from standard ANOVA and 280 genes from SVA (see Figure 2.7.3). Our method detects a total of 427 genes, that are missed by others, of which at least 6 genes deserve special mention. These genes are MLF1, BRCA2, TNF, c-MPL, CD44 and MAGE-D4.

The gene MLF1 is actively involved in the development of acute myeloid leukemia (AML). A chromosomal derangement associated with this gene is a cause

of the myelodysplastic syndrome (MDS) (Block et. al., 1953). Patients with this syndrome often develop acute anemia, which in most cases lead to low blood counts. In almost one-third of the patients, this syndrome causes progressive bone marrow failure, which in turn develops the disease into AML. Also, delayed bone marrow transplantation for patients with low risk of the myelodysplastic syndrome has been found to be connected with improved outcome (Cutler et. al., 2004).

The gene BRCA2 is an important caretaker gene (Kinzler and Vogelstein, 1997), whose inactivation initiates a tumor and the resulting genetic instability causes accelerated mutation in all genes, which in turn, may lead to the rapid progression of the tumor. Germline mutations in this gene play a dominant role in the onset of breast and ovarian cancer, pancreatic cancer, prostate cancer, Fanconi anemia and pre-B-cell acute lymphoblastic leukemia (Lancaster et. al., 1996, Murphy et. al., 2002, Ozcelik et. al., 1997, Narod et. al., 2008, Wagner et. al., 2004).

The apoptosis-inducing ligand TRAIL related to the gene TNF, plays an active role in the development of different types of cancers. Down-regulation of TRAIL-R2 inhibits the TRAIL mediated apoptosis in acute myeloid leukemia (AML) (Riccioni et. al., 2005). Monoallelic deletion of the tumor suppressing genes TRAIL-R1 and TRAIL-R2 can inactivate the TRAIL-induced apoptosis in B-cell lymphoma (Rubio-Moscardo et. al., 2005). In the development of colorectal cancer, there is a substantial increase in sensitivity to TRAIL-induced apoptosis, with the progression from benign to malignant tumors (Haque et. al., 2005).

Expression of the gene c-MPL has been found to be involved in the progression of CD34+ and M2FAB subtypes of acute myeloid leukemia (AML) (Ayala et. al., 2009).

Ligation of the gene CD44 with specific anti-CD44 antibodies (or with its natural ligand hyaluronan) can reverse the blockage in the differentiation of several subtypes of acute myeloid leukemia (AML), thereby improving the survival of patients using differentiating agents (e.g., retinoic acid) (Charrad et. al., 1999). The $8:21$ chromosomal translocation is commonly observed in acute myeloid leukemia (AML). Acute myeloid leukemia-1 transcription factor AML1-ETO and its splice variant AML1-ETO9a are capable of modulating the expression of CD44, thereby connecting the abnormal translocation 8:21 to the regulation of a cell adhesion molecule, that is involved in the nurturing of AML blast/stem cells (Peterson et. al., 2007). In the acute promyelocytic leukemia cell line NB4, over-expression of the gene CD44 receptor results in apoptosis (Abecassis et. al., 2008). In addition, down-regulation of this gene has been found to be conducive to keratoacanthoma and squamous cell carcinoma (Tataroglu et. al., 2007).

Upregulation of the gene MAGE-D4 results in the proliferation of tumor cells in non-small cell lung cancer (NSCLC) (Ito et. al., 2006).

Thus, we find that a number of the additional genes selected by our method are connected to acute myeloid leukemia or some other related type of carcinoma. These genes being found to be differentially expressed between the subjects with and without down-syndrome, can serve as important candidates for research on leukemia and down-syndrome.

2.5 **Discussion**

Hidden array-specific (subject-specific) factors in microarray analyses may constitute a substantial source of gene-expression heterogeneity. The effects of

these factors are not detectable from outside and also can't be removed by any standard normalizing method. But they can perturb the primary signals of differential gene expression and lead to erroneous conclusions on the detection of differentially expressed genes.

This problem is relatively unexplored in gene expression studies. In this paper, we have developed a novel technique for identifying these latent factors by using partial least squares and applied it to a wide variety of simulation settings characterizing different patterns of viable gene-expression profiling studies. We have shown that the technique of partial least squares, by virtue of its basic principle of projecting to latent structures, can produce precise estimates of the hidden factors causing the spurious signal heterogeneity. These estimates (surrogate variables) when incorporated in the ANOVA model enhances detection of the gene-variety interaction effects thereby leading to a large gain in sensitivity of the underlying bioinformatics screening procedure. The resulting method, SVA-PLS, also yields a reasonably high specificity for a wide range of data structures, thereby ensuring an efficient control over the incorrect detection of many silent genes. The false discovery rate is marginally higher for our method, but is sufficiently well compensated by a substantially large gain in the margin of sensitivity. Overall, SVPLS emerges as the winner when compared with two other competing methods in a range of controlled settings. The utility of our method in detecting potentially interesting genes missed by other methods is also demonstrated by an analysis of a real data set on AMKL patients.

Unaccounted sources of variation (hidden variables) in a model can adversely affect the outcomes of statistical tests. This is particularly true if the unmeasured variables are confounders, i.e., correlated with the variables in the

model whose effects on the outcomes are being tested. In a simple two group comparison, a standard assumption for the validity of the commonly used two sample pooled t-test is that the error variances in the two groups (populations) are equal. A departure from this model assumption is known as the Behrens-Fisher problem and has received a great deal of assumption in the statistics literature (see, e.g., Lehmann, 1986). A common solution to this problem is to use separate variance estimates for the error distribution in two groups and resort either to an approximate t-distribution (Welch, 1938) or to a large sample normal approximation of the distribution of the test statistics. Indeed, if one assumes (as in our formulation) the existence of an unmeasured factor contributing to the outcome in a linear model formulation of the two sample problem that is equated with the model errors, one gets a model with unequal variances in the two groups. Thus, our method may provide an alternative solution to the Behrens-Fisher problem. We plan to explore this connection in greater details elsewhere.

2.6 **Tables**

Method	Sensitivity	Specificity	FDR	FNR
	$\eta = 0.1$			
Std. ANOVA	0.319	1	0.000	0.059
SVA-PLS	1	0.986	0.078	0.000
SVA	0.999	0.993	0.040	0.000
	$\eta = 0.5$			
Std. ANOVA	0.154	1	0.000	0.078
SVA-PLS	0.813	0.993	0.047	0.029
SVA	0.640	0.995	0.045	0.055
	$\eta = 1$			
Std. ANOVA	0.104	1	0.000	0.071
SVA-PLS	0.370	0.997	0.045	0.091
SVA	0.194	0.998	0.041	0.109

Table 2.6.1 Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of independently expressed genes with similar effects of the hidden variable over the two varieties.

Method	Sensitivity	Specificity	FDR	FNR
	$\eta = 0.1$			
Std. ANOVA	0.176	1	0.000	0.031
SVA-PLS	0.926	0.990	0.061	0.012
SVA	0.290	0.997	0.045	0.035
	$\eta = 0.5$			
Std. ANOVA	0.086	1	0.000	0.034
SVA-PLS	0.539	0.995	0.048	0.065
SVA	0.208	0.998	0.046	0.084
	$\eta = 1$			
Std. ANOVA	0.026	1	0.000	0.027
SVA-PLS	0.251	0.997	0.052	0.095
SVA	0.080	0.999	0.058	0.074

Table 2.6.2 Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of independently expressed genes with different effects of the hidden variable over the two varieties.

Method	Sensitivity	Specificity	FDR	FNR
	$\eta = 0.1$			
Std. ANOVA	0.174	1.000	0.000	0.025
SVA-PLS	0.870	0.999	0.008	0.020
SVA	0.152	0.999	0.047	0.096
	$\eta = 0.5$			
Std. ANOVA	0.047	1.000	0.000	0.028
SVA-PLS	0.269	0.999	0.023	0.095
SVA	0.027	0.999	0.067	0.051
	$\eta = 1$			
Std. ANOVA	0.038	1.000	0.000	0.039
SVA-PLS	0.105	0.999	0.022	0.087
SVA	0.011	1.000	0.027	0.041

Table 2.6.3 Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of independently expressed genes, when the hidden variable has a complex differential pattern between the two varieties, resulting in a serious confounding.

Method	Sensitivity	Specificity	FDR	FNR
	$\eta = 0.5$			
Std. ANOVA	0.234	1.000	0.000	0.070
SVA-PLS	0.989	0.991	0.054	0.002

Table 2.6.4 Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of co-regulated genes with similar effects of the hidden variable over the two varieties.

Method	Sensitivity	Specificity	FDR	FNR
	$\eta = 0.5$			
Std. ANOVA	0.100	1.000	0.000	0.030
SVA-PLS	0.747	0.993	0.052	0.039

Table 2.6.5 Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of co-regulated genes with different effects of the hidden variable over the two varieties.

Method	Sensitivity	Specificity	FDR	FNR
	$\eta = 0.5$			
Std. ANOVA	0.105	1.000	0.000	0.027
SVA-PLS	0.618	0.998	0.019	0.055

Table 2.6.6 Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of co-regulated genes, when the hidden variable has a complex differential pattern between the two varieties, resulting in a serious confounding.

2.7 **Figures**

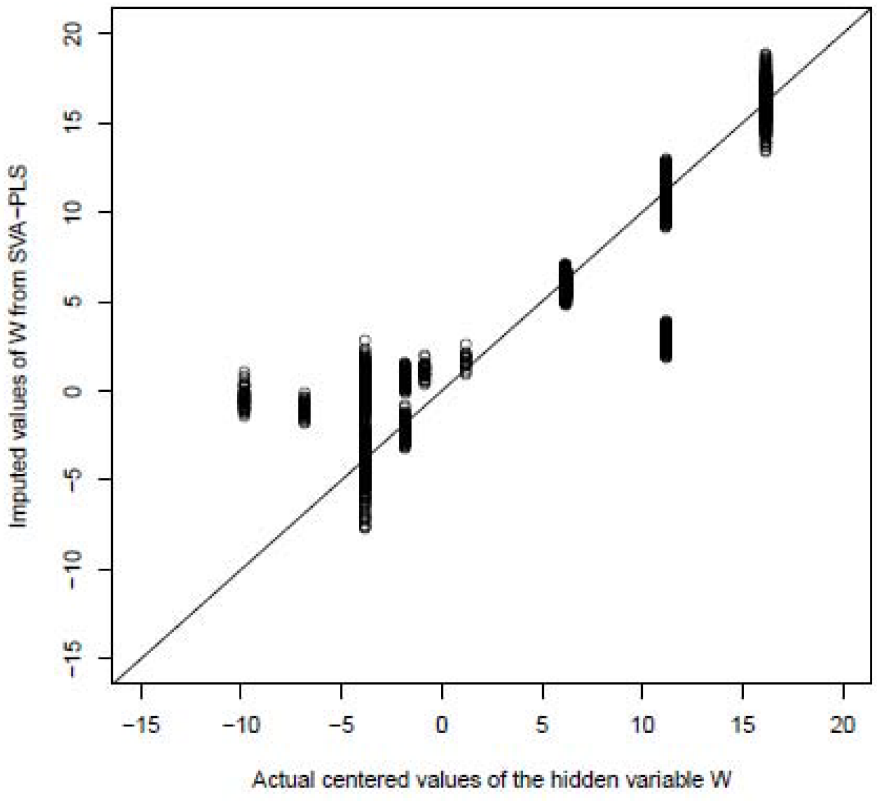

Figure 2.7.1 Plot of the PLS-imputed values of W versus the actual centered values under the setting of independently expressed genes, when the hidden variable has a complex differential pattern between the two varieties, resulting in a serious confounding.

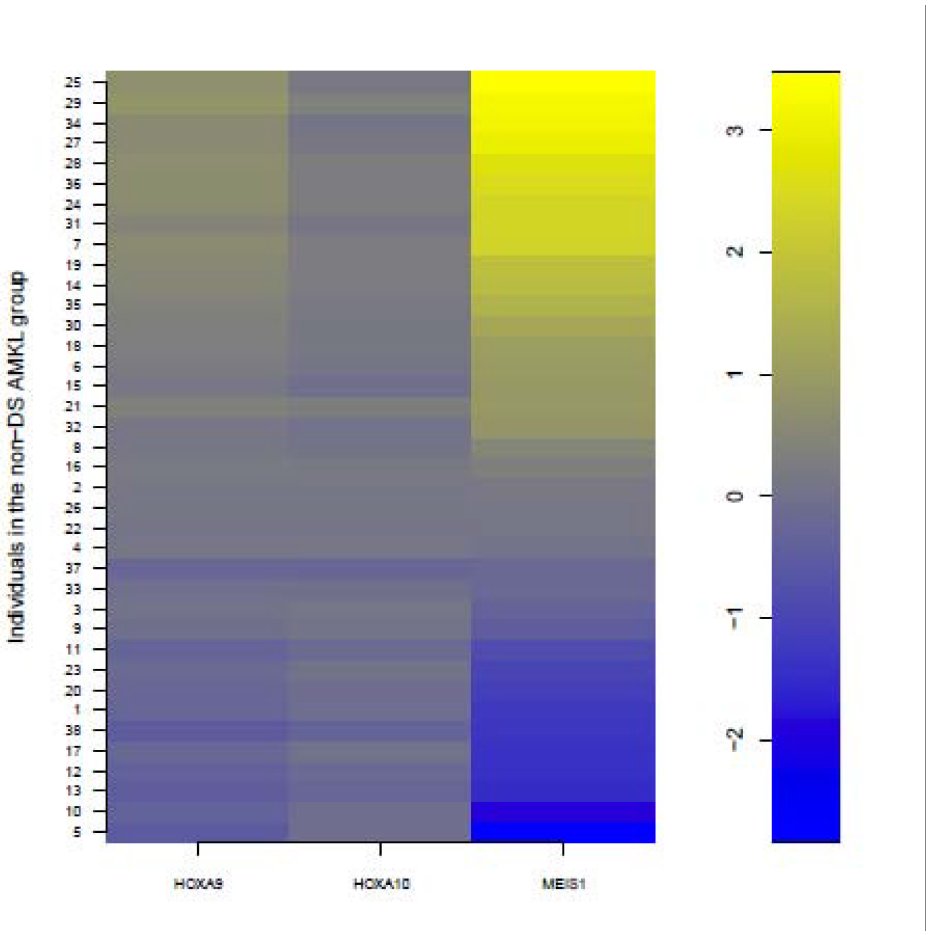

Figure 2.7.2 Heatmap of the PLS imputed W^{imp} for the three HOX/TALE family genes in the individuals under the non-DS AMKL variety showing a subgroup structure.

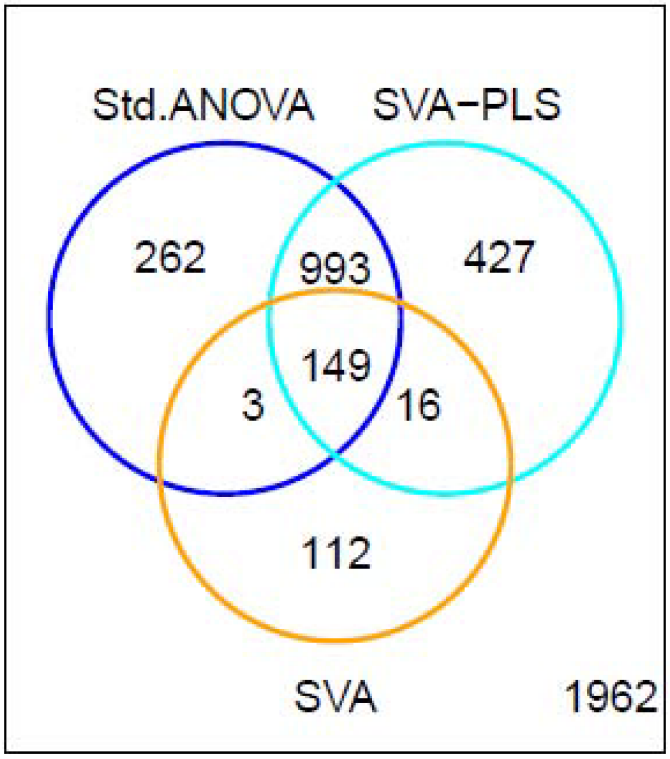

Figure 2.7.3 Venn Diagram showing the number of significant genes detected from the AMKL data by Standard ANOVA, SVA-PLS and SVA.

CHAPTER 3: *svapls* - AN R PACKAGE TO CORRECT FOR HIDDEN FACTORS OF VARIABILITY IN GENE EXPRESSION STUDIES

3.1 **Motivation**

We present an R package svapls that can be used to identify several types of unknown sample-specific sources of heterogeneity in a gene expression data and adjust for them in order to provide a more accurate inference on the original expression pattern of the genes over different varieties of samples. The proposed method implements Partial Least Squares regression to extract the hidden signals of sample-specific heterogeneity in the data and uses them to find the genes that are actually correlated with the phenotype of interest.

As discussed in Chapter 1 we know that several types of subject/sample specific factors constitute an important but often overlooked source of hidden variability in differential gene expression analyses. In a wide variety of situations these factors are triggered from certain specific biological, environmental or demographic profiles of the subjects corresponding to the collected tissue samples. The latent effects from these hidden factors can generate spurious signals of heterogeneity that may significantly distort the original differential expression pattern of the genes. In this context, a simple example is provided by the widely known batch-effect in microarray analyses, where subject tissue samples collected in separate batches can produce an additional effect of residual variation. The caveat of this effect is still manageable as composition of the batches are known

prior to analyses. But, numerous other factors may still exist that are not detectable from outside, but can potentially affect the subject-specific expression levels of the genes in different ways. They can in turn lead to complex latent expression structures in the entire genomic landscape of the data (e.g., confounded signals between the two groups of samples, correlated expression signals corresponding to a specific group of genes and samples affected by the hidden factors, etc.). The contributed impact of these factors, either acting singly or in consort can induce serious problems in multiple testing of differential expression for the genes. Thus, a number of truly significant genes can pass out undetected while many others may be wrongly flagged as positives. The consequence is a severe reduction in power (sensitivity) of the testing procedure accompanied by a substantially high rate of erroneous discoveries. Most of available softwares for differential gene expression analyses either overlook this broadly general issue of hidden variability or consider simple parametric regression approaches (linear regression, mixed effects models, etc.) to address the maladies of residual heterogeneity. However the complexity of problem necessitates the development of a more generalized and efficient technique that can identify these latent effects of variation in the data and adjust for them in order to deliver a more powerful and accurate inference on the actual expression pattern of the genes. This motivated us to construct a methodology (discussed in Chapter 1) that provides an unified framework for handling these widely different types of spurious variability in the data.

