DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1180

Topics on distance correlation, feature screening and lifetime expectancy with application to Beaver Dam Eye Study data

June 21, 2015

Jing Kong[1]
Department of Statistics
University of Wisconsin, Madison

[1]Jing Kong's research has been supported by NSF Grant DMS1308877 and NIH Grant EY09946.

TOPICS ON DISTANCE CORRELATION, FEATURE SCREENING AND LIFETIME EXPECTANCY WITH APPLICATION TO BEAVER DAM EYE STUDY DATA

by

Jing Kong

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: 05/28/2015

The dissertation is approved by the following members of the Final Oral Committee:
Grace Wahba, Advisor, IJ Schoenberg-Hilldale Professor, Statistics
Sijian Wang, Associate Professor, Biostatistics & Medical Informatics and Statistics
Menggang Yu, Associate Professor, Biostatistics & Medical Informatics
Karl Rohe, Assistant Professor, Statistics
Garvesh Raskutti, Assistant Professor, Statistics

ACKNOWLEDGMENTS

The completion of my dissertation and subsequent Ph.D. puts an end to a memorable journey. I can not suppress the mixed emotions in heart as I recall every moment of the past five years. Embracing a young mind with ambition and aspiration, I landed in a brand new world of the United States wishing all the best coming to life until the unexpected heavy stress from study and the difficulties encountered in a foreign life overwhelmed me. I am glad that the life challenges and changes followed that never beat me or swayed my determination to the accomplishment of my dreams, but equipped me with a more tolerant, patient and better prepared heart for the future.

I could not have succeeded without the invaluable support of a several. Without these supporters, especially the select few I am about to mention, I may not have reached where I am today.

I would like to give a special and heartfelt thanks to my advisor Professor Grace Wahba. One could never expect a more knowledgable, patient, flexible and genuine caring PhD advisor than Grace is. Grace has been motivating, supportive and enlightening. She has been constantly happy to discuss whenever I get stuck with my research, to instruct with insight and direction-right that help me solve problems, and to provide me with beneficial opportunities. She always brightens up my day with her stories about dancing clubs, biking in Europe, country skiing in the cold northern winter as well as her adventure and championship in the Wisconsin Senior Olympics. With Grace, I've learned so much more than intelligence, dedication, consideration, encouragement and passion. It's about advising me to live life to the fullest by showing me how it is done. For this, I cannot thank her enough. I am forever grateful. Thank You Grace!

I would like to thank Professor Sijian Wang who is one of the smartest and energetic professors I met during the PhD study. He introduced the area of survival analysis to me, guided me with his expertise, motivated my thinking with perspective questions, and shared his useful suggestions with me. Sijian provided me the TCGA data used in the second chapter of the thesis and contributed meaningful discussions about the results. Moreover, he also advised me to further explore the relationship between backward imputation and Buckley-James estimator in the fourth chapter. I want to thank Professor Sijian Wang for his kind efforts leading to the completion of this thesis.

I am very grateful to Professor Menggang Yu, Professor Karl Rohe and Professor Garvesh Raskutti for being my thesis committee members. Their academic support and questions towards my work are the drivers for my further thoughts and are greatly appreciated. I would also like to thank them for inputting interesting research topics during the Thursday

group meetings to broaden my knowledge in different areas of statistics.

I would also like to thank the professors who ever contributed in my study. I want to thank Professor Ming Yuan for challenging me with sharp questions related to my research and for encouraging me to know more academic celebrities. I want to thank Professor Yaning Yang from my undergraduate study who was the best professor teaching ideas rather than formulas. Without him, I would never be so interested in statistics and be so sure to pursue a PhD degree in statistics. Thanks all the students from Thursday group meetings, including Bin Dai, Shilin Ding, Zhigeng Geng, Tai Qin, Luwan Zhang, Han Chen, Shulei Wang, Cuize Han, Hao Zhou, Yilin Zhang, Xiaowu Dai, for presenting your work and introducing new ideas to me. Research becomes interesting and enjoyable with you together!

The last but not the least, thank my parents for their unconditional love, understanding and support. Thank you for comforting me when I was too stressful to sleep and walking me through every obstacle ever happened to life. Thank you for being such bright parents, looking into the future of the child, and suggesting the best way out of your wisdom and experiences. Thank Yun Zhai for listening to all my complaints, cheering me up, patiently enduring my long hours busy with my own study, serving as a math expert, and delighting my PhD study.

True courage is like a kite; a contrary wind raises it higher.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis includes 4 pieces of work, focusing on topics including distance correlation, smoothing spline ANOVA models and lifetime expectancy function with applications to Beaver Dam eye study data and the Cancer Genome Atlas ovarian cancer data. The following displays a general picture for each part of the dissertation with a brief introduction.