We have built an R software *svapls* that uses the multivariate Non-Linear Iterative Partial Least Squares (NIPALS) algorithm (Rosipal and Krämer, 2006) to extract the latent, unwanted effects of variation in a gene expression data and uses them to build an optimal ANCOVA model for detecting the truly differentially expressed genes. In the next section we describe the important functions in our

package along with illustrative examples that explain their practical usage in detail. The following section 3 demonstrates its comparatively superior performance with respect to three other popular softwares: *sam* (Tusher, V. et. al., 2001), *limma* (Smith, G.K., 2005) and *sva* (Leek and Storey, 2007) through a sensitivity analysis of two simulated differential gene expression datasets affected by complicated hidden variation patterns. Section 4 elucidates an application on a real-life dataset that proves the worth of our software through the detection of some phenotype-related genes that are deemed to be significant from their annotations in the literature.

3.2 **Brief Overview of the Package**

This R package consists of the three primary functions: *fitModel*, *svpls* and *hfp.* Below we give a brief outline of them. The function applications are demonstrated on a simulated dataset affected by hidden variation (*hidden_fac.dat*) that is inbuilt as a part of the R package.

The first function ***fitModel*** fits an ANCOVA model to the original log-transformed gene expression data ,with a certain number of PLS scores as surrogate variables (specified by **n.surr**) or the simple ANOVA model (Kerr et. al., 2000; Kerr et. al., 2002) if no surrogate variables are specified. This function provides an user with the flexibility of estimating the actual gene-variety interaction effects from a certain ANCOVA model with a specific choice on the number of surrogate variables, which can be selected depending on the complexity of the situation under study.

```
> data(hidden_fac.dat)
>
> ## Fitting an ANCOVA model with 5 surrogate variables
> fit <- fitModel(10,10,hidden_fac.dat,n.surr = 5)
> print(fit)
Estimated coefficients of the surrogate variables:
[1]  0.0425701446  0.0134271227  0.0012466815  0.0041702000 -0.0007253327
```

```
Estimated Mean Squared Error of the fitted model:
[1] 9.053331

AIC value of the fitted model:
[1] 51791.02
```

The second function ***svpls*** calls the first function ***fitModel*** to fit a number of ANCOVA models (specified by **pmax**) to the data and selects the optimal model as the one with the minimum value of the Akaike's Information Criterion (AIC) (Hirotsugu, 1974). This model is then used to predict the actual pattern of differential expression of the genes over the two sample varieties by performing a multiple hypothesis testing at specified value of the false discovery rate (FDR) (Benjamini and Hochberg, 1995) (specified by **fdr**).

```
> ## Fitting the optimal ANCOVA model to the data gives:
> fit <- svpls(10,10,hidden_fac.dat,pmax = 5)
>
> ## The optimal ANCOVA model, its AIC value and the positive genes detected from
it are given by:

> fit$opt.model
[1] 5
>
> fit$AIC.opt
[1] 51789.12
>
> fit$genes
 [1]  31  38  42  43  65  33  57  54  30  34  25  29  41  61  68  51  62  50  55
[20]  46  52  53  63  60  28  69  24  59  40  66  21  44  27  26  37  45  48  23
[39]  39  67  36  56  49  14  47  64  35   1  70   6   4 455  58  12   8  13  32
[58]   7  10   3  18  22  11 184
>
> ## The corrected gene expression matrix obtained after removing the effects of
the hidden variability is given by:
> Y.corrected <- fit$Y.corr
> pval.adj <- fit$pvalues.adj
```

While the Benjamini-Hochberg correction is used by default in our R package the p-values returned by the *svpls* object provides an user with the flexibility of

applying several other FDR controlling techniques and also performing the more specifically targeted gene set enrichment analyses.

A side-by-side plot of the histograms of the p-values obtained by a differential testing of the genes with the estimated effects from standard ANOVA and the optimal ANCOVA model selected by our R package clearly demonstrates its efficacy in terms of the proximity of the null p-values towards the uniform distribution (Figure 3.7.1).

The third function ***hfp*** produces a heatmap for the PLS-imputed estimate of the residual expression heterogeneity corresponding to an user-specified set of genes and samples (specified by **gen** and **ind** respectively). This enables us to understand how intensely the latent factors from a certain set of subjects affect the true expression levels of a specified set of genes.

```
## Specifying the set of genes and subjects
> genes <- c(1,20,55,70,100,150,250,450)
> subjects <- c(1,4,7,10,11,15,17,20)
>
> hfp(fit,genes,subjects,hidden_fac.dat)
```

This produces a plot revealing the way the hidden variable affects the expression pattern of the selected group of genes over the specified subjects (Figure 3.7.2). Clearly, we can observe a substantial difference in the expression variability caused by the latent factor for subjects $1, 4$ 7 and the rest specified under the selected group. The effect of the hidden variability from the subjects $1, 4$ and 7 is consistent over the first and last four genes in the gene set, while the impact of the other subjects varies alternately between the two gene groups.

3.3 **Comparative evaluation with other available softwares**

In this section we illustrate the application of the R package along with the other three popular softwares through a family of simulation analyses conducted under a set of noise-to-signal ratios (η) controlling the relative intensity of the random error and primary signal variances (Chakraborty et. al. 2012). In each simulation study we generate expression measurements on 1000 genes over n subjects classified equally into two groups 1 and 2. We consider two different choices of n as 20 and 40. The genes are considered to be correlated and affected by a highly complex subject-specific confounder (Chakraborty et. al., 2012). Overall, the first 70 genes are considered to be truly differentially expressed over the two varieties while the rest are chosen as non-significant. The simulation study is based on the computation of the average values of two right decision indicators (sensitivity, specificity) and two wrong detection indicators (false discovery rate and false non-discovery rate) for the two different sample sizes, evaluated from 500 Monte-Carlo replications (Tables 3.6.1 and 3.6.2). The obtained results clearly reveal the superior sensitivity of *svapls* compared to the other three packages *sam, limma* and *sva* along with an expected improvement on a larger sample size (Table 3.6.2). Especially, *limma* and *sva* perform very poorly in terms of the detecting power. In addition the average error rates of falsely detecting some non-significant genes (FDR) and not identifying some truly positive genes (FNR) are much lower for *svapls* compared to the other three softwares. The sensitivity of *sam* is comparable to our method but is adversely impacted by the significantly elevated false discovery rate. The specificity rate is the best for *svapls* closely followed by *sva*, while *sam* and *limma* are less efficient in this context. Thus, overall the function *svpls* in our R package is capable of detecting the truly differentially expressed genes with more power along with an efficient control over the wrong decisions with comparatively smaller error rates.

3.4 **Application on the Golub Data**

Now, we explore the performance of *svapls* on the pre-processed ALL/AML dataset (Golub et. al., 1999, Dudoit et. al., 2002). It contains the log-transformed expression levels of 7129 genes over two groups of patients: 47 having Acute Lymphoblastic Leukemia (ALL) and 25 suffering from Acute Myeloid Leukemia (AML). The patient tissue samples were obtained from the following four sources: (1) Dana-Farber Cancer Institute (DFCI), (2) St-Jude's Children's Research Hospital (St-Jude), (3) Cancer and Leukemia Group B (CALGB) and (4) Children's Cancer Study Group (CCG). This inherent classification in the data can potentially generate significant batch effects that may distort the original expression pattern of the genes. This motivated the implementation of our R package on this dataset. The corrected gene expression matrix returned from the use of the *svpls* function on this data demonstrates that the batch effects due to variability in the sample sources have been removed effectively (Figure 3.7.3). The haphazard distribution of the samples from the four batches in the corrected gene expression matrix returned by the function *svpls* in our package wipes out the additional effects owing to the observed batch-specific clustering in the original data. In this context *svapls* fares equally well compared to another popular R package *ber* for removing batch effects in microarray data (Giordan, M., 2012).

Overall, *limma* detects 7128 genes followed by 3307 genes from *sam*, 1015 genes from our *svapls* and 412 genes from *sva.* A Venn diagram (Figure 3.7.4) represents the extent of overlap between the genes detected by the four softwares. Specifically, *limma* detects all the genes that are found to be significant from the other three softwares. This may be attributable to its high false discovery rate (FDR)

as was observed in the simulation study. Interestingly, *svapls* detected 24 genes that are missed by both *sam* as well as *sva.* Among them the genes *CD74, TNFRSF1A, LCN2* and *GSN* deserve special mention. All these genes are either related to some type of cancer or regulate cell growth/apoptosis. *CD74* plays an important role in multiple myeloma and its higher expression induces tumor cell malignancy (Burton, J.D. et. al., 2004). An isoform of the tumor necrosis factor *TNFRSF1A* is associated with the development of Acute Lymphoblastic Leukemia (ALL) in children (Wu et. al. 2003). Specifically, *LCN2* has been found to be connected with Acute Myelogenous Leukemia (AML) (Shimada et. al., 2002). *GSN* plays a significant role of suppressing tumorigenicity in lung cancer (Sagawa et. al., 2003) and has a diminuted expression in bladder cancer cells (Haga, 2003).

3.5 **Discussion**

Various hidden sources of variation are found to exist in a gene expression data that cannot be removed by the standard normalization procedures. But, their effect may be substantial enough to change the expression pattern of the genes over two different varieties of samples. The immediate consequence is a large reduction in the detection power of the testing procedure employed to find the truly significant genes, followed by highly elevated error rates. In this project we discuss the development and usage of an R package *svapls* that can tackle a wide variety of hidden effects in a gene expression analysis and can deliver a more accurate inference on the differential expression variability of the genes between two groups of samples (tissues). We illustrate the superior performance of our R package in comparison to other popular softwares available for differential gene expression analyses. The high detection power (sensitivity) of our package *svapls* along with the reasonably small error rates provides it a significantly better edge over the competing softwares. Specifically, *sva* is outperformed by our package in terms of

the sensitivity (power), while *sam* comes close although its competence is severely marred by the considerably high false discovery rate (FDR). In addition the graphical representation of the hidden variation (by the function *hfp*) from our package enables the user to understand the pattern in which the hidden sources of variability affect the expression signals of any specified subset of genes over a selected group of subjects. This paves the way to more sophisticated analyses of subject-set specific gene expression variability in the data. Application of our package on the Golub data demonstrates its efficacy in removing the significant batch effects from the collected/analyzed samples. Moreover our package detects four additional genes (missed by both *sva* and *sam*) that have been found to be connected to Leukemia or some other type of cancer.

Our R package provides the user with a simplified framework for analyzing gene expression data with a wide range of hidden variation patterns and delivering a differential gene expression analysis with substantially improved power and accuracy.

3.6. **Tables**

Method	Sensitivity	Specificity	FDR	FNR
	$\eta = 0.05$			
LIMMA	0.2439	0.6136	0.4436	0.2131
SAM	0.9215	0.5908	0.7451	0.0132
SVA	0.3522	0.9985	0.0473	0.0454
SVAPLS	0.9461	0.9997	0.0034	0.0039
	$\eta = 0.10$			
LIMMA	0.2382	0.6523	0.3949	0.2171
SAM	0.9053	0.6376	0.6588	0.0142
SVA	0.3139	0.9988	0.0488	0.0480
SVAPLS	0.9257	0.9994	0.0086	0.0054
	$\eta = 0.15$			
LIMMA	0.2088	0.6542	0.3786	0.2288
SAM	0.8451	0.6813	0.5955	0.0193
SVA	0.2481	0.9991	0.0477	0.0518
SVAPLS	0.8636	0.9991	0.0129	0.0099

Table 3.6.1 : Average performance measures from a sensitivity analysis on the simulated gene expression data on 20 subjects (10 being in each group), with the four softwares *limma, sam, sva* and *svapls.*

Method	Sensitivity	Specificity	FDR	FDR
	$\eta = 0.05$			
LIMMA	0.7951	0.2086	0.7939	0.4963
SAM	0.9924	0.5784	0.7569	0.0011
SVA	0.5892	0.9976	0.0499	0.0294
SVAPLS	0.9989	0.9998	0.0024	0.0008
	$\eta = 0.10$			
LIMMA	0.7825	0.2202	0.7582	0.5918
SAM	0.9791	0.6133	0.7051	0.0024
SVA	0.5854	0.9939	0.0449	0.0295
SVAPLS	0.9868	0.9953	0.0087	0.0007
	$\eta = 0.15$			
LIMMA	0.7355	0.2335	0.7499	0.5615
SAM	0.9627	0.6525	0.6504	0.0047
SVA	0.5289	0.9941	0.0452	0.0336
SVAPLS	0.9734	0.9949	0.0150	0.0017

Table 3.6.2 : Average performance measures from a sensitivity analysis on the simulated gene expression data on 40 subjects (20 being in each group), with the four softwares *limma, sam, sva* and *svapls.*

3.7. **Figures**

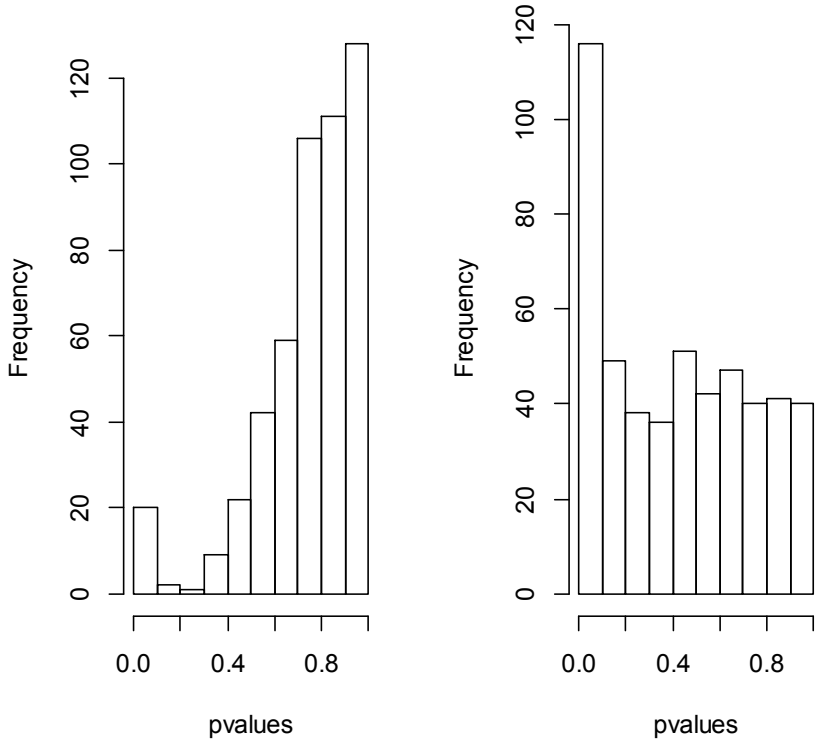

Figure 3.7.1 Histograms of the unadjusted (left) and adjusted (right) p-values obtained respectively, from the application of standard ANOVA and our R package *svapls* on the inbuilt dataset *hidden_fac.dat*.

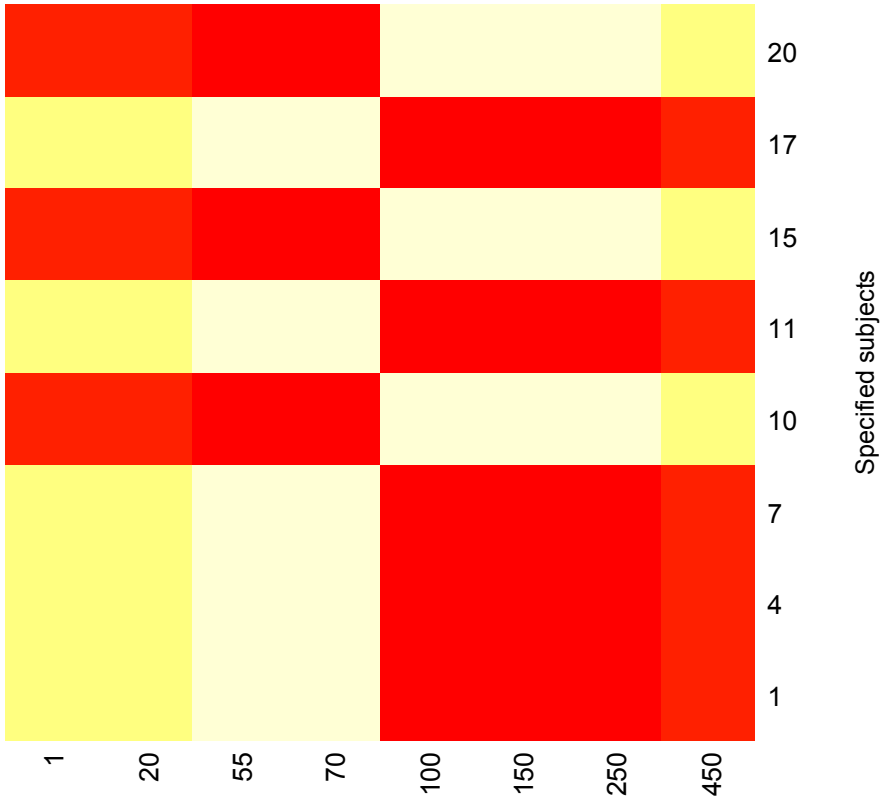

Figure 3.7.2: Heatmap showing the hidden variability in gene expression for a specified group of subjects and genes.

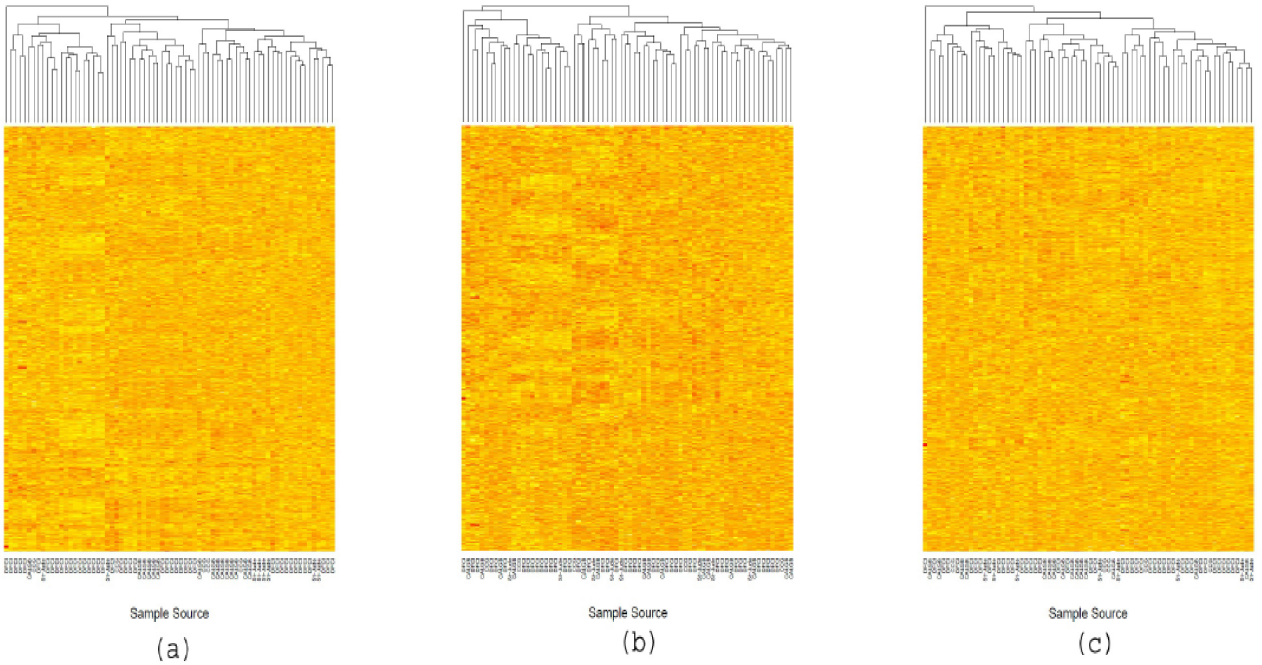

Figure 3.7.3: (a), (b), (c) - Heatmaps showing the original and corrected expression levels for the first 1000 genes in the Golub data.

(a) Heatmap for the first 1000 genes in the original Golub expression data.

(b) Heatmap for the first 1000 genes in the adjusted Golub expression data obtained by use of the R package **ber**.

(c) Heatmap for the first 1000 genes in the adjusted Golub expression data obtained by the use of our R package **svapls**.

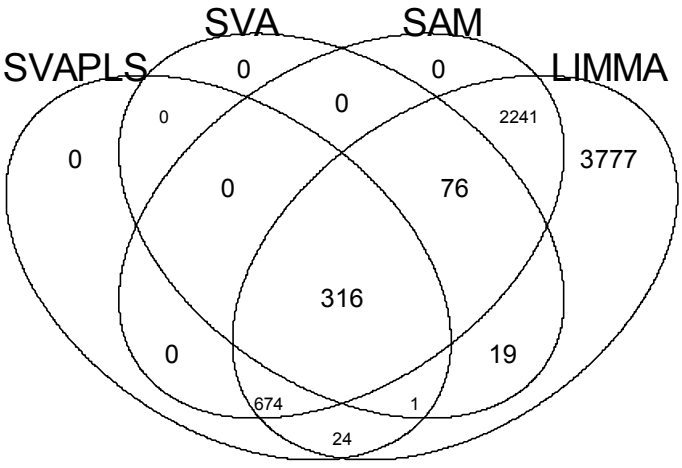

Figure 3.7.4: A Venn-diagram showing the overlap pattern of the genes detected by *svapls*, *sva*, *sam* and *limma*.

CHAPTER-4: NONPARAMETRIC REGRESSION OF TEMPORAL FUNCTIONS IN A MULTISTATE MODEL UNDER RIGHT CENSORING VIA ADDITIVE MODELS FOR COUNTING AND NUMBER AT RISK PROCESSES

4.1 The Proposed Methodology

4.1.1 Data Structure and Notations

We start by giving a brief outline of a multistate model data structure along with the relevant notations that have been used throughout the rest of this project. Suppose we have n individuals moving through an interconnected network of J stages $0, 1,...(J-1)$ in a certain multistate model.

Let $S_i(t)$ denote the stage occupied in the multistate model by the individual i at time t. Define T_i^* to be the actual final transition time for the individual i. C_i is the corresponding right censoring time and $T_i = min(T_i^*,\ C_i)$ is the observed final transition time. In case the individual i moves from stage j to j' we define $U_{jj'}^i$ to be the corresponding time of transition (taken to be ∞ if that transition is not made at all). $I(C_i > T_i^*)$ is the censoring indicator for individual i and $K_i(t)$ is the conditional survival function for the censoring distribution assuming it to be solely dependent on the baseline covariates. But, under a more general framework it may not be a survival function. A more precise definition of $K_i(t)$ along with the computation of its estimate $\widehat{K}_i(t)$ is discussed in Section 2.4. Let X_{iu} be the observed value of the covariate X_u for individual i. $u = 1, 2, ...p$. Overall we have $X_i = (X_{i1}, X_{i2}, \cdots X_{ip})$.

4.1.2. Additive Models

In this work we have developed an IPCW modified version of the standard backfitting technique (Hastie and Tibshirani, 1990) to estimate the conditional transition and at-risk processes for the different states in a progressive multistate model, given the observed values for a set of covariates. Our objective is to use these conditionally estimated processes to derive the state occupation probabilties for the different stages in the model at specific values of these covariates.

Let $N_{jj'}$ denote the counting process for transitions from stage j to j', with jumps equal to $\Delta N_{jj'}(t) = \sum_{i=1}^{n} I(S_i(t-) = j, S_i(t) = j')$ and $Y_j(t) = \sum_{i=1}^{n} I(S_i(t-) = j)$ be the at-risk process for individuals occupying stage j just prior to time t. But these two processes are not completely observable due to the presence of censoring in the data. Recently, Mostajabi et. al., (2012) estimated these two processes at a specified value of a single covariate using a IPCW (Datta and Satten, 2001) based locally weighted kernel smoother. In our method we extend this approach to the case of multiple covariates by using an IPCW based backfitting regression technique (Hastie and Tibshirani, 1990).

Imitating the structure of the simple backfitting framework (Hastie and Tibshirani, 1990) we consider the following two non-parametric regression models corresponding to the two processes $N_{i,jj'}(t)$ and $Y_{i,j}(t)$:

$$E(N_{i,jj'}(t)|\underset{\sim}{x}) = \alpha(t) + f_1^t(x_1) + f_2^t(x_2) + \cdots + f_p^t(x_p) \quad (1)$$

$$E(Y_{i,j}(t)|\underset{\sim}{x}) = \gamma(t) + g_1^t(x_1) + g_2^t(x_2) + \cdots + g_p^t(x_p), \quad (2)$$

where $\underset{\sim}{x} = (x_1, x_2 \cdots x_p)$ is a specified vector of values for $\underset{\sim}{X}$, $\alpha(t)$ and $\gamma(t)$ are the intercepts and $f_1^t, f_2^t, \cdots f_p^t$ and $g_1^t, g_2^t, \cdots g_p^t$ are respectively, two different sets of p unknown arbitrary functions, corresponding to the covariates $X_1, X_2, \cdots X_p$, at a certain time point t. Let us denote $\mu_1(t|\underset{\sim}{x}) = E(N_{i,jj'}(t)|\underset{\sim}{x})$ and $\mu_2(t|\underset{\sim}{x}) = E(Y_{i,j}(t)|\underset{\sim}{x})$. Now, Using the ordinary least squares (OLS) principle the estimates $\widehat{\mu}_1$ and $\widehat{\mu}_2$ can be obtained by minimizing the following two criteria:

$$\phi_1 = \sum_{i=1}^{n} (N_{i,jj'}(t) - \alpha(t) + f_1^t(x_{1i}) + f_2^t(x_{2i}) + \cdots + f_p^t(x_{pi}))^2$$

$$\phi_2 = \sum_{i=1}^{n} (Y_{i,j}(t) - \gamma(t) + g_1^t(x_{1i}) + g_2^t(x_{2i}) + \cdots + g_p^t(x_{pi}))^2$$

Minimization of ϕ_1 and ϕ_2 lead to a system of p equations that can be solved by the highly time consuming methods of matrix decomposition (e.g., QR decomposition, , where a matrix is expressed a product of an orthogonal matrix Q and an upper triangular matrix R). Here we present an easier and faster alternative way that is based on a IPCW-reweighted version of the simple backfitting algorithm (Hastie and Tibshirani, 1990)

Suppose we have n observations $H_1(t), H_2(t) \cdots H_n(t)$ on an arbitrary right censored stochastic process $H(t)$ from n individuals with corresponding censoring indicators $\delta_1, \delta_2, \cdots \delta_n$. We denote corresponding probabilities of censoring by $k_1, k_2, \cdots k_n$. Further details on these weights for specific choices of the stochastic process $H(t)$ will be discussed later on. Then for the triplet

$\Omega = \{(H_i(t), \delta_i, k_i), i = 1, 2 \cdots n\}$, the IPCW-backfitting algorithm operates in the following manner:

Step 1: For each observed transition time t, we start with an initial choice of p mean zero functions $f_1^t, f_2^t, \cdots f_p^t$, say $f_1^{t,0} = 0, f_2^{t,0} = 0, \cdots f_p^{t,0} = 0$ and set $\widehat{\alpha}(t) = \frac{1}{n}\sum_{i=1}^{n} H_i(t)\delta_i/k_i$.

Step 2: For each $u = 1, 2, \cdots p$, compute an updated estimate of f_u^t as $f_u^{t,1}(v) = \sum_{i=1}^{n} S_u^i(v)[H_i(t) - \sum_{k \neq u} f_k^{t,0}(x_{ik})]\delta_i/k_i - \widehat{\alpha}$, where $S_u^i(v)$ is the value of a smoothing function for individual i corresponding to a certain value v of the covariate X_u $(u = 1, 2 \cdots p)$.

Step 3: For $u = 1, 2, \cdots p$, set $f_u^{t,1} = (f_u^{t,1}(x_{1u}), f_u^{t,1}(x_{2u}), \cdots f_u^{t,1}(x_{nu}))$.

Repeat **Steps 2-3**, until for some integer l, $\sum_{u=1}^{p} \frac{||f_u^{t,l} - f_u^{t,l-1}||}{||f_u^{t,l-1}||} < \epsilon$, or l crosses a certain pre-specified maximum number of iterations $max.it$. For our purpose we have taken $\epsilon = 10^{-5}$ and $max.it = 100$.

Thus, a conditional estimate for the mean value of the stochastic process $H_i(t)$ for the i-th individual at a certain time t given the specified value $\underset{\sim}{x} = (x_1, x_2 \cdots x_p)$ of the covariate vector $\underset{\sim}{X} = (X_1, X_2 \cdots X_p)$ is provided by $\widehat{E}(H_i(t)|\underset{\sim}{x}) = \widehat{\alpha}(t) + \sum_{u=1}^{p} f_u^{t,l}(x_u)$.

Below we present a theorem on the convergence of our algorithm for the case of two covariates. To that end we let $M_j = S_j\Lambda$, where

$$S_j = \begin{pmatrix} S_j^1(x_{1j}) \\ S_j^1(x_{2j}) \\ \vdots \\ S_j^n(x_{nj}) \end{pmatrix}_{n \times n}$$

is the smoother matrix scanning all the n observations on the covariate $X_j, j = 1, 2$ and $\Lambda = diag(\delta_1/k_1, \delta_2/k_2 \cdots \delta_n/k_n)$ is a diagonal matrix of the censoring indicators $\delta_1, \delta_2 \cdots \delta_n$ for the process H weighted by inverse of their corresponding probabilities of censoring $k_1, k_2 \cdots k_n$.

Theorem 1. For two covariates and any sample size n the IPCW reweighted backfitting algorithm converges to an unique solution if $||M_1 M_2|| < 1$ and $||M_2 M_1|| < 1$, where $||\,||$ denotes an operator norm.

A formal proof of the theorem in case of two covariates $(p = 2)$ is provided in the appendix. The situation of more than two covariates is highly complicated and demonstration of convergence requires the strict assumption of symmetric smoothing matrices for all the covariates with corresponding eigenvalues being in the interval $[0, 1]$ (Hastie and Tibshirani, 1992). For the sake of simplicity we so not discuss that case under the present context.

We define $d_i = I(C_i \geq U_{jj'}^i), b_i = I(C_i \geq t)$ as the censoring indicators for the process of transition from stage j to $j^{'}$ and the process for individuals at risk of transition from stage j.

Now, following the algorithm with $\Omega = \{(N_{i,jj'}(t), d_i, \widehat{K}_i(U^i_{jj'} -)), i = 1, 2\cdots n\}$ we get $\widehat{\mu}_1(t|\underset{\sim}{x})$ which provides an estimate of the conditional mean of the stochastic process for transition from stage j to j' at a certain time t as $\widehat{E}(N_{jj'}(t)|\underset{\sim}{x}) = n\widehat{\mu}_1(t|\underset{\sim}{x})$. Now, the stage-to-stage conditional transition counts are supposed to be non-decreasing over time and hence estimated means of the conditional transition processes are monotonized by using isotonic regression with the generalized pooled adjacent violators algorithm (Barlow et. al., 1972). The corresponding conditional mean at-risk process at time t can be estimated in a similar way by $\widehat{E}(Y_j(t)|\underset{\sim}{x}) = n\widehat{\mu}_2(t|\underset{\sim}{x})$, using $\Omega = \{(Y_{i,j}(t), b_i, \widehat{K}_i(t -)), i = 1, 2\cdots n\}$. For our purpose we specifically use the smoothing function $S^i_u(v) = \phi((v - x_{iu})/h_u)/\sum_{r=1}^{n}\phi((v - x_{ru})/h_u)$ for covariate X_u, where ϕ is the standard Gaussian kernel and the h_u is the corresponding bandwidth that is selected from the observed values of the covariate by a method due to Wand and Jones, (1995).

4.1.3. Conditional Transition Hazard Rates and State Occupation Probabilities

In a general uncensored multistate model with J states $0, 1\cdots(J-1)$ the conditional hazard of transition from stage j to j' $(j \neq j'; j = 1, 2\cdots J)$ given a specific value $\underset{\sim}{x}$ of the covariate vector $\underset{\sim}{X}$ is given by $\alpha_{jj'}(t|\underset{\sim}{x}) = \lim_{dt\to 0} Pr(S(s) = j',$ for some $s \in [t, t + dt)|S(t -) = j, \underset{\sim}{X} = \underset{\sim}{x})/dt$

Thus the cummulative conditional stage-to-stage transition hazard matrix for the multistate model can be represented as:

$$A_{jj'}(t|\underset{\sim}{x}) = \begin{cases} \int_0^t \alpha_{jj'}(u|\underset{\sim}{x})du\,, & \text{if } j \neq j' \\ -\sum\limits_{j \neq j'} A_{jj'}(t|\underset{\sim}{x}), & \text{if } j = j'. \end{cases}$$

The corresponding estimator obtained from our method is given by:

$$\widehat{A}_{jj'}(t|\underset{\sim}{x}) = \begin{cases} \int_0^t d\widehat{E}(N_{jj'}(u|\underset{\sim}{x}))/\widehat{E}(Y_j(u|\underset{\sim}{x}))\,, & \text{if } j \neq j' \\ -\sum\limits_{j \neq j'} \widehat{A}_{jj'}(t|\underset{\sim}{x}), & \text{if } j = j'. \end{cases}$$

Therefore for any two time points s and t, $(s < t)$, the estimator of the $J \times J$ conditional transition probability matrix P, in case the multistate system is conditionally Markov given $\underset{\sim}{X}$ can be represented in an Aalen-Johansen form as:

$$\widehat{P}(s,t|\underset{\sim}{x}) = \prod_{(s,t]} (I + d\widehat{A}(u|\underset{\sim}{x}))$$

Following Mostajabi and Datta, (2012) even without the assumption of conditional Markovity, the stage occupation probability for stage j in the model, given the covariate can be estimated as

$$\widehat{p}_j(t|\underset{\sim}{x}) = \sum_{k=0}^{J-1} \frac{\widehat{Y}_k(0+|\underset{\sim}{x})}{n} \widehat{P}_{kj}(0,t|\underset{\sim}{x}), \tag{3}$$

where $\widehat{P}_{jk}(0,t|\underset{\sim}{x})$ is the (j,k)-th element of the matrix $\widehat{P}(0,t|\underset{\sim}{x})$.

4.1.4. Censoring Hazards and Estimation of the Weights $K_i(t)$

We define a generalized covariate $Z_i(t)$ corresponding to each individual i at time t, in such a way that it can involve both baseline ($X_u,\ u = 1, 2, \cdots p$) as well as internal potentially time varying covariates (for example the state occupation indicators of the individuals), that may significantly affect the censoring hazard. Let $\overline{Z}_i(t) = \sigma(Z_i(s) : 0 \leq s < t)$ be the observed history of Z_i till time t. We assume that given the record of $\bar{Z}_i(t)$, for an individual i until a certain time t, the corresponding censoring hazard ($\lambda_C^i(t)$) is independent of the stage occupied at t. Mathematically, this boils down to letting for each individual i :

$$\lambda_C^i(t|S_i(t), \bar{Z}_i(t)) = \lambda_C^i(t|\bar{Z}_i(t))$$

where $\lambda_C^i(t|.) = \lim_{dt \to 0} Pr(C_i \in [t, t + dt)|T_i \geq t, .)/dt$.

Then $K_i(t) = exp(-\Lambda_C^i(t|\overline{Z}_i(t)))$, where $\Lambda_C^i(t|\overline{Z}_i(t)) = \int_0^t \lambda_C^i(u|\overline{Z}_i(t))du$

The IPCW weights for the individuals can be estimated in a variety of ways using different hazard models. The Aalen's linear hazard model (Aalen, 1980; 1989) provides a flexible and generalized way that can incorporate the effects of both external as well as internal covariates, on the risk of the censoring. Using this model the censoring hazard for the i-th individual in the model can be represented as:

$$\lambda_C^i(t|\bar{Z}_i(t)) = \sum_{k=0}^{m} \beta_k(t) U_{ik}(t)$$

where $U_{i0}(t) = 1$ and $U_{ik}(t) = f_k(\bar{Z}_i(t)), k = 1, 2, \cdots m$, are certain functions of the past history of the generalized covariate process $\bar{Z}_i(t)$, that can represent different types of complex effects on the censoring hazard. $U_{ik}(t)$ can be constructed in several ways depending on the situation under study. As in our case we have considered two different choices of $U_{ik}(t)$. For the simulation studies $U_{ik}(t)$ equals

the value of the kth covariate X_k for individual i while for the real-life data analyses it is an indicator variable showing whether the kth stage was occupied by individual i at time t. β_k's are the corresponding regression coefficients, measuring the impact of these covariates functions on the overall censoring hazard. Define $U_i(t) = (U_{i1}(t), U_{i2}(t), \cdots U_{ip}(t))$. Then the Aalen's estimator (Aalen, 1980) of the cummulative censoring hazard for the i-th individual is given by $\widehat{\Lambda}_C^i(t|\overline{Z}_i(t)) = \int_0^t \widehat{\lambda}_C^i(u|\overline{Z}_i(t))du \quad = \sum_{j=1}^{n} I(T_j \leq t)(1-\delta_j)U_i(T_j)R^{-1}(T_j)U_j(T_j)$, where $R(t) = \sum_{i=1}^{n} I(T_i \geq t)U_i(t)U_i^T(t)$. Finally, the IPC weight for the i-th individual can be expressed as $\widehat{K}_i(t) = exp(-\widehat{\Lambda}_C^i(t|\overline{Z}_i(t)))$.