In Chapter 1, we present the work in Kong et al. (2012) with a method for examining mortality as it is seen to run in families, and lifestyle factors that are also seen to run in families, in a subpopulation of the Beaver Dam Eye Study that has died by 2011. We observe that pairwise distance between death age in related persons is on average less than pairwise distance in death age between random pairs of unrelated persons. Our goal is to examine the hypothesis that pairwise differences in lifestyle factors correlate with the observed pairwise differences in death age that run in families. Szekely and coworkers, Székely et al. (2007), have recently developed a method called distance correlation, that is suitable for this task with some enhancements relevant to the particular task at hand. We build a Smoothing Spline ANOVA (SS-ANOVA) model for predicting death age based on four major lifestyle factors generally known to be related to mortality and four of the major diseases contributing to mortality, to develop a lifestyle mortality risk vector and a disease mortality risk vector. We then examine to what extent pairwise differences in these scores correlate with the pairwise differences in mortality as they occur between family members and between unrelated persons. We find significant distance correlations between death ages, lifestyle factors, and family relationships. Considering only sib pairs compared to unrelated persons, distance correlation between siblings and mortality is, not surprisingly, stronger than that between more distantly related family members and mortality. The overall methodological approach here easily adapts to exploring relationships between multiple clusters of variables with observable (real-valued) attributes, and other factors for which only possibly nonmetric pairwise dissimilarities are observed.

Chapter 2 introduces a feature screening procedure in Kong et al. (2015) with the use of distance correlation and covariance in Székely et al. (2007). With Pearson's correlation, Fan and Lv proposed the sure independence screening (SIS) in Fan and Lv (2008) and showed that the Pearson correlation ranking procedure possessed a sure screening property for linear regression with Gaussian predictors and responses. Distance correlation generalizes Pearson's correlation in that it captures multivariate and nonlinear dependence and hence can be used for feature screening with general relationship between the response and predictors and is robust to model mis-specification. A new feature screening procedure for high dimensional data based on distance correlation, named DC-SIS, was presented

in Li et al. (2012). However, both SIS and DC-SIS rely on a user-specified model size d which decides the number of predictors being selected and may influence the screening results. To address this problem, we demonstrate a property for distance covariance, which is incorporated in a novel feature screening procedure based on distance correlation as a stopping criterion. The approach is further implemented to two real examples. The first one is the famous small round blue cell tumors (SRBCT) data, which have been extensively studied and are relatively easy to deal with due to the significant distinguish among the 4 types of tumor. The second is the Cancer Genome Atlas (TCGA) ovarian cancer data, which are much more challenging due to the large number of genes and limited sample size. We illustrate the selected genes out of our procedure though prediction power using support vector machine with reject option (SVM-R), Wegkamp et al. (2011), to adapt a subgroup of hard-to-classify patients.

Chapter 3 pays attention to the right censored human longevity data and the estimation of lifetime expectancy. The conditional lifetime expectancy function (LEF) is the expected lifetime of a subject given survival past a certain time point and the values of a set of explanatory variables. This function is attractive to researchers since it summarizes the entire residual life distribution and has an easy interpretation compared to the popular used hazard function. In this chapter, we propose a general framework of backward multiple imputation for estimating the conditional LEF and the variance of the estimator in the right censoring setting. We prove that the proposed method is equivalent to estimating the LEF with Kaplan-Meier estimator for the survival function in the case without any covariate. Moreover, when covarites information is available, using our backward imputation method with kernel regression gives the nonparametric estimation for the conditional LEF proposed in McLain and Ghosh (2011). Simulation studies are conducted to investigate the empirical properties of the proposed estimator and the corresponding variance estimator. With the recently updated survival information up to December 31, 2013, the Beaver Dam eye study data again provide us an excellent opportunity to study human longevity, where the expected human lifetime are modeled with smoothing spline ANOVA based on the covariates including baseline age, gender, lifestyle factors and disease variables. The effects of different risky components towards human lifetimes are explicitly illustrated with plots and are observed to vary with the time that one individual survives past.

Chapter 4 compares two imputation methods for right censored data, namely the famous Buckley-James estimator and the backward imputation method proposed in Chapter 3. Buckley-James estimator introduced in Buckley and James (1979) is a popular alternative to Cox's proportional hazard model as the usual least square regression adapted to censored data. Simulated data demonstrate that the original Buckley-James method fails when

its two assumptions, namely linearity and homoscedasticity, are moderately violated. To make things worse, checking these assumptions under censoring is difficult, if at all possible. Backward imputation, however, are shown to be less biased and more robust with nonlinear and heterogeneous data in the conducted simulation studies, especially under small sample size and high censoring rate. To further compare the two methods in real examples, we evaluate the performances on two well-known survival data, the Stanford heart transplantation data and the veteran's administration lung cancer data. It turns out that backward imputation with SS-ANOVA model outperforms Buckley-James with linear model or SS-ANOVA with respect to the mean squared error of predicted censored times and the bias in linear coefficients estimated from the imputed data.

1 USING DISTANCE CORRELATION AND SS-ANOVA TO ASSESS ASSOCIATIONS OF FAMILIAL RELATIONSHIPS, LIFESTYLE FACTORS, DISEASES AND MORTALITY

1.1 Introduction

Multiple studies have reported that collectively lifestyle factors, including smoking, low or high body mass index (bmi), low educational attainment and low socio-economic status, are associated with earlier mortality. Diseases, such as diabetes, cardiovascular disease, cancer and chronic kidney diseases, are leading causes of death. Longevity is generally believed to run in families. Furthermore, there is evidence showing that the lifestyle factors all tend to run in families. The goal of this paper is to capture the association of familial relationships, lifestyle factors, diseases and mortality. It is possible that some of the lifestyle variables may be or turn out to be related to genetic factors. Current research interest involves searches for “longevity genes" but this work is not related to that quest. We are not assessing to what extent genetics is involved in longevity.

The Beaver Dam Eye Study (BDES), Klein et al. (1991), is an ongoing population-based study of age-related ocular disorders. Subjects at baseline, examined between 1988 and 1990, were a group of 4926 people aged 43-86 years who lived in Beaver Dam, WI. Many group members have relatives in the study, and pedigree information was collected. Mortality information was updated to March 2011. BDES provides an excellent opportunity to attempt to examine and quantify the above associations.

A pair of landmark papers ,Székely et al. (2007, 2009) proposed the distance correlation as a measurement of multivariate independence, and others have recently built upon it, see Tran et al. (2012); Li et al. (2012); Khoshgnauz (2012); Lyons et al. (2013). The method is extremely general in that it is applicable to random vectors of arbitrary and not necessarily equal dimension and only involves Euclidean pairwise distance. If the two variables are sampled from a bivariate normal distribution, the distance correlation behaves very much like the Pearson’s correlation coefficient. Since only Euclidean pairwise distances enter, the method may be applied to inherently unobservable variables with only Euclidean pairwise distances observable. The “genetic distances" defined on pairs of persons representing their familial relationships are generally not Euclidean. However, it is shown that the use of genetic dissimilarity in the distance correlation is still validated since the genetic dissimilarity can be well approximated by Euclidean pairwise distances obtained by embedding the subjects into Euclidean spaces through Regularized Kernel

Estimation (RKE), see Lu et al. (2005); Bravo et al. (2009).

Smoothing Spline ANOVA (SS-ANOVA) models have a successful history for modeling various aspects of BDES data, two examples are Wahba et al. (1995); Gao et al. (2001). In this study, we focus on modeling the mortality (death ages) of the form

$$\begin{aligned} \textit{death age}_i = & g_0(\textit{baseline age}_i, \textit{gender}_i) + \\ & g_1(\textit{lifestyle factor}_i) + g_2(\textit{disease}_i), \end{aligned}$$

where g_0 is a term involves fixed characteristics, baseline age and gender, for the individuals, g_1 is a term that includes only lifestyle factors and g_2 is a term containing only disease variables, namely diabetes, cancer, cardiovascular disease and chronic kidney disease. In the paper, the fitted values of g_1 and g_2 are treated as scores for the individuals and to be used to assess the association with familial relationships.

1.2 Pedigrees

The genetic relationships between pedigree members can be described by Malecot's Malécot et al. (1948) kinship coefficient φ which defines a pedigree dissimilarity measure. The kinship coefficient φ between individuals i and j in the pedigree is defined as the probability that a randomly selected pair of alleles, one from each individual, is identical by descent, that is, they are derived from a common ancestor. For a parent-offspring pair, $\varphi_{ij} = 0.25$ since there is a 50% chance that the allele inherited from the parent is chosen at random for the offspring, and a 50% chance that the same allele is chosen at random for the parent.

Pedigree Dissimilarity

The pedigree dissimilarity between individuals i and j is defined for this study as $d_{ij} = 1 - 2\varphi_{ij}$, where φ is the kinship coefficient. Thus, for $i \neq j$, the pedigree dissimilarity here falls in the interval $[\frac{1}{2}, 1]$. Note that Bravo et al. (2009) define pedigree dissimilarity for that study as $-log_2(2\varphi)$, which ranges from 1 to ∞ for $i \neq j$, which is not appropriate for the way we will be using pedigree dissimilarity.

In BDES, not all family members are included in the study and not all the subjects have pedigree records.