4.2. **Simulations**

4.2.1. The Simulation Design

We illustrate the performance of our method by designing a simulation study with n individuals moving through the different branches of a 5-stage acyclic multistate model (see Figure 4.6.1). For the sake of simplicity we assume that all the individuals head out from the intitial stage 0. The flow of the individuals along the different branches of the model is controlled by generating a Bernoulli (0.4) random variable s, for each of them. The individuals corresponding to $s = 1$ follow the 0-1 arm, whereas the ones with $s = 0$ take the path to stage 2. Similarly, we generate values of s for the individuals reaching stage 1 and use them as before to decide their subsequent transition to stage 3 or 4. The individual state-to-state transition times are generated from a Weibull distribution with different shape and scale parameters for the different branches in the model. Additionally, for each individual we also consider information on two different covariates X_1 and X_2, where both of them are generated by taking the absolute value of a random variable following a $N(\mu, 1)$ distribution, with $\mu \sim N(10, 1.5)$. The individual transitions in different

branches of the model are assumed to be affected by right censoring and its rate is varied from moderate (25%) to heavy (50%), for each choice of the sample size n.

The entire simulation study is conducted under two different structural settings: semi-markov and markov. Under each scenario, the censoring is allowed to depend on the two covariates $X_1$and X_2, by generating its corresponding time of occurrence for an individual i, from an exponential distribution with mean $\frac{1}{d(X_1^i+X_2^i)}$, where d depends on the censoring rate. We consider two different values for the parameter d, in order to incorporate respectively, 25% (moderate) and 50% (heavy) censoring in the data.

4.2.2. Conditionally Semi-Markov Network

Under the conditionally semi-markovian structure, the future transition times of the individuals are allowed to be partially dependent on their previous transitions, given a particular set of values for the covariate vector. This is achieved by generating the waiting times for the individuals at each state in the model, and using them to simulate their times of any possible future transition from that state.

As discussed earlier, the Bernoulli random variable s is used to allocate the n individuals to the different paths of transition in the 5-stage model. Let B_{jk} denote the set of individuals making a transition from stage j to stage k. Then for each individual $i \in B_{jk}$, the waiting time W_{jk}^i at stage j of the model is generated from a Weibull distribution with shape r_{jk} and scale q_{jk}. Specifically, we take $r_{01} = r_{02} = 0.5, q_{01} = q_{02} = 1;\ r_{13} = r_{14} = 0.5, q_{13} = q_{14} = 5.$

The semi-markovity is incorporated in the model by generating the transition times for each individual i moving to stages 3 or 4 in the network, by setting $T_i = W_{01}^i + W_{13}^i$ or $T_i = W_{01}^i + W_{14}^i$. Now adding that to the means of their corresponding covariate values $(X_{i1} + X_{i2})/2$ generates an actual time of transition to stage 3 or 4 in the network as $V_i^* = T_i + (X_{i1} + X_{i2})/2, i = 1, 2, \cdots, n$. In this way for each individual i, the corresponding covariate values are enabled to affect his/her movement from one stage to another.

4.2.3. Conditionally Markov Network

In the conditionally markov setting, the future transitions of the individuals are allowed to depend directly on their previous transitions, by using a functional connection between their corresponding times of occurrence. Thus, if V_1 is the time of transition for an individual from stage 0 to 1, then the second transition time (V_2), to stage 3 or 4 is generated as $V_2 = D^{-1}[D(V_1) + R_2(1 - D(V_1))]$ where $D()$ is the cummulative distribution function of a Weibull $(0.5, 1)$ random variable and R_2 is randomly generated from an $U(0, 1)$ distribution. This association between the two transition times characterizes the markovity in the model.

Here, for each individual i, the effects of the two covariates are incorporated in a similar way as before, by generating the first transition time as $V_{i1} = G_{i1} + (X_{i1} + X_{i2})/2$, where $G_{i1} \sim$ Weibull$(0.5, 1)$.

4.2.4. Study of the Censoring Bias

We study the censoring bias by comparing the performances of our method between the censored and uncensored versions of the data, using an average $L1$-distance between two estimates of the occupation probabilities for each stage j in the model: one computed from the original uncensored data $(\widehat{p}_j^U(t|\underset{\sim}{x}))$ and the other

obtained from the observed right censored data $(\widehat{p}_j^C(t|\underset{\sim}{x}))$, where $\underset{\sim}{x} = (x_1, x_2)$ is a specified vector of values for the two covariates $X_1$and X_2. As for our purpose, we have taken x to be the vector of medians for the generated covariate values. We define:

$$\Delta_1 = E\int |\widehat{p}_j^U(t|\underset{\sim}{x}) - \widehat{p}_j^C(t|\underset{\sim}{x})| d\widehat{F}_n(t)$$

as the average L1-distance for comparing the two sets of estimated stage occupation probabilities, where $\widehat{F}_n(t)$ is the empirical cummulative distribution function (CDF) for the observed transition times generated under the model. Intuitively, a gradual decrease in this distance with the increase in sample size is expected to demonstrate the consistency of our method.

We perform the entire simulation study over 1000 Monte-Carlo replications and report the average L1-distances Δ_1 in Tables 4.5.1 and 4.5.2. For the semi-markov model with 25% right censoring we observe that the average linear difference between the estimated stage occupation probabilities (from the censored and uncensored data versions) decrease gradually as the number of individuals in the model increase from 100 through 500 upto 1000. Specifically, an average relative decrement of 43% is found in the difference between the two sets of estimates as the sample size increases from 100 to 500, while the rate is 16% for an even larger set of 1000 individuals. Under 50% censoring the average L1-distances follow a similar pattern as before, but have higher values as is expected from the impact of the substantially large proportion of censored observations in the data. In addition, the estimated standard errors for these distances are found to be less than 0.02 under both the moderate (25%) and heavy (50%) censoring scenarios, for all the five stages in the model. Under the markov setting we observe a similar pattern of gradual decrement in the average distances along with the increase in sample size.

In this setting under moderate censoring (25%) we achieve an average relative decrease of 41% in the L1-distance as the number of individuals increase from 100 to 500, which shifts to 19% as the sample sizes grows to 1000. We observe a similar pattern for the case of heavy censoring (50%) as well. Here, the estimated standard errors are all observed to be less than 0.03 thereby justifying the robust performance of our method as in the previous scenario. Interestingly, under the semi-markov setting our method produces estimates that are closer to their corresponding versions based on the uncensored data, compared to the estimators that are obtained under a markov model. Thus the detailed overall conclusions clearly reveal that under varying censoring rates (moderate as well as heavy), for both markov as well as semi-markov models our method robustly handles the underlying right censoring in the data and demonstrates a consistent performance in terms of producing estimators that are found to converge asymptotically towards their corresponding full-data (uncensored) versions.

4.2.5. Study of the Overall Estimation Bias

In this section we perform an analysis to study the deviation (bias) of our estimated stage occupation probabilities from their population counterparts that generate the original multistate model data under both of the two structural simulation settings as discussed in the previous section. To that end, under each setting, we generate a set of state-to-state transition times for a very large number of individuals $(N$=10000$)$ and we order the distinct transition times as $t_1 < \cdots < t_M$, $(M > 10000)$, say. We use the Nadaraya-Watson estimator (Nadaraya, 1964) based on a bivariate gaussian kernel to smooth the observed transition and at-risk processes at a specific covariate value $\underset{\sim}{x}$ for the different stages in the simulation model. From these smoothed estimates for each of the observed transition time points we derive their corresponding empirical stage occupation probabilities $P^{\underset{\sim}{x}}_{t_i0}, P^{\underset{\sim}{x}}_{t_i1}, P^{\underset{\sim}{x}}_{t_i2}, P^{\underset{\sim}{x}}_{t_i3}, P^{\underset{\sim}{x}}_{t_i4}$; $i = 1, 2, \cdots, M$, which provide sufficiently close approximations of their population counterparts that generate the original multistate model data. In general for a particular time point t we define the empirical estimates of the stage occupation probabilities as

$$\widehat{p}^{E}_{j}(t|\underset{\sim}{x}) = \begin{cases} = U_1 & \text{if } t \leq t_1 \\ = P^{\underset{\sim}{x}}_{t_{i-1},j} & \text{if } t_{i-1} \leq t < t_i \\ = P^{\underset{\sim}{x}}_{t_M,j} & \text{if } t \geq t_M \end{cases}$$

where $j = 0, 1 \cdots 4$ represent the five stages in the model and and $U_1 = (1, 0, 0, 0, 0)$.

For this study we generate the multistate model data under both the semi-markov as well as markov settings with 25% right censoring and evaluate the estimated stage occupation probabilities for the three different sample sizes 100, 500 and 1000, as done before in section 4.2.4. Following the same notations, we

define the average L1-distance for comparing the estimated probabilities from our method with their corresponding empirical counterparts:

$$\Delta_2 = E\int |\widehat{p}_j^{E}(t|\underset{\sim}{x}) - \widehat{p}_j^{C}(t|\underset{\sim}{x})| d\widehat{F}_n(t).$$

We use the estimators based on a Cox's model (Cox, 1972) as a benchmark. In this alternative technique, the state-to-state transition hazards in the multistate model are estimated from the Cox's regression, with the baseline hazard function being approximated by the Breslow's method (Breslow, 1972). These estimated local state-to-state transitions are then used to build the cummulative transition hazard matrix that is in turn used to derive the estimates of the stage occupation probabilities $(\widehat{p}^{Cox}(t|\underset{\sim}{x}))$ at a specified value $\underset{\sim}{x}$ for the two covariates $X_1$and X_2, from the standard Aalen-Johansen's formula. We aim to conduct a comparative evaluation of this altenative technique with our method by exploring the average difference of their corresponding estimates from the empirical stage occupation probabilities conditioned upon $\underset{\sim}{x}$ $(\widehat{p}_j^{E}(t|\underset{\sim}{x}); j = 0, 1\cdots 4)$. To that end, we define the following average L1-distance to visualize the average bias of these estimators from the true conditional stage occupation probabilities:

$$\Delta_3 = E\int |\widehat{p}_j^{E}(t|\underset{\sim}{x}) - \widehat{p}_j^{Cox}(t|\underset{\sim}{x})| d\widehat{F}_n(t).$$

We perform the entire simulation study with 500 Monte-Carlo replications and jointly report the average L1-distances Δ_2, Δ_3 in Tables 4.5.3 and 4.5.4. For our purpose we specifically take $\underset{\sim}{x}$ to be the vector of medians for the generated values of the two covariates. Evidently, under 25% right censoring, for both the

simulation settings our IPCW-backfitting method gives a comparable performance with respect to the semi-parametric technique based on the Cox's model. In particular our method exhibits a consistently better performance for several stages under a relatively smaller sample size $(100, 500)$ and gets closer to the method based on the Cox's model as the number of subjects in the network increase upto 1000. Specifically, under the conditionally markov model the performance of our method seems to have a better edge over the alternative technique. Additionally, under both the settings (with the only exception of stage 4 in the markov model), the estimated average L1-distances from both the two methods followed a gradually decreasing pattern along with the increase in sample size from 100 to 1000. This demonstrates the convergence of the conditional stage occupation probabilities estimated from the two methods towards their corresponding empirical versions evaluated at the median covariate values. Moreover, the standard deviations of the distances were all less than 0.15 and decreased gradually along with the increase in sample size from 100 to 1000, thereby demonstrating the improving precision of the estimates with an increment in the number of individuals/subjects in the model. Overall from the perspective of a comparative evaluation we find that our method is capable of producing reasonably accurate and precise estimates of the occupation probabilities for the different stages in the multistate network and competes well with the alternative semi-parametric technique, even if the hazard functions of the underlying data generating model doesn't follow a simple pattern.

4.2.6. Bootstrap Confidence Intervals

We illustrate the precision of our method by constructing a set of pointwise bootstrap confidence bands for the occupation probabilities of the different stages in the simulated multistate model. For this purpose we collapse the two stages 3 and 4 in the simulation model to avoid the heavy computational burden.

Thus the model now consists of the three stages $0, 1$ and 2. We generate the state-to-state transition times from this model under 25% right censoring controlled by the two covariates as discussed before in the beginning of section 3. In addition we now generate the first covariate X_1 as the absolute value of a random variable following a $N(\mu, 1)$ distribution, with $\mu \sim N(10.5, 50^2)$ and X_2 from an exponential distribution with mean 40.

In this study we adopt a resampling scheme where bootstrap samples of the individuals are drawn repeatedly to generate their corresponding set of state-to-state transition times and the covariate values in the concerned network. The generated covariate values are perturbed by a $N(0, \widetilde{h})$ random variable, where h is the original kernel smoothing bandwidth selected by the method due to Wand and Jones, (1995) and $\widetilde{h} = h^p$ ($p = 0.8$, if $h < 1$ and $= 1.1$, if $h > 1$, to provide a larger bandwidth for the bootstrap scheme). The transition times and censoring times corresponding to the resampled individuals coupled with the resampled covariate values give rise to the combined bootstrapped dataset for the model. At each iteration b we conditionally re-estimate the state occupation probability at the median covariate vector $\underset{\sim}{x}$ $(\widehat{p}_j^b(t|\underset{\sim}{x}, h))$ for every stage j $(j = 1, 2, 3)$ based on the b-th resample $(b = 1, 2 \cdots 1000)$.

For the j-th stage let $\Delta_\alpha^j(t)$ be the α-th bootstrap percentile for the distribution of $[\widehat{p}_j^b(t|\underset{\sim}{x}, h) - \widehat{p}_j(t|\underset{\sim}{x}, \widetilde{h})]$ where $\widehat{p}_j(t|\underset{\sim}{x}, \widetilde{h})$ is the counterpart of the original estimated probability $\widehat{p}_j(t|\underset{\sim}{x})$, with h being replaced by the larger bandwidth vector $\widetilde{h}$. In essence, the selection of a different bandwidth in the estimation method compensates for the inherent estimation bias owing to the bootstrapping mechanism (Datta and Sundaram, 2006). Then the $(1 - \alpha) \times 100\,\%$

pointwise confidence interval for the true conditional probability $p_j(t|\underset{\sim}{x})$ of occupying stage j in the multistate model at time t is given by

$$[max(0, \widehat{p}_j(t|\underset{\sim}{x}) - \Delta^j_{1-\alpha/2}(t)), min(1, \widehat{p}_j(t|\underset{\sim}{x}) - \Delta^j_{\alpha/2}(t))].$$

We take $\alpha = 0.05$ and construct 95% bootstrap confidence intervals for the empirical mean of $\widehat{p}_j(t|\underset{\sim}{x})\,(j = 0, 1, 2)$ at the first (q_1^t), second (q_2^t) and third quartiles (q_3^t) of the generated time points and estimate their corresponding empirical coverage proportions. The inherent model bias being unavoidable, these intervals are expected to give a reasonably precise idea of the true coverage probabilities. We consider two different sample sizes: $100, 500$ and perform a simulation study with 500 Monte-Carlo replications (using 500 bootstrap iterations inside each of them) to get the respective coverage values at the medians of the two covariates (Table 4.5.5). Evidently, the bootstrap confidence intervals constructed with the estimates from our method cover the empirical conditional stage occupation probabilities with a reasonably high precision that gradually gets better as the sample size increases from 100 to 500.

4.2.7. Tests for Regression Effects and a Power Study

Using our model, we test whether the covariates indeed have any significant effect on the stage occupation probabilities in the simulated multistate model. For this purpose we use the reduced network discussed in the previous sub-section 4.2.6 and estimate the conditional stage occupation probabilities at two different values of the covariates X_1 and X_2, namely $\underset{\sim}{z}_1 = (x_{11}, x_{12})$ and $\underset{\sim}{z}_2 = (x_{21}, x_{22})$. To that end we define the following L1-distance for each stage j in the model, $j = 0, 1, 2$:

$$\Delta_4 = E\int |\widehat{p}_j^C(t|\underset{\sim}{z}_1) - \widehat{p}_j^C(t|\underset{\sim}{z}_2)| d\widehat{F}_n(t)$$

Now we repeatedly generate two independent resamples from the transition and covariate distributions in the 3-stage network. This ensures that the samples are being effectively drawn from the null distribution where the state-to-state transition times are not affected by the covariates. Let Δ_4^{*b} be the value of Δ_4 computed from the b-th bootstrap resample, $b = 1, 2, \cdots B$. Then the p-value for testing the effect of the two covariates on the stage occupation probabilities is given by $p = \frac{1}{B}\sum_{b=1}^{B} I(\Delta_4^{*b} \geq \Delta_4)$ and the null hypothesis of no covariate effect is rejected at 5% level of significance if $p < 0.05$.

We perform this test using 1000 Monte Carlo replications with $B = 1000$ bootstrap iterations inside each of them, using $n = 100$ and 500 individuals and 25% censoring in the model. Specifically, we take $\underset{\sim}{z}_1$ and $\underset{\sim}{z}_2$ to be the first and third quartiles of the generated values for the two covariates X_1 and X_2, respectively. We compute the power by using two values of the parameter $c\,(10, 20)$ that is multiplied with the mean of the two covariates to generate their additive effect on the individual transition times in the network under the two alternative models. The sizes of the test for all the three stages $0, 1$ and 2 are found to be reasonably low $(\text{close to the nominal level of } 0.05)$ while the powers are observed to fall towards the higher end with an expected rise along with an increase in the number of individuals in the network (Table 4.5.6). Thus the results being in agreement with our expectations we conclude that the two covariates indeed have a significant

impact on the state-to-state transitions of the individuals and consequently cast a substantial effect on the occupation probabilities of all the three stages in the model.

4.3. Real Data Applications

4.3.1. Bone Marrow Transplant Data

In this section we illustrate the performance of our method with the well known bone marrow transplant study (Copelan et.al., 1991). The data gathered from this study provides the status information and their corresponding time of onset, for a set of 137 patients with leukemia (ALL/AML), who have been followed upto death/relapse, after receiving an initial bone marrow transplant. This data has been extensively cultivated in previous research works with varying formulations of the underlying model structure. In our analysis, we visualise the progression of these patients through different intermediate stages as a multistate network with the following 7 states, namely : 1 - root state/node denoting the receival of the bone marrow transplant, 2 - Developing Acute Graft versus host disease (GVHD), 3 - Returning of the platelets to normal levels (platelet recovery), 4 - Returning of the platelets to normal levels after developing Acute GVHD, 5 - Developing Acute GVHD after platelet recovery, 6 - Developing Chronic GVHD and 7 - Relapse/Death. (see Figure 4.6.2). Along with the time to event data for the different stages in the network, we also have information on a number of covariates for all the 137 individuals under study. Out of them we have selected the two continuous covariates: patients age and donor's age for demonstrating the efficacy of our method. For a detailed description of the entire dataset we refer the reader to Klein and Moeschberger (Klein and Moeschberger, 1997).