1.3 Smoothing-Spline ANOVA Models

SS-ANOVA models, Wahba (1990); Gu (2013); Wang (2011), estimate the responses $y_i, i = 1, ..., n$ to be a function of the covariates $f(x_i)$, by assuming that f is a function in a reproducing kernel Hilbert space (RKHS) of the form $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$. $\mathcal{H}_0$ is a finite dimensional space spanned by a set of functions $\{\phi_1, ..., \phi_m\}$, and $\mathcal{H}_1$ is an RKHS induced by a given kernel function $k(\cdot, \cdot)$ with the property that $\langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{H}_1} = k(x_i, x_j)$. Thus, the function f has a semiparametric form of

$$f(x) = \sum_{j=1}^{m} d_j \phi_j(x) + g(x),$$

for some coefficients d_j, where the functions ϕ_j's are of parametric linear form and $g \in \mathcal{H}_1$. $\mathcal{H}_1$ is further decomposed by assuming that it is the direct sum of multiple RKHSs. Hence, $g \in \mathcal{H}_1$ is defined to be

$$g(x) = \sum_{\alpha} g_\alpha(x_\alpha) + \sum_{\alpha<\beta} g_{\alpha\beta}(x_\alpha, x_\beta) + \cdots,$$

where $\{g_\alpha\}$ and $\{g_{\alpha\beta}\}$ satisfy side conditions that generalize the standard ANOVA side conditions. Functions g_α are the "main effects" and $g_{\alpha\beta}$ are the "second-order interactions", and so on. The RKHS $\mathcal{H}_\alpha$ is associated with each component in the above sum, along with its corresponding kernel function k_α. In this case, the reproducing kernel function for $\mathcal{H}_1$ is defined to be

$$k(\cdot, \cdot) = \sum_{\alpha} \theta_\alpha k_\alpha(\cdot, \cdot) + \sum_{\alpha<\beta} \theta_{\alpha\beta} k_{\alpha\beta}(\cdot, \cdot) + \cdots,$$

where the coefficients θ's are tuning parameters that weigh the relative importance of each term in the decomposition.

The SS-ANOVA estimates f given data $\{(x_i, y_i), i = 1, ..., n\}$ by the solution of a penalized likelihood problem of the form

$$\min_{f\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) + J_{\lambda,\theta}(f), \tag{1.1}$$

where $l(y_i, f(x_i)) = (y_i - f(x_i))^2$ and

$$J_{\lambda,\theta}(f) = \lambda[\sum_{\alpha} \theta_\alpha^{-1} \|P_\alpha f\|^2_{\mathcal{H}_\alpha} + \sum_{\alpha<\beta} \theta_{\alpha\beta}^{-1} \|P_{\alpha\beta} f\|^2_{\mathcal{H}_{\alpha\beta}} + \cdots],$$

with $P_\alpha f$ the projection of f into RKHS $\mathcal{H}_\alpha$ and λ a non-negative regularization parameter. The penalty $J_{\lambda,\theta}(f)$ is a seminorm in RKHS $\mathcal{H}$ and penalizes the complexity of f using the norm of RKHS $\mathcal{H}_1$ to avoid overfitting f to the training data.

According to Kimeldorf and Wahba (1971), the minimizer of the problem in equation [1] has a finite representation taking the form of

$$f(\cdot) = \sum_{j=1}^{m} d_j \phi_j(\cdot) + \sum_{i=1}^{n} c_i k(x_i, \cdot),$$

where $\|P_1 f\|^2_{\mathcal{H}_1} = c^T K c$ for kernel matrix K with $K_{ij} = k(x_i, x_j)$. Therefore, for a given value of the regularization parameter λ, the minimizer f_λ can be estimated by solving the following convex optimization problem:

$$\min_{c \in R^n, d \in R^m} \sum_{i=1}^{n} (y_i - f(x_i))^2 + n\lambda c^T K c, \tag{1.2}$$

where $f = [f(x_1), ..., f(x_n)]^T = Td + Kc$ with $T_{ij} = \phi_j(x_i)$. The hyperparameters, λ and θ's, are to be chosen by the generalized cross validation (GCV) as described in Golub et al. (1979); Craven and Wahba (1977).

1.4 Distance Correlation

For a random sample $(X, Y) = \{(X_k, Y_k) : k = 1, ..., n\}$ of n i.i.d random vectors (X, Y) from the joint distribution of random vectors X in R^p and Y in R^q, the Euclidean distance matrices $(a_{ij}) = (|X_i - X_j|_p)$ and $(b_{ij}) = (|Y_i - Y_j|_q)$ are computed. Define the double centering distance matrices

$$A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, \quad i, j = 1, ..., n,$$

where

$$\bar{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \bar{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij},$$

similarly for $B_{ij} = b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}, \quad i, j = 1, ..., n.$

Sample Distance Covariance

The sample distance covariance $\mathcal{V}_n(X, Y)$ is defined by

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}.$$

Sample Distance Correlation

The sample distance correlation $\mathcal{R}_n(X, Y)$ is defined by

$$\mathcal{R}_n^2(X, Y) = \begin{cases} \dfrac{\mathcal{V}_n^2(X, Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0; \\ 0, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0, \end{cases}$$

where the sample distance variance is defined by

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X, X) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^2.$$

The nonnegativity of $\mathcal{V}_n^2$ and $\mathcal{R}_n^2$ is guaranteed, see Székely et al. (2009). The theory in Székely et al. (2009) is based on dissimilarities being actual distances between objects embedded in a Euclidean space, although it is mentioned in the rejoinder to the discussion there that the results hold in certain other metric spaces, see also Lyons et al. (2013). The pedigree dissimilarity (d_{ij}) cannot be considered as coming from some metric space, however, since, at least in our study, it does not satisfy the triangle inequality. But we could still treat the pedigree dissimilarity as though it were a distance, since we will see that it can be well approximated by a Euclidean distance obtained by RKE, which we discuss in the next section.

1.5 Regularized Kernel Estimation

The Regularized Kernel Estimation (RKE) framework was introduced in Lu et al. (2005) as a robust method for estimating dissimilarity measures between objects from noisy, incomplete, inconsistent, and repetitious dissimilarity data. RKE is useful in settings where object classification or clustering is desired but objects do not easily admit description by fixed-length feature vectors, but instead, there is access to a source of noisy and incomplete dissimilarity information between objects. It estimates a symmetric positive semidefinite

kernel matrix K which induces a real squared distance admitting of an inner product $d_{ij}^2 = K_{ii} + K_{jj} - 2K_{ij}$.

Assume dissimilarity information is given for a subset Ω of the $\binom{n}{2}$ possible pairs occurring in a training set of n objects, with the dissimilarity between objects i and j denoted as $d_{ij} \in \Omega$. RKE estimates an $n \times n$ symmetric positive semidefinite kernel matrix K of size n such that the fitted squared distance between objects induced by K, $\hat{d}_{ij}^2 = K_{ii} + K_{jj} - 2K_{ij}$, is as close as possible to the square of the observed dissimilarities $d_{ij} \in \Omega$. RKE solves the following optimization problem with semidefinite constraints:

$$\min_{K \succeq 0} \sum_{d_{ij} \in \Omega} w_{ij} |d_{ij}^2 - \hat{d}_{ij}^2| + \lambda_{rke} trace(K). \tag{1.3}$$

The parameter $\lambda_{rke} \geqslant 0$ is a regularization parameter that trades off fit of the dissimilarity data, as given by absolute deviation, and a penalty, trace(K), on the complexity of K. The trace may be seen as a proxy for the rank of K. Thus, RKE is regularized by penalizing high dimensionality of the space spanned by K. RKE requires that Ω satisfies a connectivity constraint that the undirected graph consisting of objects as nodes and edges between them, such that an edge between nodes i and j is included if $d_{ij} \in \Omega$, is connected. Additionally, optional weights w_{ij} may be associated with each $d_{ij} \in \Omega$. A method for choosing the regularization parameter λ_{rke} is required. In this work λ_{rke} is fixed at 1. Unlike in many regularization models, results in the RKE tend to be remarkably insensitive to λ_{rke} over a wide range of values, as can be seen in Figure 1.1 of Lu et al. (2005).

The solution to the RKE problem is a symmetric positive semidefinite matrix K from which an embedding $Z \in R^{n \times r}$ in $r-$dimensional Euclidean space is obtained by decomposing K as $K = ZZ^T$ with $Z = \Gamma_r \Lambda_r^{\frac{1}{2}}$, where the $n \times r$ matrix Γ_r and the $r \times r$ diagonal matrix Λ_r contains the r leading eigenvalues and eigenvectors of K respectively. The ith row of Z is regarded as the vector of "pseudo" coordinates $z(i)$ for subject i. A method for choosing r is required.

The fact that RKE operates on inconsistent dissimilarity data, rather than distances, fits into pedigree studies significantly where the distance correlation depends on Euclidean distances. The pedigree dissimilarity defined above does not satisfy the triangle inequality for general pedigrees, thus is not Euclidean distance. The Euclidean distances induced by the embedding resulting from RKE provides an approximation of the pedigree dissimilarities in our case. This allows us to validate our result of involving the non-metric pedigree dissimilarity in distance correlation by comparing with that obtained by using the embedded Euclidean distances.

1.6 Beaver Dam Eye Study

The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age-related ocular disorders. Subjects at baseline, examined between 1988 and 1990, were a group of 4926 people aged 43-86 years. Pedigree information was available for 2356 of the subjects. Although we will only use data from the baseline study for our experiments, five, ten, fifteen and twenty year follow-ups were also obtained. Familial relationships of participants were ascertained and pedigrees of different sizes were constructed for the subset of 1004 subjects who were dead prior to March 2011 with death ages ranging from 46 to 101 years.