Our fundamental objective is to estimate the occupation probabilties for the 7 states in the model, by using the IPCW- based backfitting technique with the two

selected covariates. We use the internal time-dependent covariate of stage occupation for estimating the IPCW weights via the Aalen's linear hazard model as discussed before. Using these estimated weights we compute the estimated probabilities $\widehat{p}_j(t|\underset{\sim}{x})$ at the median covariate vector $\underset{\sim}{x} = (28, 28)$ for every stage j in the model and plot them along the scale of all the observed transition times in the data along with their corresponding 95% pointwise bootstrap confidence intervals .

Figure 4.6.4 exhibits a graph of 7 plots showing the estimated conditional occupation probabilities of an individual of age 28 years with bone-marrow transplant received from a donor of the same age to occupy each of the 7 stages in the multistate model. The plot for stage 1 clearly represents a gradually decreasing pattern along with the progression of time as is expected from the fact that with an increase in the time span more and more individuals move out of the initial starting state. On the other hand the estimated probabilities for the terminal stage 7 (absorbing state) reflects a gradually increasing trend that is evident from a higher risk of transition for individuals towards the absorbing state (death/relapse) along with time. For the transient stages $2, 3, 4, 5, 6$ we observe a mixed-pattern (increment in the beginning followed by a gradual decline) that is controlled by their intermediate positions in the model leading to a varying flow of individuals through them along with time. In addition, the 95% confidence intervals demonstrate the reasonably good precision of our method in terms of estimating the occupation probabilities for the different states in the model.

We construct another plot representing the conditional stage occupation probabilities for the individuals on a specified grid of the two covariates (constructed by taking 10 values along each co-ordinate direction) at a specific time point (105 days, which is closest to the median time of the dataset) (see Figure

4.6.6). From the resulting graph we can observe the substantial differences between the estimated probability surfaces along with the varying co-ordinate positions of the two covariates for all the stages in the model. The plots reveal interesting features about the stage occupation probabilities at different combinations of the covariate values. For stage 1 we find that the occupation probability increases considerably with a gradual increase in the donor's age (upto 0.07). This implies that patients who receive bone marrow transplants from highly aged donors have comparatively lower chance of moving out of the initial state right after the transplant. For stage 2 we observe a gradual increase in the occupation probability as the patient and donor ages move towards their corresponding median values (28 years). Thus patients of age close to 28 years and receiving transplants from donors in similar age brackets have the highest risk (close to 0.015) of developing an acute Graft vs. Host Disease (Acute GVHD). The surface plot for stage 3 exhibits a mixed pattern which depends heavily on the donor's age. Thus we find patients receiving transplants from donors of lower ages (smaller than the median) have higher chances of platelet recovery (attaining a maximum of 0.5) that drops gradually as the donor's age moves closer to the median and again goes up moderately (upto 0.35) for even higher ages. In case of stage 4 we find that patients of very high ages (50 or more), receiving transplants from highly aged donors (close to 40) have the highest chance of platelet recovery (0.05) after developing the Acute Graft vs. Host Disease. For stage 5 we observe a gradually elevating trend in the estimated probability surface along with an increase in the patient and donor ages. This clearly reveals that highly aged patients receiving transplants from highly aged donors have a reasonably high chance of developing the Acute Graft vs. Host Disease (reaching a maximum of 0.2) after experiencing an initial platelet recovery following the transplant. The pattern of the probability surface plot for stage 6 implies that patients of age moderately lower

than the median age and getting transplants from donors of age less than 10 years have a substantially high risk of developing the Chronic Graft vs. Host Disease (Chronic GVHD) (upto 0.4) that decreases with the increment in both of their ages. For stage 7 the occupation probability is very high (about 0.8) for patients with smaller ages (≤ 10 years) receiving transplants from donors in a similar age group and rises gradually as their ages increase (upto 0.7 for patients and donors of ages close to 50 years). Thus overall we find that the two variables: patient's age and donor's age cast a significant effect on the probability of occupying a stage in the bone marrow transplant model.

4.3.2. Spinal Cord Injury Data

We present another illustration of our method on the Spinal Cord Injury data (Harkema, et. al., 2011; a,b). The data consists of the measurements on different performance criteria for a set of 326 individuals who are enrolled in a locomotor training program after suffering a spinal cord injury. From the time of enrollment into the program the gradual recovery of these individuals is monitored by their repeated evaluations on the basis of several functional indicators, till they are discharged. Walking speed is a significant indicator in this context and is repeatedly measured for the individuals over their follow up time via two separate walking tests: one based on a six minute walk and the other being a 10 m distance walk. Now, depending on their performances in these tests in terms of the maximum walking speed these individuals are subsequently classified into different speed categories. Following clinical benchmarks these categories are represented as specific speed limits. For example a minimum walking speed of 0.44 m/s is required for being able to walk in the community, which increases to 0.7 m/s for walking without any supporting device and is further higher (1.2 m/s) for being able to cross a stoplight (van Hedel and Dietz, 2010). From a graphical perspective the

transitions of the individuals along these different speed benchmarks can be visualized as an example of a multistate model, rather more specifically as a five-state tracking model (Figure 4.6.3). In addition, we also consider three continuous covariates that can play a significant role in controlling the movement of the individuals along the different states in the model. They are: (1) time from the spinal cord injury to enrollment in the program, (2) lower motor score from the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI) exam and (3) treatment intensity given by the ratio of the cummulative number of training sessions received by the individual and his/her duration of enrollment in the program. Our objective is to estimate the conditional occupation probabilities given a specific set of values of these three covariates for all the 5 states in the model. In this analysis we use the internal covariate of stage occupation for building the IPCW weights from the Aalen's linear hazard model. We evaluate the conditional occupation probabilities for all the states in the model at the median values of the covariates and represent them in Figure 4.6.5. An overview of the plots illustrates an expected pattern for the occupation probabilities of the five states in the model. Likewise, the terminal stages 1 and 5 exhibit a monotonic trend that is gradually decreasing for the former and increasing for the latter, as is expected from the enhanced movement of the individuals along with time from the non-ambulatory phase (stage 1) to the state of maximal recovery with the highest walking speed (stage 5). For the transient stages $2, 3$ and 4 we observe a mixed pattern that is attributable to the varying intensity of individuals passing through them along with the progression of time. Specifically, the erratically spiked pattern of the estimated conditonal probability curves for stages 3 and 4 can be potentially attributable to the presence of multiple transition paths connecting them with the other stages in the model.

4.4. Discussion

Non-parametric estimation of the conditional occupation probability distributions for the different stages in multistate networks with multiple covariates is a relatively unexplored area of research. Past studies related to this field have mostly delved upon specific parametric/semi-parametric approaches, applied under simple hazard model assumptions (like Cox's proportional hazards model). These methods provide reasonable answers only when the underlying data generating model is in coherence with the underlying structural assumptions. But in a more general situation, the state-to-state transition hazards in the model can depend upon various individual specific covariates in a lot of different ways. In such cases, the stage occupation probability distributions can change with the covariates and their conditional estimation at specified covariate values can indeed be of considerable statistical importance. As an illustration, for disease state models, this mode of estimation can have biologically significant implications in terms of the individuals falling inside certain particular covariate brackets.

We have developed a convenient technique that can estimate the conditional transition and at-risk processes corresponding to the stage occupation probability distributions by using an IPCW-reweighted version of the backfitting regression principle (Hastie and Tibshirani, 1990). It demonstrates an elegant fusion of a highly flexible regression method that can incorporate several types of complicated covariate effects on the hazards of the estimated stochastic processes and the Inverse-Probability of Censoring Weighted (IPCW) methodology (Datta and Satten, 2001; 2002) that can effectively handle the underlying censoring in the data. The efficacy of the method has been demonstrated through rigorous simulation studies conducted under two different structural settings: Markov and Semi-Markov, with the underlying right censoring being varied from moderate (25%) to heavy (50%). The first study illustrates the consistency of the method in terms of the gradual

decrease in the average L1-distance between our estimates computed for the original uncensored and the observed right censored data, as the sample size is increased from 100 to 1000. The second study compares the performance of our method with a competing technique based on the Cox's proportional hazards model, where the derived results clearly reveal the superiority of our method in terms of producing estimates much closer to the empirically estimated conditional stage occupation probabilities in the model. Moreover the use of the Aalen's linear hazards model in estimating the weights for the IPCW scheme provides a reasonably flexible way of modeling widely different types of censoring hazards.

In several biological applications we may have multistate model data on individuals having a high-dimensional covariate. In such cases use of an efficient dimensional reduction technique under a censored data setup along with our method can indeed provide appropriate non-parametric estimates for the conditional occupation distributions for the different states in the model. We intend to pursue this idea in some future research work.

4.5. Tables

Table 4.5.1: L_1-distances between the estimated stage occupation probabilities for the censored and uncensored data generated from a conditionally semi-Markov model with covariate dependent censoring, at the medians of the given covariate values.

Stage	25% Censoring			50% Censoring		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
0	0.0232	0.0105	0.0078	0.0412	0.0190	0.0141
1	0.0290	0.0177	0.0152	0.0483	0.0318	0.0277
2	0.0285	0.0136	0.0101	0.0479	0.0225	0.0177
3	0.0070	0.0048	0.0041	0.0099	0.0074	0.0069
4	0.0098	0.0060	0.0059	0.0140	0.0108	0.0099

Table 4.5.2: L_1-distances between the estimated stage occupation probabilities for the censored and uncensored data generated from a conditionally Markov model with covariate dependent censoring, at the medians of the given covariate values.

Stage	25% Censoring			50% Censoring		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
0	0.0287	0.0132	0.0096	0.0496	0.0238	0.0169
1	0.0263	0.0169	0.0150	0.0466	0.0323	0.0288
2	0.0314	0.0156	0.0108	0.0512	0.0254	0.0186
3	0.0115	0.0073	0.0063	0.0184	0.0137	0.0138
4	0.0126	0.0090	0.0078	0.0228	0.0164	0.0141

Table 4.5.3: Comparison of the L_1-distances between the estimated stage occupation probabilities and their empirical values, evaluated from our method and the Cox's proportional hazards model at the median covariate values, for a conditionally semi-Markov model with 25% censoring.

Stage	Cox's method			IPCW-backfitting		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
0	0.1203	0.1115	0.1107	0.1458	0.1237	0.1222
1	0.1885	0.1302	0.1175	0.1459	0.1235	0.1042
2	0.1987	0.1156	0.0979	0.1822	0.1551	0.1379
3	0.0442	0.0260	0.0223	0.0221	0.0188	0.0179
4	0.0382	0.0249	0.0210	0.0264	0.0251	0.0219

Table 4.5.4: Comparison of the L_1-distances between the estimated stage occupation probabilities and their empirical values, evaluated from our method and the Cox's proportional hazards model at the median covariate values, for a conditionally Markov model with 25% censoring.

Stage	Cox's method			IPCW-backfitting		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
0	0.1782	0.1561	0.1566	0.1665	0.1448	0.1434
1	0.2151	0.2029	0.1945	0.1554	0.1358	0.1335
2	0.1811	0.1127	0.0985	0.1788	0.1483	0.1483
3	0.0396	0.0390	0.0374	0.0234	0.0211	0.0193
4	0.1113	0.0952	0.0907	0.1026	0.1057	0.1035

Table 4.5.5: Estimated coverage proportions for the empirically estimated conditional stage occupation probabilities at the median covariate values, for the first, second (median) and third quartiles of the generated time points, in the three stage model with 25% censoring.

Stage	$n = 100$			$n = 500$		
	q_1^t	q_2^t	q_3^t	q_1^t	q_2^t	q_3^t
0	0.762	0.998	1.000	0.962	0.998	1.000
1	0.418	0.966	0.998	0.924	0.982	1.000
2	0.586	0.978	1.000	0.924	0.998	0.998

Table 4.5.6: Size and power for testing the significance of the covariate effects under the three stages with 25% censoring.

		$n = 100$			$n = 500$	
	Size	Power		Size	Power	
Stage		$c = 10$	$c = 20$		$c = 10$	$c = 20$
0	0.058	0.974	0.986	0.064	0.998	1.000
1	0.038	0.638	0.766	0.048	0.972	0.972
2	0.058	0.874	0.918	0.042	0.962	0.962

4.6. FIGURES

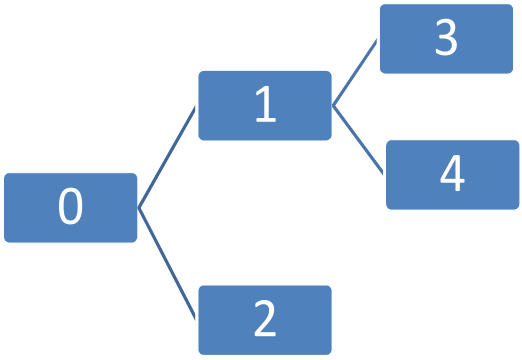

Figure 4.6.1: Network showing the 5 stages and the transition paths interconnecting them used in the simulation studies.

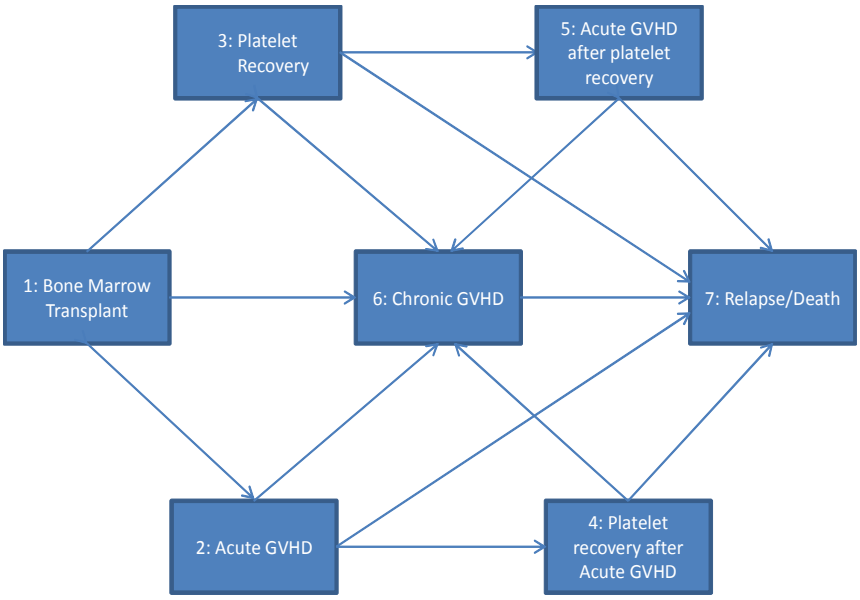

Figure 4.6.2: Network showing the different stages and their mutual transition paths for the Bone-marrow Transplant data.

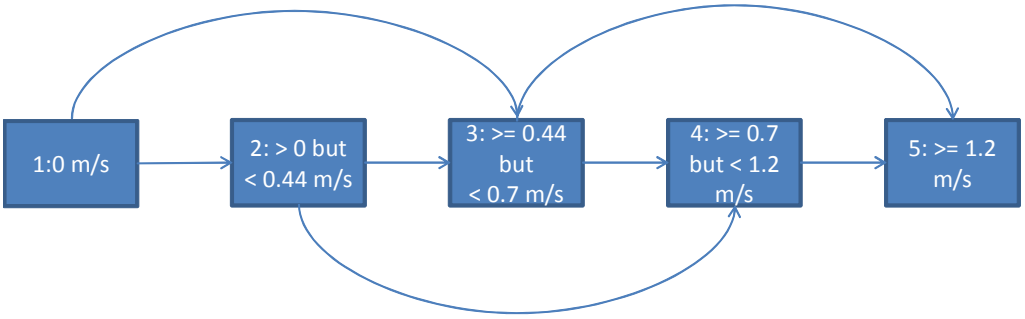

Figure 4.6.3: Network showing the different stages and their mutual transition paths for the Spinal-Cord Injury data.

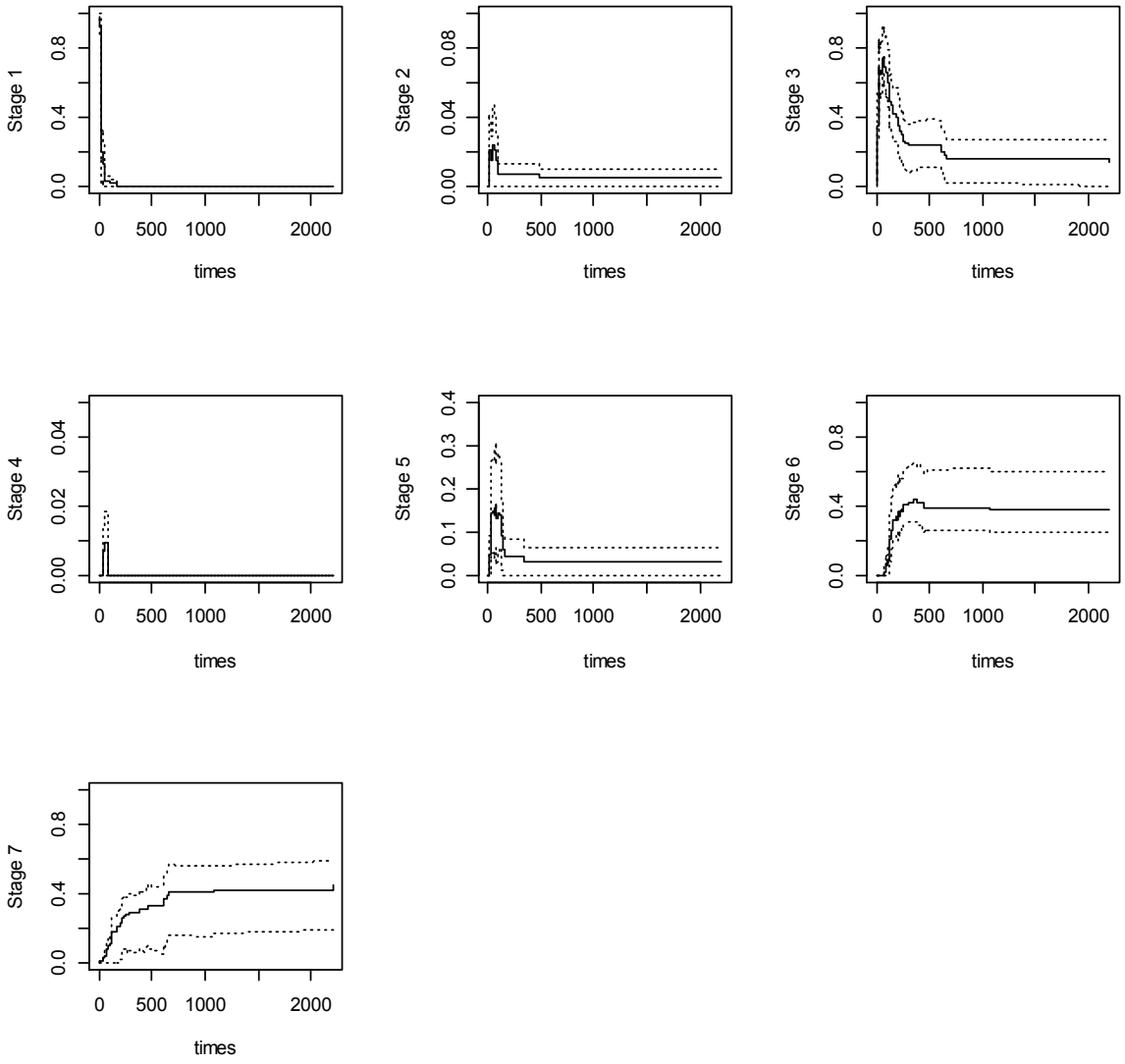

Figure 4.6.4: Plot of the estimated conditional occupation probabilities from our method at the median covariate values along with their corresponding 95% bootstrap confidence intervals (represented by the dotted lines) for the 7 stages in the Bone-Marrow Transplant Data.

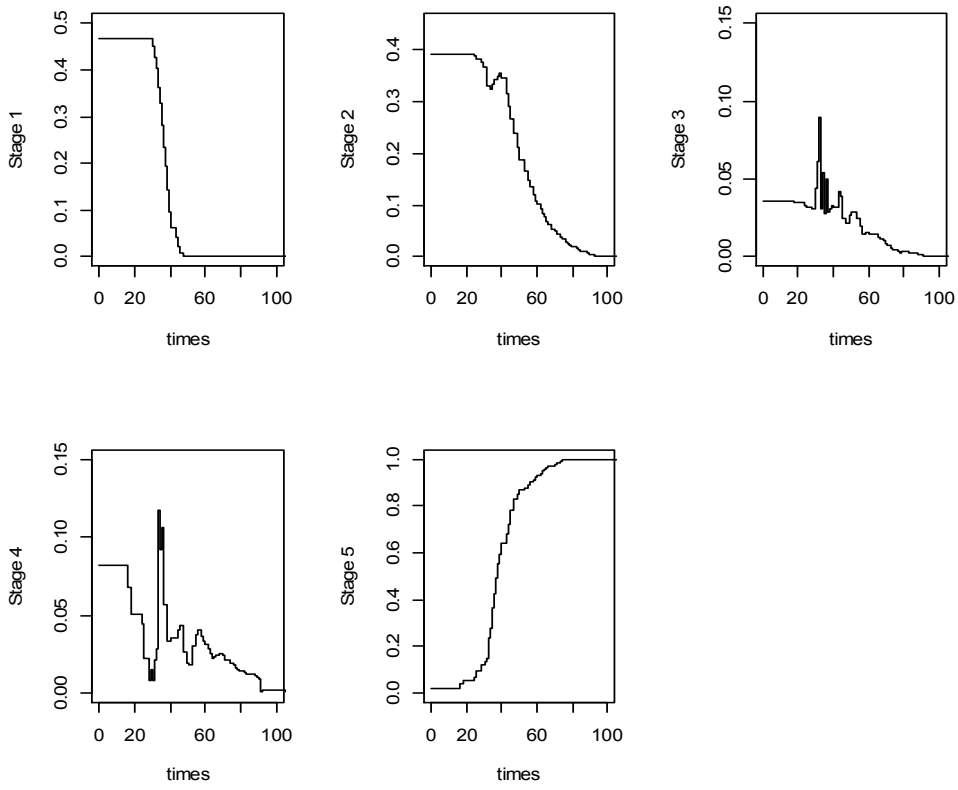

Figure 4.6.5: Plot of the estimated conditional occupation probabilities from our method at the median covariate values along for all the 5 stages in the Spinal-cord Injury Data.

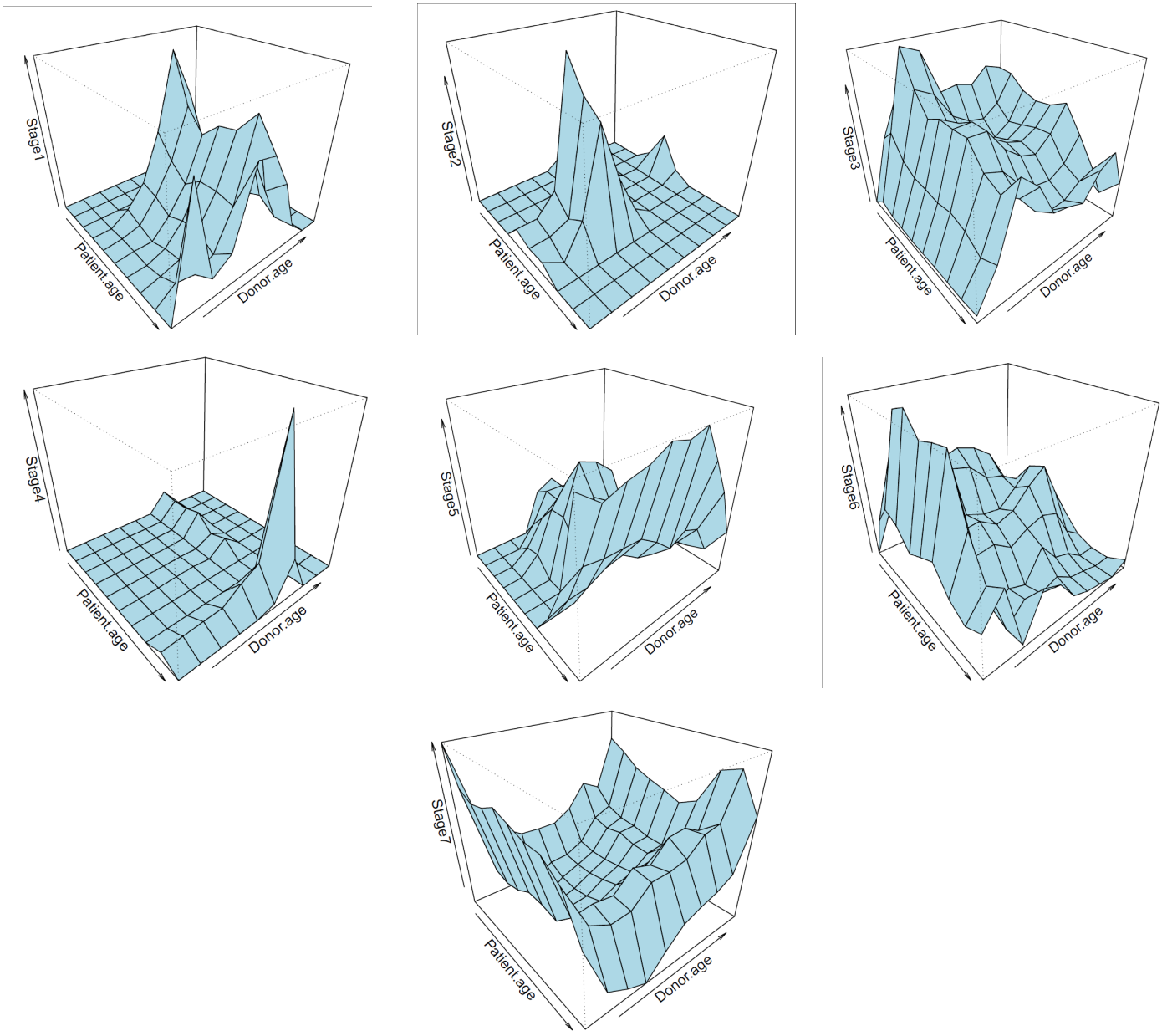

Figure 4.6.6: Bivariate plot of the estimated conditional occupation probability surfaces at a specified grid of covariate values for all the 7 stages in the Bone Marrow Transplant Data.

CHAPTER 5: TESTING THE EQUALITY OF THE WAITING TIME DISTRIBUTIONS BETWEEN TWO GROUPS OF INDIVIDUALS USING AN IPCW-BASED MANN-WHITNEY U-STATISTIC AFTER ADJUSTING FOR AVAILABLE COVARIATES

5.1 Data structure and notations

Let us envisage a scenario where we have data on the right censored entry and exit times of individuals from two independent populations (groups). Let $X_{i,j}^*$ and $V_{i,j}^*$ denote the original uncensored entry and exit times for the i-th individual in the j-th group, $C_{i,j}$ be a common censoring variable which affects both of them and is assumed to be independent of the pair $(X_{i,j}^*, V_{i,j}^*)$. In addition we have information on a covariate Z for all the individuals in the model. To that end, let $Z_{i,j}$ denote the observed value of Z for the i-th individual in the j-th group. Thus overall, our entire observed dataset is composed of the 5-tuples $(X_{i,j}, \delta_{i,j}, V_{i,j}, \eta_{i,j}, Z_{i,j}), (i = 1, 2 \cdots n_j;\ j = 1, 2;\ n_1 + n_2 = n)$, where $X_{i,j} = min(X_{i,j}^*, C_{i,j})$ and $V_{i,j} = min(V_{i,j}^*, C_{i,j})$ are the observed right-censored entry and exit times for the i-th individual in the j-th group and $\eta_{i,j} = I(C_{i,j} \geq X_{i,j}^*)$ and $\delta_{i,j} = I(C_{i,j} \geq V_{i,j}^*)$ are the corresponding censoring indicators.

Define $W_{i,j}^* = V_{i,j}^* - X_{i,j}^*$ as the actual uncensored waiting time for the i-th individual in the j-th group and $W_{i,j}$ be its corresponding observed version in the right censored data. Clearly, $W_{i,j}$ is uncensored and equals $W_{i,j}^*$ if and only if $\eta_{i,j} = 1$.

In the absence of censoring the Mann-Whitney U-statistic to be used for comparing the marginal waiting time distributions between the two different groups is given by

$$U = \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} I(W^*_{i_1,1} \leq W^*_{i_2,2}).$$

But, in the presence of right censoring not all waiting times can be observed in both the two groups and hence they need to be replaced by their corresponding right censored values. Fan and Datta (2013) proposed a modified Mann-Whitney U-statistic that compensates for this selection bias by using the IPCW reweighting principle (Datta and Satten, 2001). In this work we propose an extension of the classical Mann-Whitney U-statistic that can be used to build a test for comparing the waiting time distributions between the two groups, after adjusting for subject (and group) level covariates Z. For this purpose we pursue a regression approach to build a set of model residuals which can in turn be used to build such a modified U-statistic.

Although other types of regression models (both parametric as well as non-parametric) can be used for the purpose of covariate adjustment, we choose a transformation model for the waiting times in order to calculate the residuals. To this end, we define the following two accelerated failure time (AFT) models corresponding to the waiting times of the individuals from the two groups:

$$log\, W_{i_1,1} = \alpha_1 + Z_{i_1,1}\beta_1 + \epsilon_1,$$

$$log\, W_{i_2,2} = \alpha_2 + Z_{i_2,2}\beta_2 + \epsilon_2,$$

where α_1, α_2 are the intercepts, β_1, β_2 are the regression coefficients corresponding to the two covariates Z_1, Z_2 and ϵ_1, ϵ_2 are respectively the random error terms for the two models.

Now, as the waiting times of the subjects in the two groups are right censored, the least-square fitting equations for these two models need to be modified following the IPCW reweighting principle (Datta and Satten, 2001). Thus the estimated coefficients $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are derived from the score equations obtained after minimizing the following two criteria respectively:

$$\Delta_1 = \sum_{i_1=1}^{n_1} (log\, W_{i_1,1} - \alpha_1 - Z_{i_1,1}\beta_1)^2 \frac{\delta_{i_1,1}}{K_1(V_{i_1,1} -)}$$

$$\Delta_2 = \sum_{i_2=1}^{n_2} (log\, W_{i_2,2} - \alpha_2 - Z_{i_2,2}\beta_2)^2 \frac{\delta_{i_2,2}}{K_2(V_{i_2,2} -)}$$

If the original AFT models corresponding to the two population groups were known then, in absence of censoring, the actual residuals from the two groups after eliminating the covariate effects would have been given by $R^*_{i_j,j} = log(W^*_{i_j,j}) - Z_{i_j,j}\beta_j; j = 1, 2.$ We denote the corresponding censored versions estimated by minimizing the IPC weighted OLS criteria Δ_1 and Δ_2 by $R_{i_j,j} = log(W_{i_j,j}) - Z_{i_j,j}\widehat{\beta}_j; j = 1, 2.$

Our main objective is to estimate the parameter $\theta = P(R^{*}_{i_1,1} \leq R^{*}_{i_2,2})$. We propose a modified IPCW-based Mann-Whitney U-statistic that can be used to develop an inferential framework on θ.

5.2 The modified Mann-Whitney U-statistic

We define the following U-statistic based on the residuals $R_{i_1,1}$ and $R_{i_2,2}$ obtained after fitting the reweighted accelerated failure time (AFT) models on the waiting times of the individuals in the two groups:

$$\widehat{U} = \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \frac{I(R_{i_1,1} \leq R_{i_2,2}) \delta_{i_1,1} \eta_{i_2,2}}{\widehat{K}_1(V_{i_1,1} -) \widehat{K}_2(X_{i_2,2} + C(Z, \widehat{\beta}) W_{i_1,1} -)}$$

$$= \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \frac{I(e^{-Z_{i_1,1}\widehat{\beta}_1} W_{i_1,1} \leq e^{-Z_{i_2,2}\widehat{\beta}_2} W_{i_2,2}) \delta_{i_1,1} \eta_{i_2,2}}{\widehat{K}_1(V_{i_1,1} -) \widehat{K}_2(X_{i_2,2} + C(Z, \widehat{\beta}) W_{i_1,1} -)} \quad (1)$$

where $C(Z, \widehat{\beta}) = e^{-(Z_{i_1,1}\widehat{\beta}_1 - Z_{i_2,2}\widehat{\beta}_2)}$, $\delta_{i_1,1} = I(C_{i_1,1} \geq V^{*}_{i_1,1})$ and $\eta_{i_2,2} = I(C_{i_2,2} \geq X^{*}_{i_2,2})$ are the two censoring indicators corresponding to the two groups as defined earlier.

5.3 **Simulation Studies**

We primarily focus on two different simulation studies concerned with this newly developed U-statistic. Both the analyses are conducted on a right censored multistate data under two different structural models: uncorrelated (semi-markov) and correlated. In the first study, the objective is to explore the average bias and variance of $\widehat{U}$, while the second is based on the implementation of a modified Mann-

Whitney U-test using $\widehat{U}$ in order to compare the waiting time distributions between the two groups of individuals after adjusting for the available information on their respective covariate values.

5.3.1 **An uncorrelated model**

In this setting the state waiting times for the individuals under both the two groups are generated independently of their state entry times. To start with, we generate data on a covariate $Z_{ij}\,(i = 1, 2\cdots n_j; j = 1, 2)$ for all the n_1 individuals in group 1 and the n_2 individuals in group 2. Additionally, we consider two different settings: one in which the covariate distributions are similar in the two groups and the other in which they are different. Specifically, under the first setting we simulate Z_{ij} from a $N(1.5, 1)$ distribution $(i = 1, 2\cdots n_j; j = 1, 2)$, while under the second setting $Z_{i1}\,(i = 1, 2\cdots n_1)$ is generated from a $N(1.5, 1)$ and $Z_{i2}\ (i = 1, 2\cdots n_2)$ from a $N(2.5, 1)$ distribution. Now, using the simulated covariate information the waiting time for the individual i in group j, W^*_{ij}, is generated from a lognormal distribution with log-mean parameter $\alpha_j + Z_{ij}\beta_j; i = 1, 2\cdots n_{_j}, j = 1, 2.$ and unit log-scale parameter. Here $\alpha_1 = \alpha_2 = 0.5; \beta_1 = \beta_2 = 0.3$ for setting 1 and $\beta_1 = 0.3$ and $\beta_2 = 0.5$ for setting 2. In both the cases the state entry times (X^*_{ij}) for the individuals from both the groups are generated independently, from a standard log-normal distribution.

The censoring times in the two groups (C^*_{ij}) are also generated from two lognormal distributions with unit log-scale, but with varying log-mean parameters depending on the desired censoring rates in the two groups.

Overall, under each setting we consider two different group sizes: 25 and 50. Additionally, we consider two different censoring rates: moderate (25%) and heavy (50%) in order to incorporate two different censoring patterns in the data.

5.3.2 **A correlated model**

In this setting, we first generate the log-entry and log-waiting times for the individual i in group j (X_{ij}^{*}) from a bivariate normal distribution with the log-mean parameter vector $(0, \alpha_j + Z_{ij}\beta_j)$ (with Z and β being defined in a similar way as in the previous section) and dispersion matrix $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. In this way the marginal distributions for the state entry and waiting times are univariate log-normal and a functional dependence is established between them owing to the underlying correlation factor.

The censoring times are generated in a similar way as discussed in the previous section, using an unit scale but with varying log-mean parameters in order to achieve different censoring patterns in the data.

5.3.3 **Bias and Variance Study**

We perform the first simulation study under all the proposed settings using 1000 Monte-Carlo replications. Tables 5.6.1 and 5.6.2 report the average bias and variances of $\widehat{U}$ for the two sample sizes 25 and 50, under the uncorrelated (semi-markov) setting with varying censoring patterns in case the covariate distributions are same or different in the two groups. Tables 5.6.3 and 5.6.4 report the results from a similar study when the multistate model is generated under the correlated

setting. In both the settings under similar covariate distributions, we observe that the average bias (empirical bias) and empirical standard deviation of our IPCW-modified U-statistic $\widehat{U}$ decrease with an increase in the group sample sizes. But with different covariate distributions in the two groups the biases seem to have a slightly inconsistent pattern (which could be due to Monte Carlo errors) although their magnitudes are small in all cases. The standard deviations keep following a monotonic trend as in the previous setting. Additionally, in all the cases, the bias and variance of $\widehat{U}$ increase as the censoring rates go up in the two groups, attributing to a larger proportion of unobserved waiting times in the data.