Our goal is to use the data to study the association of familial relationships, lifestyle factors, diseases and mortality. The strategy is to first estimate the effects of lifestyle factors and diseases on mortality, i.e. death ages, based on the 1004 subjects using an SS-ANOVA model. The distance correlation is then applied to capture the associations with the estimated effects for a subgroup of 843 people coming from pedigrees containing two or more members. This results in 222 pedigrees in the data set, with sizes ranging from 2 to 23 subjects. Note that it is possible for two persons in one pedigree to be genetically unrelated. They become relatives because of their relationships with other members in the pedigree. The pedigree dissimilarity for such a pair is 1 as previously defined.

It is necessary to notice that the covariates can be continuous, binary and of different magnitude. In addition, the effects of the variables may not be linear in mortality, in which case a large pairwise distance of the covariates values may not result in a large pairwise distance of the death ages. Body mass index (bmi) is such an example in that both underweight and obesity are unhealthy and risky to longevity. In this case, the distance of bmi for two individuals, one with low value and the other with high value, is quite large, however, their death age distance may be small. Thus, instead of the original covariates, the estimated effects are preferred in the calculation of distance correlation because the fitted values are naturally assigned with weights and transformations.

For the above purpose, we fit an SS-ANOVA model of the form

$$
\begin{aligned}
deathage =& \mu + f_1(baseage) + \beta_{gender} I_{\{gender=F\}} && \Big\} fixed \\
&+ f_2(edu) + f_{12}(baseage : edu) + f_3(bmi) && \Big\} lifestyle \\
&+ \beta_{smoke} I_{\{smoke=no\}} + \beta_{inc} I_{\{inc>20T\}} && \\
&+ \beta_{diabetes} I_{\{diabetes=no\}} + \beta_{cancer} I_{\{cancer=no\}} && \Big\} disease \\
&++ \beta_{heart} I_{\{heart=no\}} + \beta_{kidney} I_{\{kidney=no\}} &&
\end{aligned}
$$

variable	units	description
deathage	years	death age
baseage	years	age at baseline
gender	F/M	gender
edu	years	highest year school/college completed
bmi	kg/m^2	body mass index
smoke	yes/no	history of smoking
inc	yes/no	household personal income $>$ 20T
diabetes	yes/no	history of diabetes
cancer	yes/no	history of cancer
heart	yes/no	history of cardiovascular disease
kidney	yes/no	history of chronic kidney disease

Table 1.1: Variable description in the SS-ANOVA model

gender = F	smoke = no	inc $>$ 20T	
1.141	1.349	0.546	
diabetes = no	cancer = no	heart = no	kidney = no
2.000	0.888	1.131	1.303

Table 1.2: Fitted effects of linear terms in the SS-ANOVA model

with variables being described in *Table 1.1* based on 1004 people. The terms in lines one, two to three, and four to five of the above equation are the fixed characteristics, lifestyle factors and disease variables respectively. Functions f_1, f_2 and f_3 are cubic splines and f_{12} uses the tensor product construction. The remaining covariates are unpenalized and modeled as linear terms with $I_{\{\cdot\}}$ as indicator functions. The fitted effects for *edu* and *bmi* are shown in *Figure 1.1*. The fitted effects of the linear terms are listed in *Table 1.2*.

Distance correlation, relying on pairwise distances, is the tool for measuring the association among the lifestyle factors, disease variables, mortality and pedigree. The cohort was restricted to the subgroup of 843 people coming from pedigrees with two or more

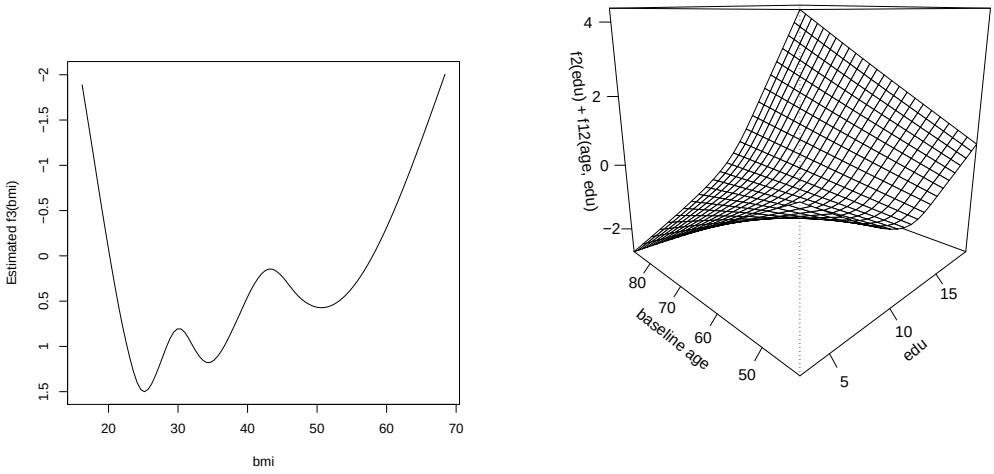

Figure 1.1: $f_3(bmi)$ (flipped y-axis) (top), and $f_2(edu) + f_{12}(baseage, edu)$ (bottom) are the fitted effects for bmi and education.