5.3.4 **Testing the equality of waiting time distributions between the two groups of individuals**

In this simulation study our objective is to perform a size and power analysis using a test-statistic based on $\widehat{U}$ in order to examine whether the waiting time distributions in the two groups exhibit any significant difference after adjusting for the individual specific covariates. For this purpose we consider an uncorrelated model (as discussed in Section $5.3.1$) under 25% censoring, with the covariate distributions being different in the two groups of individuals. To that end we define the following test-statistic T :

$$T = 0.5[\widehat{U}(1,2) + 1 - \widehat{U}(2,1)],$$

where $\widehat{U}(1,2)$ is the value of $\widehat{U}$ computed with the observations on the 5-tuples, from the two groups 1 and 2: $(X_{i,j}, \delta_{i,j}, V_{i,j}, \eta_{i,j}, Z_{i,j}), (i = 1, 2 \cdots n_j;\ j = 1, 2;\ n_1 + n_2 = n)$, being in their natural order. $\widehat{U}(2,1)$ is the version of $\widehat{U}$ with this order being reversed. Our objective is to use T for testing the null hypothesis H_0 : the waiting time distributions after covariate adjustment are the same in the

two groups vs. the alternative hypothesis H_1 : the waiting time distributions after covariate adjustment are different in the two groups.

Now from the large sample theory of U-statistics, T is expected to follow an asymptotically normal distribution under H_0, with mean 0.5 and variance σ_T^2, say. We estimate the asymptotic variance of T by implementing the bootstrap resampling technique. For this purpose we generate a resample of size 500 (with replacement) from the observations on the 5-tuple $(X_{i,j}, \delta_{i,j}, V_{i,j}, \eta_{i,j}, Z_{i,j}); i = 1, 2 \cdots n_j; j = 1, 2,$ simulated in each of 1000 Monte-Carlo replications. We compute the values of T for each of the bootstrap samples and take their average as the estimated asymptotic variance of T $(\widehat{\sigma}_T^2)$. With these estimates we construct the 95% bias corrected confidence interval for the actual population mean of T (μ_T) as:

$$[\widehat{\mu}_T - 1.96\widehat{\sigma}_T, \widehat{\mu}_T + 1.96\widehat{\sigma}_T].$$

We calculate the proportion of times the mean of T (μ_T) under the null distribution (0.5) is not covered by this interval, in order to get the size and power values for the corresponding model settings controlled by the parameters α_1 and α_2. As a competing method, we construct a similar test using the Fan-Datta U-statistic (Fan and Datta, 2011):

$$\widehat{U}_{FD} = \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \frac{I(W_{i_1,1} \leq W_{i_2,2}) \delta_{i_1,1} \eta_{i_2,2}}{\widehat{K}_1(V_{i_1,1} -) \widehat{K}_2(X_{i_2,2} + W_{i_1,1} -)}$$

Results from the size and power analyses for both the two methods, under the uncorrelated model are graphically represented in Figure 5.7.1.

Clearly, from the figure we can see that under both of the two group sample sizes $(50$ and $100)$ the test maintains a reasonable size that is just marginally higher than the nominal level (5%) and the power increases gradually as the waiting time distributions in the two groups differ more and more owing to the extent of variation in the intercept parameters α_1 $(=0.5)$ and α_2 corresponding to groups 1 and 2, respectively. Moreover, the size gets closer to the nominal level and the power values rise up as the number of subjects in the two groups increases to 100. In contrast, the Mann-Whitney U-test based on the Fan-Datta U-statistic (Fan and Datta, 2011) yields inflated size for a large number of subjects and substantially lower power values compared to the test based on our U-statistic.

5.4 **Application to the Spinal Cord Injury Data**

In this section we illustrate an application of the Mann-Whitney test based on our IPCW-modified U-statistic $\widehat{U}$ on the well-known Spinal Cord Injury data (Harkema, et. al. 2011; a,b). The dataset has been discussed in detail in Chapter 4. Overall, it consists of the performance measurements from different functional indicators for 326 individuals who are enrolled in a locomotor training program after suffering a spinal cord injury. Continued evaluation of the walking speed constitutes a fundamental part of the entire monitoring process of the patients following their time of enrollment in the program. On the basis of the performances in two separate walking tests these individuals are subsequently classified into different speed categories by virtue of their maximum walking speeds in the two tests. Following clinical benchmarks these categories are represented as specific speed limits which can be jointly visualized as a five-state tracking model (Figure 4.7.3). For our illustrative purpose, we collapse the stages 4 and 5 in Figure $4.7.3$. to generate a reduced model with the three states $0, 1$ and 2. We consider three

individual specific covariates that may potentially control the movement of the individuals along the different states in the model. These covariates are: (1) time from the spinal cord injury to enrollment in the program, (2) lower motor score from the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI) exam and (3) treatment intensity given by the ratio of the cummulative number of training sessions received by the individual and his/her duration of enrollment in the program.

Now, we create two different groups of injured patients depending on their initial phase at the time of enrollment (1 or 2). Our objective is to use the modified IPCW-based U-statistic to compare the sojourn time distributions at stage 2 between these two categories of patients after adjusting for their information on the three covariates discussed earlier.

Application of the Mann-Whitney test based on our U-statistic (as discussed in Section 5.3.4) gives the absolute value of the test-statistic T as 0.639 (< 1.96). But, using the Fan-Datta U-statistic (Fan and Datta, 2013) we get $|T| = 4.950$ (> 1.96). This demonstrates that treatment intensity (covariate) indeed casts a significant effect in creating a substantial difference between the sojourn time distributions for the injured patients enrolled in the initial phases 1 and 2. However, this effect can either be due to the difference between the covariate distributions in the two groups of patients or a variation in its impact over the two groups (characterized by the regression coefficients) or a combined effect from both of them.

5.5 **Discussion**

U-statistics are fundamental objects in theoretical statistics and provide a broad generalization of different types of commonly used measures in the statistical analyses (sample mean, variance, etc.). Different types of statistics with complicated expressions (that are not readily amenable to algebraic treatments) can be expressed as U-statistics, or approximate U-statistics, thereby facilitating their asymptotic treatments (consistency, asymptotic normality, etc.) in an unified fashion.

Mann-Whitney U-statistics (Mann and Whitney, 1947) are well known in this context and can be used to test the equality of two probability distributions by formulating an indicator kernel function in terms of the observed sample values on their corresponding random variables. Fan and Datta (2013) initiated the development of a modified Mann-Whitney U-statistic from a right censored data on the sojourn times of individuals classified into two groups. Specifically, their work was focused on the use of this modified statistic to compare the stage waiting time distributions between two groups of subjects/individuals progressing through the different branches of a multistate network affected by right censoring. In the present context we have pursued an extension of this work to build a different version of the Mann-Whitney U-statistic that uses the concept of Inverse Probability of Censoring (Fan and Datta, 2013) to tackle the censoring in the data and adjusts for respective subject-specific covariates to ensure a more accurate inference on the comparison of the waiting time distributions between the two groups of individuals.

We have demonstrated the improved performance of our modified U-statistic in terms of its lower empirical bias and standard deviation through extensive simulation studies conducted under both semi-markov and markov settings with

similar and varying covariate effects in the two groups of individuals affected by different censoring patterns. In addition we have illustrated the usefulness and superior performance of the test based on our modified Mann-Whitney U-statistic (compared to the test based on the Fan-Datta U-statistic), by the detailed power analyses under both the structural settings. Moreover, application of our modified Mann-Whitney U-test on the Spinal-Cord Injury Data do not show a significant difference between the covariate adjusted sojourn time distributions of the two categories of patients starting in the initial phases 1 and 2.

In future, we aim to derive the asymptotic distribution of our proposed U-statistic and apply it to compare the waiting time distributions between two groups of individuals in a right censored multistate model data using available high-dimensional covariate information.

5.6 TABLES

Measure	Censoring $= 25\%$		Censoring $= 50\%$	
	$n_1/n_2 = 25$	$n_1/n_2 = 50$	$n_1/n_2 = 25$	$n_1/n_2 = 50$
Empirical Bias	-0.007	-0.006	-0.025	-0.019
Emprical SD	0.166	0.122	0.183	0.149

Table $5.6.1$: Empirical bias and standard deviation of our U-statistic $\widehat{U}$ under an uncorrelated model with similar covariate distributions in the two groups of individuals.

Measure	Censoring $= 25\%$		Censoring $= 50\%$	
	$n_1/n_2 = 25$	$n_1/n_2 = 50$	$n_1/n_2 = 25$	$n_1/n_2 = 50$
Empirical Bias	-0.005	0.012	-0.014	0.007
Emprical SD	0.182	0.140	0.208	0.168

Table $5.6.2$: Empirical bias and standard deviation of our U-statistic $\widehat{U}$ under an uncorrelated model with different covariate distributions in the two groups of individuals.

Measure	Censoring $= 25\%$		Censoring $= 50\%$	
	$n_1/n_2 = 25$	$n_1/n_2 = 50$	$n_1/n_2 = 25$	$n_1/n_2 = 50$
Empirical Bias	-0.006	-0.004	-0.029	-0.012
Emprical SD	0.167	0.120	0.194	0.153

Table $5.6.3$: Empirical bias and standard deviation of our U-statistic $\widehat{U}$ under a correlated model with similar covariate distributions in the two groups of individuals.

Measure	Censoring $= 25\%$		Censoring $= 50\%$	
	$n_1/n_2 = 25$	$n_1/n_2 = 50$	$n_1/n_2 = 25$	$n_1/n_2 = 50$
Empirical Bias	0.00005	0.008	-0.004	0.002
Emprical SD	0.18569	0.141	0.212	0.165

Table $5.6.4$: Empirical bias and standard deviation of our U-statistic $\widehat{U}$ under a correlated model with different covariate distributions in the two groups of individuals.

5.7 Figures

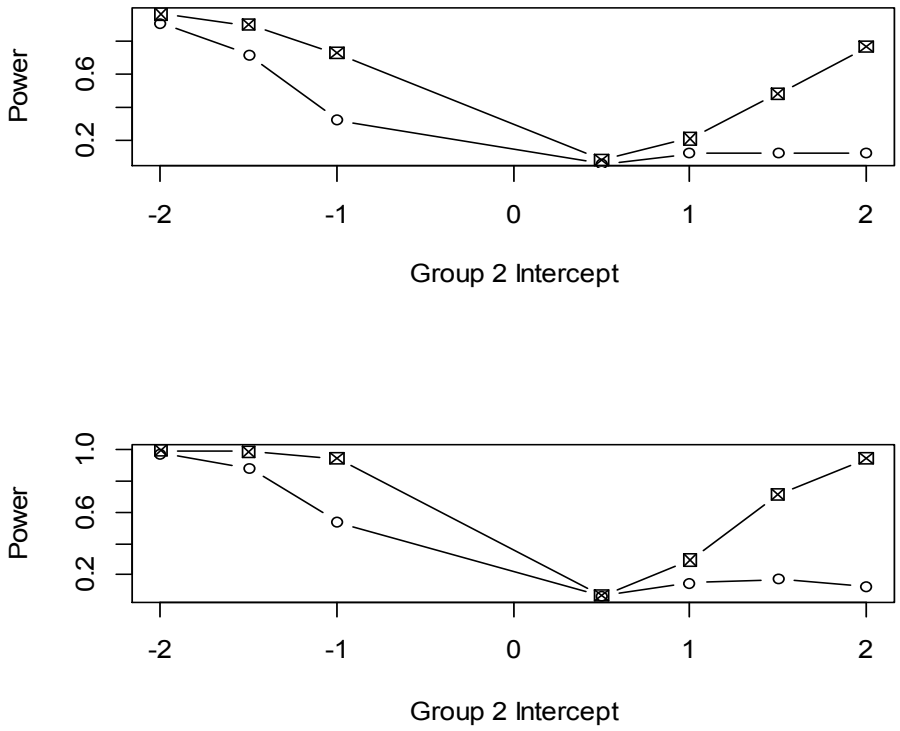

Figure 5.7.1: Power curves for the covariate adjusted IPCW based Mann-Whitney U-test under the uncorrelated model with 25% censoring, using different covariate distributions in the two groups. The upper figure corresponds to a group sample size of 50 while the lower one is for a sample size of 100: the solid curve with checked squares represents the power curve for the test with our U-statistic $\widehat{U}$, while the one with the white circles corresponds to the Fan-Datta U-statistic $\widehat{U}_{FD}$.

CHAPTER 6: EXTENSIONS AND FUTURE RESEARCH

Four different research projects have been discussed in this thesis that are interconnected by the common framework of a regression approach. Novel methods have been developed for each project that are demonstrated through a wide variety of simulation studies along with contexual real-life applications. Different other aspects of the projects still remain that can be pursued under future extensions.

In project-1 there is a scope for two primary developments. Firstly, the theoretical foundation behind the superior performance of our method SVA-PLS can be constructed. This will involve a rigourous study of the statistical properties for the NIPALS algorithm along with the estimates that are obtained from the ANCOVA model after incorporating the PLS based scores as surrogate variables. Secondly, a discrete version of the PLS algorithm can be used under a generalized linear model (GLM) framework (instead of ANCOVA) to develop an alternative natural application on next generation sequence data of gene expression (e.g., RNA-seq data).

The R package *svapls* discussed in project-2 can be developed even more by incorporating several other functions that can perform a wide array of genomic analyses like: serial analysis of gene expression (SAGE) or expression quantitative trait locus (e-QTL) mapping, by using the extracted signatures of the hidden expression heterogeneity from the partial least squares (PLS) algorithm. Also different other FDR controlling techniques can be included in the package to provide

the user with a more generalized multiple testing framework. In addition the selection of the maximum number of surrogate variables in the function *svpls* can also be automated by using a cross-validation technique.

The method developed in project-3 can be significantly extended by constructing an inverse probability of censoring weighted backfitting algorithm under a generalized linear model framework. This will lead to the genesis of a more appropriate methodology for estimating the two right censored counting processes for the transitions and at-risk set of subjects. Additionally, the properties of the estimated functions from the algorithm can be studied over the domains of different time points and covariate values to understand the pattern in which the external variables affect the probabilities of stage occupation for the corresponding subjects in the model.

In project-4, a theoretical derivation of the asymptotic distribution of the U-statistic can be illustrated in some future work. This project can provide a unified framework for studying the impact of different subject-specific covariates on their sojourn time distributions under a multistate network. As for example, a more flexible regression model may be used with multiple covariates to generate the residuals from the two groups of individuals. The statistical properties of the resultant U-statistic can then be explored through detailed simulation studies and real-life data analyses.

REFERENCES

1. Aalen, O.O. Non-parametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics* 1976; **3**: 15-27.

2. Aalen, O.O. Non-parametric inference for a family of counting processes. *Annals of Statistics* 1978; 6: 701-726.

3. Aalen, O.O & Johansen, S. An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 1978; **5**: 141-150.

4. Aalen, O.O. Nonparametric inference for a family of counting processes. *Annals of Statistics* 1978; **6**: 701-726.

5. Aalen, O.O. A model for nonparametric regression analysis of counting processes. *Springer Lecture Notes in Statistics* 1980, 2, 1-25.

6. Aalen, O.O. A linear regression model for the analysis of lifetimes. *Statistics in Medicine* 1989; **8**: 907-925.

7. Aalen, O.O., Borgan, Ø & Fekjaer, H. Covariate adjustment of event histories estimated from Markov chains: The adaptive approach. *Biometrics* 2001; **57**: 993-1001.

8. Abdi, H. Partial least squares regression (PLS-regression). In Lewis-Beck, M., Bryman, A., Futing, T. (ed.), *Encyclopedia for research methods for the social sciences* 2003, Thousand Oaks (CA): Sage.

9. Abecassis, I. et. al. Re-expression of DNA methylation-silenced CD44 gene in a resistant NB4 cell line: rescue of CD44-dependent cell death by cAMP. *Leukemia* 2008; **22(3)**: 511-520.

10. Anderson, P.K., Borgan, Ø, Gill, R.D. & Kieding, N. Statistical Methods Based on Counting Processes. *New York:Springer-Verlag.*

11. Anderson, P.K. & Kieding, N. Multistate models for event history analysis. *Statistical Methods in Medical Research* 2002; **11**: 91-115.

12. Anderson, P.K. & Klein, J.P. Regression analysis for multistate models based on a pseudo-value approach, with applications to bone-marrow transplantation studies. *Scandinavian Journal of Statistics* 2007; **34**: 3-16.

13. Ayala, R.M. et. al. Clinical significance of Gata-1, Gata-2, EKLF, and c-MPL expression in acute myeloid leukemia. *American Journal of Hematology* 2009; **84(2)**: 79-86.

14. Barlow, R.E., Bartholomew, J.M., Bremner, J.M. and Brunk, H.D. *Statistical Inference under Order Restrictions* 1972. New York: Wiley.

15. Beck, G.J. Stochastic survival models with competing risks and covariates. *Biometrics* 1979; **35**: 427-438.

16. Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* 1995; **57(1)**: 289-300.

17. Beran, R. Nonparametric regression with randomly censored survival data. *Technical Report, University of California, Berkeley* 1981.

18. Block, M. et. al. Preleukemic acute human leukemia. *JAMA* 1953; **152(11)**: 1018-1028.

19. Bourquin, J-P. et. al. Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proceedings of The National Academy of Sciences* 2006; **103(9)**: 3339-3344.

20. Buja A, Hastie T and Tibshirani R. Linear Smoothers and Additive Models. *The Annals of Statistic*s 1994; **17(2)**: 453-510.

21. Burton, J. D. et. al. CD74 is expressed by multiple myeloma and is a promising target for therapy. *Clinical Cancer Research* 2004; **10(19)**: 6606-6611.

22. Chakraborty, S., Datta, S. and Datta, S. Surrogate Variable Analysis Using Partial Least Squares (SVA-PLS) in Gene Expression Studies. *Bioinformatics* 2012; **28(6)**: 799-806.

23. Charrad, R-S. et. al. Ligation of the CD44 adhesion molecule reverses blockage of differentiation in human acute myeloid leukemia. *Nature Medicine* 1999; **5(6)**: 669-676.

24. Copelan EA, Biggs JC, Thompson JM, Crilley P, Szer J, Klein JP, Kapoor N, Avalos BR, Cunningham I, Atkinson K, Downs K, Harmon GS, Daly MB, Brodsky I, Bulova SI, Tutschka PJ. Treatment for acute myelocytlic leukemia with allogeneic bone marrow transplantation following preparation with Bu/Cy. *Blood* 1991; **78**: 838-843.

25. Cox, D.R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B* 1972; **34**: 187-220.

26. Crowley, J and Hu, M. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association* 1977; **72**: 27-36.

27. Cutler, C.S. et. al. A decision analysis of allogeneic bone marrow transplantation for the myelodysplastic syndromes: delayed transplantation for low-risk myelodysplasia is associated with improved outcome. *Blood* 2004; **104(2)**: 579-585.

28. Dabrowska, D.M. Nonparametric regression with censored survival time data. *Scandinavian Journal of Statistics* 1987; **14**: 181-197.

29. Dabrowska, D.M. Uniform consistency of the Kernel Conditional Kaplan-Meier Estimate. *The Annals of Statistics* 1989; **17**: 1157-1167.

30. Datta, S. and Satten, G.A. Non-parametric estimation for the three stage irreversible illness-death model. *Biometrics* 2000; **56**: 841-847.

31. Datta, S. Satten, G.A. & Datta, S. Estimation of stage occupation probabilities in multistate models. In: Balakrishnan, N eds. *Advances on Theoretical and Methodological Aspects of Probability and Statistics* 2000. New York: Gordon and Breach, 493-506.

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

33. Datta, S. & Satten, G.A. Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics* 2002; **58**: 792-802.

34. Datta, S. & Sundaram, R. Nonparametric estimation of state occupation probabilities in a multistate model with current status data. *Biometrics* 2006; **62**: 829–837.

35. Datta, S., Lan, L. & Sundaram, R. Non-parametric estimation of waiting time distributions in a Markov model based on current status data. *Journal of Statistical Planning and Inference* 2009; **139**: 2885-2897.

36. Doksum, K.A. & Yandell, B.S. Properties of regression estimates based on censored survival data. A *Festschrift for Erich L. Lehmann, ed. by PJ Bickel, KA Doksum, JL Hodges, Jr, Wadsworth: Belmont, CA*, 140–156.

37. Fan, J. and Datta, S. On Mann–Whitney tests for comparing sojourn time distributions when the transition times are right censored. *Annals of the Institute of Statistical Mathematics,* 2013; **65(1)**: 149-166.

38. Haga, K. The mechanism for reduced expression of gelsolin, tumor suppressor protein, in bladder cancer. *Hokkaido Igaku Zasshi.* 2003; **78(1)**: 29-37.

39. Haque, A. et. al. Increased sensitivity to TRAIL-induced apoptosis occurs during the adenoma to carcinoma transition of colorectal carcinogenesis. *British Journal of Cancer* 2005; **92(4)**: 736-742.

40. Harkema, S.J., Schmidt-Read, M., Behrman, A.L., Bratta, A., Sisto, S.A. and Edgerton, V.R. Establishing the NeuroRecovery Network: multisite rehabilitation centers that provide activity-based therapies and assessments for neurologic disorders. *Archives of Physical Medicine and Rehabilitation* 2011; **93**: 1498-1507.

41. Harkema, S. J., Schmidt-Read, M., Lorenz, D., Edgerton, V.R. and Behrman, A.L. Balance and ambulation improvements in individuals with chronic incomplete spinal cord injury using locomotor training based rehabilitation. *Archives of Physical Medicine and Rehabilitation* 2011; **93**: 1508-1517.

42. Hastie T and Tibshirani R. Generalized Additive Models. *Monographs on Statistics and Applied Probability* 43, *Chapman and Hall,* 1990.

43. Helland, I.S. Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 1999; **58**: 97-107.

44. Hirotsugu, A. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19(6)**: 716-723.

45. Hirotsugu, A. Likelihood and the Bayes procedure. *Bayesian Statistics* 1980, 143-146.

46. Ito, S. et. al. Expression of MAGE-D4, a novel MAGE family antigen, is correlated with tumor-cell proliferation of non-small cell lung cancer. *Lung Cancer* 2006; **51(1)**: 79-88.

47. Kang, H.M. et. al. Efficient control of population structure in model organism association mapping. *Genetics* 2008; **178(3)**: 1709-1723.

48. Kang, H.M. et. al. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 2008; **180(4)**: 1909-1925.

49. Kay, R. The analysis of transition times in multistate stochastic processes using proportional hazard regression models. *Communications in Statistics-Theory and Methods* 1982; **11**: 1743-1756.

50. Kerr, M.K. and Churchill, G.A. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences* 2001; **98(16)**: 8961-8965.

51. Kerr, M.K. et. al. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 2000; **7**: 819-837.

52. Kinzler, K.W. and Vogelstein, B. Gatekeepers and caretakers. *Nature* 1997; **386**: 761-763.

53. Klein, J.P., Klotz, J.H. & Grever, M.R. A biological marker model for predicting disease transitions. *Biometrics* 1984; **40**: 927-936.

54. Klein, J.P. & Moeschberger, M.L. Survival Analysis: techniques for censored and truncated data. *New York: Springer-Verlag* 1997.

55. Lagakos, S.W. A stochastic model for censored-survival data in the presence of an auilliary variable. *Biometrics* 1976; **52**: 551-559.

56. Lan, L. & Datta, S. Nonparametric estimation of state occupation, entry and exit times with multistate current status data. *Statistical Methods in Medical Research* 2010; **19**: 147-165.

57. Lancaster, J.M. et. al. BRCA2 mutations in primary breast and ovarian cancers. *Nature Genetics* 1996; **13**: 238-240.

58. Leek, J.T. and Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* 2007; **3(9)**: e161.

59. Lehmann, E.L. Testing Statistical Hypotheses, 2nd edn. *Wiley, New York*. 1986.

60. Li, G. & Doss, H. An Approach to Nonparametric Regression for Life History Data Using Local Linear Fitting. *Annual Statistics* 1995; **23**: 787-823.

61. Li, G. & Datta, S. A bootstrap approach to nonparametric regression for right censored data. *Annals of Institute of Statistical Mathematics* 2001; **53**: 708-729.

62. Lin, D.Y. & Ying, Z. Semiparametric analysis of the additive risk model. *Biometrika* 1994; **81**: 61-71.

63. Lin, D.Y., Sun, W. & Ying, Z. Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* 1999; **45**: 497-507.

64. Listgarten, J. et. al. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences* 2010; **107(38)**: 16465-16470.

65. Mann, H.B. and Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947; **18**: 50-60.

66. McKeague, I.W. & Utikal, K.J. Inference for a Nonlinear Counting Process Regression Model. *Annual Statistics* 1990; **18**: 1172-1187.

67. Mostajabi, F. and Datta, S. Nonparametric regression of state occupation, entry. exit and waiting times with multistate right censored data. *Statistics in Medicine* 2012.

68. Murphy, K.M. et. al. Evaluation of candidate genes MAP2K4, MADH4, ACVR1B, and BRCA2 in familial pancreatic cancer: deleterious BRCA2 mutations in 17. *Cancer Research* 2002; **62(13)**: 3789-3793.

69. Nadaraya, E. On estimating regression. *Theory of Probability and Its Applications* 1964; **9**: 141-142.

70. Narod, S.A. et. al. Rapid progression of prostate cancer in men with a BRCA2 mutation. *British Journal of Cancer* 2008; **99(2)**: 371-374.

71. Özcelik, H. et. al. Germline BRCA2 6174delT mutations in Ashkenazi Jewish pancreatic cancer patients. *Nature Genetics* 1997; **16**: 17-18.

72. Peterson, L.F. et. al. The multi-functional cellular adhesion molecule CD44 is regulated by the 8;21 chromosomal translocation. *Leukemia* 2007; **21(9)**: 2010-2019.

73. Plevritis, S.K., Salzman, P., Sigal, B.M. & Glynn, P.W. A natural history model of stage progression applied to breast cancer. *Statistics in Medicine* 2007; **26**: 581-595.

74. Price, et. al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 2006; **38**: 904-909.

75. Riccioni, R. et. al. TRAIL decoy receptors mediate resistance of acute myeloid leukemia cells to TRAIL. *Haematologica* 2005; **90(5)**: 621-624.

76. Rosipal, R. and Krämer, N. Overview and Recent Advances in Partial Least Squares. In *Subspace, Latent Structure and Feature Selection Techniques, Lecture Notes in Computer Science* 2006, Springer, pp. 34-51.

77. Rubio-Moscardo, F. et. al. Characterization of 8p21.3 chromosomal deletions in B-cell lymphoma: TRAIL-R1 and TRAIL-R2 as candidate dosage-dependent tumor suppressor genes. *Blood* 2005; **106(9)**: 3214-3222.

78. Sacks, S.T & Chiang, C.L. A transition probability model for the study of chronic diseases. *Mathematical Biosciences* 1977; **60**: 197-206.

79. Sagawa, N. et. al. Gelsolin suppresses tumorigenicity through inhibiting PKC activation in a human lung cancer cell line, PC10. *British Journal of Cancer* 2003; **88(4)**: 606-612.

80. Satten, G.A. & Datta, S. Marginal estimation for multi-stage models: waiting time distributions and competing risks analyses. *Statistics in Medicine* 2002; **21**: 3-19.

81. Scheid, S. and Spang, R. Compensating for unknown confounders in microarray data analysis using filtered permutations. *Journal of Computational Biology* 2007; **14(5)**: 669-681.

82. Shimada, H. et. al. Potential involvement of the AML1-MTG8 fusion protein in the granulocytic maturation characteristic of the t(8;21) acute myelogenous leukemia revealed by microarray analysis. *Leukemia* 2002; **16(5)**: 874-885.

83. Smith, G. K. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, 2005.

84. Stegle, O. et. al. Accounting for non-genetic factors improves the power of eQTL studies. *Research in Computational Molecular Biology* 2008, 411-422.

85. Stegle, O. et. al. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLoS Computational Biology* 2010; **6(5)**: e1000770.

86. Shu, Y. & Klein, J.P. Additive hazards Markov regression models illustrated with bone marrow transplant data. *Biometrika* 2005; **92**: 283-301.

87. Tataroglu, C. et. al. Beta-catenin and CD44 expression in keratoacanthoma and squamous cell carcinoma of the skin. *Tumori* 2007; **93(3)**: 284-289.

88. Tusher, V., Tibshirani, R. and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 2001; **98**: 5116-5121.

89. Wagner, J.E. et. al. Germline mutations in BRCA2: shared genetic susceptibility to breast cancer, early onset leukemia, and Fanconi anemia. *Blood* 2004; **103(8)**: 3226-3229.

90. Wand, M.P. & Jones, M.C. Kernel Smoothing. *London: Chapman & Hall*, 1995.

91. Welch, B.L. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1938; **29(3/4)**: 350-362.

92. Wold, H. Path models with latent variables: The NIPALS approach. *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building* 1975, 307-357.

93. Wold, H. Partial least squares. In Kotz, S. and Johnson, N.L. (ed.), {\it *Encyclopedia of the statistical sciences*, *Wiley, New York* 1985; **6**: 581-591.

94. Wolfinger, R.D. et. al. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* 2001; **8(6)**: 625-637.

95. Wu, S. et. al. Levels of the soluble, 55-kilodalton isoform of tumor necrosis factor receptor in bone marrow are correlated with the clinical outcome of children with acute lymphoblastic leukemia in first recurrence. *Cancer* 2003; 98(3): 625-631.

96. Wu, S.C. A semi-Markov model for survival data with covariates. *Mathematical Biosciences* 1982; **60**: 197-206.

97. Yu, J.M. et. al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 2006; **38(2)**: 203-208.

Appendix: Proof of Theorem 1

For the case of two covariates $(p = 2)$ the two estimating equations for the i-th iterated estimates of f_1^t and f_2^t are given by:

$$f_1^{t,i} = M_1(\boldsymbol{H}(t) - f_2^{t,(i-1)}) - \widehat{\alpha}(t)$$

$$f_2^{t,i}(t) = M_2(\boldsymbol{H}(t) - f_1^{t,i}(t)) - \widehat{\alpha}(t),$$

where $\boldsymbol{H}(t) = (H_1(t), H_n(t) \cdots H_n(t))^T$ is a vector of n observations on the process H at time t. Now, after running the algorithm upto l iterations we get the estimators for the two time dependent covariate specific functions as:

$$f_1^{t,l} = [\boldsymbol{H}(t) - \widehat{\boldsymbol{\alpha}}(t)] - \sum_{j=0}^{l-1}(M_1M_2)^j(\boldsymbol{I} - M_1)\boldsymbol{H}(t) +$$

$$B_l\widehat{\boldsymbol{\alpha}}(t) - (M_1M_2)^{l-1}M_1f_2^{t,0},$$

$$f_2^{t,l} = M_2\sum_{j=0}^{l-1}(M_1M_2)^j(\boldsymbol{I} - M_1)\boldsymbol{H}(t) - \sum_{u=0}^{l-1}(M_2M_1)^u(I - M_2)\widehat{\boldsymbol{\alpha}}(t) +$$

$$M_2(M_1M_2)^{l-1}M_1f_2^{t,0}$$

where

$$B_l = \begin{cases} \mathbf{0}_{n\times n} & \text{if } l = 1 \\ M_1(I - M_2) & \text{if } l = 2 \\ M_1\sum_{u=1}^{l-1}(M_2M_1)^{u-1}(I - M_2) & \text{if } l > 2 \end{cases}$$

$\boldsymbol{I} = diag((1, 1, \cdots 1))$ is the identity matrix of order n and $\widehat{\boldsymbol{\alpha}}(t) = (\widehat{\alpha}(t), \widehat{\alpha}(t) \cdots \widehat{\alpha}(t))^T$.

Now, generalizing the two iterated estimates through $l = 1, 2, \cdots \infty$ we get two infinite geometric series for each function, which in the case of $||M_1 M_2|| < 1$ and $||M_2 M_1|| < 1$ converge to the following two solutions:

$$f_1^{t,\infty} = [\boldsymbol{I} - (\boldsymbol{I} - M_1 M_2)^{-1}(1 - M_1)]\boldsymbol{H}(t) - [\boldsymbol{I} - M_1(\boldsymbol{I} - M_2 M_1)^{-1}(I - M_2)]\widehat{\boldsymbol{\alpha}}(t),$$

$$f_2^{t,\infty} = M_2(\boldsymbol{I} - M_1 M_2)^{-1}(\boldsymbol{I} - M_1)\boldsymbol{H}(t) - (\boldsymbol{I} - M_2 M_1)^{-1}(\boldsymbol{I} - M_2)\widehat{\boldsymbol{\alpha}}(t).$$

Thus the estimator of the mean of $\boldsymbol{H}(t)$ is given by :

$$\widehat{E}(\boldsymbol{H}(t)) = \widehat{\boldsymbol{\alpha}}(t) + f_1^{t,\infty} + f_2^{t,\infty}$$

$$= (\boldsymbol{I} - G_1)Y - G_2\widehat{\boldsymbol{\alpha}}(t)$$

where

$$G_1 = (\boldsymbol{I} - M_2)(\boldsymbol{I} - M_1 M_2)^{-1}(\boldsymbol{I} - M_1)$$

$$G_2 = (\boldsymbol{I} - M_1)(\boldsymbol{I} - M_2 M_1)^{-1}(\boldsymbol{I} - M_2)$$

which are both symmetric in M_1 and M_2.

Hence, we find that if $||M_1 M_2|| < 1$ and $||M_2 M_1|| < 1$, the conditionally estimated mean of $\boldsymbol{H}(t)$ is uniquely defined and independent of the starting values for the functions f_1^t and f_2^t.

CURRICULUM VITAE

Sutirtha Chakraborty

CONTACT INFORMATION
786 Raymond Kent Court, Apt-3
Louisville, KY-40217
Mobile: 502 415-5676
Gmail: statistuta

RESEARCH INTERESTS
Statistical genetics, Survival Analysis, Bayesian Statistics

EDUCATION AND TRAINING

University of Louisville, Louisville, Kentucky, USA
Ph.D. student (2009 - present)

- Ph.D. dissertation: Novel Methods Based on Regression Techniques to Analyze Multistate Models and High dimensional Omics Data
- Advisors: Dr. Somnath Datta (Professor, ASA fellow, IMS fellow) and Dr. Susmita Datta (Professor, ASA fellow).

Indian Statistical Institute, India
MSTAT, Applied Statistics and Data Analysis (2009)

- Dissertation: Problems involved in Multiple Hypotheses Testing and their feasible remedies.
- Advisor: Dr. Debapriya Sengupta.

Presidency University, India
B.Sc. in Statistics honors (2007)

- Graduated with First Class
- Minors in Physics and Mathematics.

AWARDS

Silver medal and cash prizes for scoring the highest marks in B.Sc Part I and II examinations in Statistics Honors, among the students in the university.

University Fellow in the University of Louisville from 2007 - 2009.

Received the University of Louisville Dissertation completion award in Fall 2012.

PROFESSIONAL EXPERIENCE

University of Louisville

Statistical Consultant

- Served as a statistical analyst in several collaborative projects with biomedical researchers, providing meaningful validations of designed methodologies in different contexts like brain cortex clustering, computational fluid dynamics (CFD) modeling, gene-expression studies and peptide fragmentation pattern detection in Proteomics.

Teaching Assistant

- Advanced Statistical Inference (PHST 762)
- Advanced Survival Analysis (PHST 783)

Referee Service

- BMC Bioinformatics.

PUBLICATIONS

Chakraborty, S., Datta, S. and Datta, S. (2012) Surrogate Variable Analysis using Partial Least Squares (SVA-PLS) in Gene Expression Studies. *Bioinformatics*, 28(6): 799-806.

Chakraborty, A., **Chakraborty, S.**, Jala, V. R., Haribabu, B., Sharp, M.K., Berson, R. E. (2012) Effects of Biaxial Oscillatory Shear Stress on Endothelial Cell Proliferation and Morphology. *Biotechnology and Bioengineering*, 109(3): 695-707.

Datta, S., Datta, S., Kim, S., **Chakraborty, S.**, Gill, R. S. (2010) Statistical Analyses of Next-Generation Sequence Data: a partial overview. *Journal of Proteomics and Bioinformatics*, 3(6): 183-190.

UNDER PREPARATION

Chakraborty, S., Datta, S., Datta, S. (2013) *svapls*: An R package to correct for hidden factors of variability in gene expression studies. Submitted to *BMC Bioinformatics.*

Chakraborty, S., Datta, S. and Datta, S. (2013) Nonparameteric Regression of Temporal Functions in a Multistate Model Under Right Censoring via Additive Models for Counting and Number at Risk Processes.

Chakraborty, S. and Datta, S. (2013) Testing the equality of the waiting time distributions between two groups of individuals using an IPCW-based Mann-Whitney U-test after adjusting for available covariates.

Chakraborty, A., Jala, V. R., **Chakraborty, S.**, Sharp, M. K., Berson, R. E., Haribabu, B. (2012) Biaxial Oscillatory Fluidic Shear Stress increases Pro- Atherogenic Gene Expression in Endothelial Cells. Submitted to *PLoS One.*

SOFTWARE SKILLS

R, MATLAB, SAS, C.