members. Up to now, the pedigree dissimilarities and Euclidean pairwise death age distances are ready for the calculation of the distance correlation. Lifestyle factors and disease variables get involved as the form of lifestyle factor scores and disease scores. The lifestyle factor score for an individual is the vector of the fitted effects for *smoke*, *bmi*, *edu* and *inc*. Similarly, the disease score is defined to be the vector of the fitted effects for the four disease variables. The Euclidean pairwise distances of the lifestyle factor scores and disease scores are constructed as the input information for lifestyle factors and disease variables in the distance correlation. Permutation tests are implemented to obtain the p-values of the distance correlations. The network in *Figure 1.2* summarizes the results. Both mortality and lifestyle factors are associated with familial relationships significantly. Heart disease and some cancers are known to run in families. However, the relationship between pedigree and disease variables in this part of the study is not significant at level 0.05. Included here are some pairs of relatives as distant as second cousins, which may be the cause of the weak signal. However, lifestyle factors, disease variables and mortality are closely associated with each other.

The theory of distance correlation is based on Euclidean pairwise distance. However, three of the above six distance correlations involve the non-Euclidean pedigree dissimilarity. The strategy is to validate the results by showing that the pedigree dissimilarity can be well approximated by Euclidean distances through embedding the subjects in Euclidean spaces by RKE. It is possible to establish the embedding effectively in the RKE framework

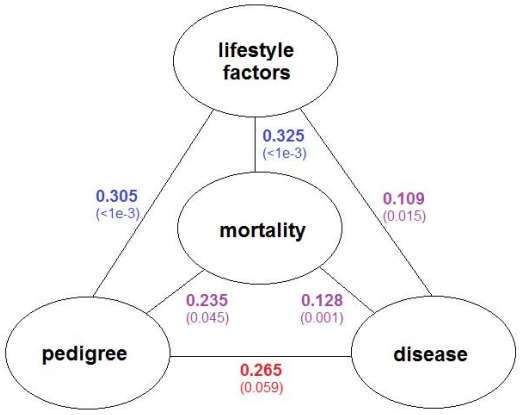

Figure 1.2: The network of lifestyle factors, disease variables, mortality and pedigree with distance correlations. The p-values obtained from permutation tests with 1000 replicates are presented in parenthesis. The significance level is distinguished by color: blue for p-value < 0.001, purple for p-value in $(0.001, 0.05)$, red for p-value > 0.05.

for a moderate sample size of subjects. However, it is too time consuming to solve the RKE semidefinite problem with the full dissimilarity information for 843 people in our case.

Alternatively, we break down the embedding into two steps. The first step only takes care of the within-pedigree dissimilarity. That is, we feed the familywise pedigree dissimilarities to RKE family by family so that it embeds the subjects into Euclidean spaces pedigree by pedigree. The kernel matrices obtained from RKE are then truncated to those leading eigenvalues that account for 95% of the matrix trace to create the "pseudo"-attribute embedding. The resulted familywise coordinates are put together in a way that each pedigree is assigned its own subspace which is orthogonal to the others. This ends up with a coordinate matrix being a horizontal concatenation of the familywise coordinates. The second step is to take into account of the out-pedigree dissimilarity, which requires pedigree specific variables. We assign one extra dimension to the coordinate matrix for each pedigree. The entries of this extra dimension are the pedigree specific variable for the family members and 0 for the rest of the subjects. This leads to a coordinate matrix being a function of the pedigree specific variables. Thus, the augmented coordinate matrix for the rth member in the pth pedigree takes the form of $(0, ..., 0, v^p, x^p_{r1}, ..., x^p_{rq}, 0, ..., 0)$, where v^p is the pedigree specific variable for the pth pedigree and q is the dimension of the subspace for the pth pedigree. The way to choose the pedigree specific variables is to maximize the Pearson's correlation between the vector form of the double centered pedigree dissimilarities and the vector form of the Euclidean pairwise distances resulting from the above coordinate matrix. The optimal value of the Pearson's correlation is 0.9907. *Figure*

1.3 shows a comparison of the embedded Euclidean pairwise distances and the pedigree dissimilarities for a subset of 100 subjects. It turns out that the non-Euclidean pedigree dissimilarities are well approximated by the embedded Euclidean distances.

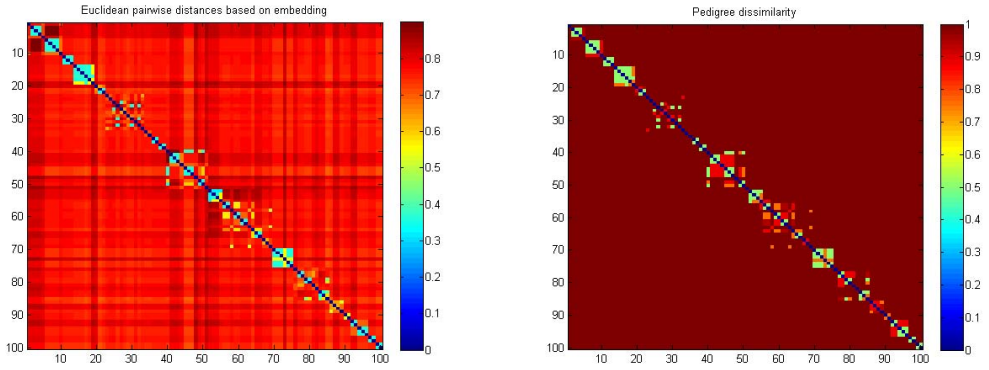

Figure 1.3: The comparison of the Euclidean pairwise distances by embedding and the pedigree dissimilarity for a subset of 100 subjects.

We could establish the distance correlations among the lifestyle factors, disease variables, mortality and pedigree based on the embedded Euclidean pairwise distances. The results are presented in *Figure 1.4* where the p-values are also obtained through permutation tests with 1000 replicates. Both the values of the distance correlation and the p-values are similar to those from the pedigree dissimilarity in *Figure 1.2*. The embedded results are slightly weaker than the original ones due to the shrinkage of RKE by penalizing high dimensionality of the space spanned by the kernel.

In addition to the study of all relatives, the analysis focusing on the full siblings shows that the signal of running in families gets stronger as the familial relationships become closer. The cohort are further restricted to 462 subjects who had at least one full sibling in the group of 843 people. To simplify the procedure, we change the pedigree dissimilarity for the full sibling pairs, which is shown to be Euclidean. The pedigree dissimilarity is assigned to be 0 for two full siblings and 1 for two unrelated persons. Suppose the subjects who are full siblings to each other are collected to different clusters and there are in total m such clusters. The members in the ith full sibling cluster are assigned the coordinates of length m, $(0, ..., 0, \frac{1}{\sqrt{2}}, 0, ..., 0)$, where the ith element is $\frac{1}{\sqrt{2}}$ and the rest are 0. The corresponding Euclidean pairwise distances are unchanged with the above pedigree dissimilarity being defined for full siblings. The distance correlations and p-values are summarized in *Figure 1.5* for the full siblings study. The three distance correlation values and related p-values involving familial relationships are strengthened compared to the

all relatives study, indicating that the signal of running in families is getting stronger as the subjects are closer. The other three associations are weaker due to the shrinkage of the sample size.

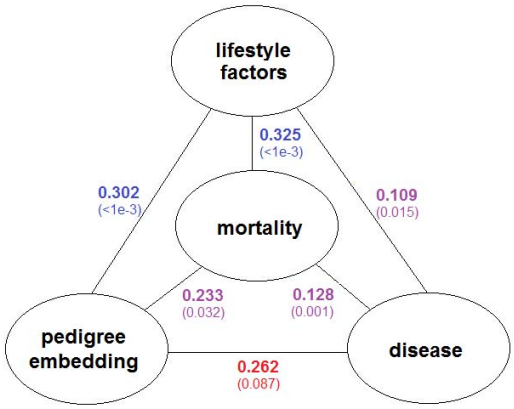

Figure 1.4: The network of lifestyle factors, disease variables, mortality and pedigree with distance correlations using the embedded Euclidean distances. The p-values obtained from permutation tests with 1000 replicates are presented in parenthesis.

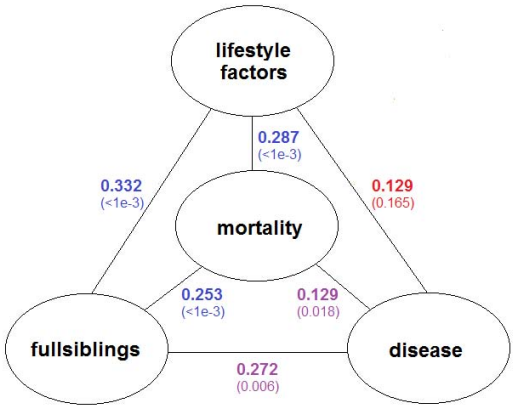

Figure 1.5: The distance correlations for full siblings study. The p-values obtained from permutation tests with 1000 replicates are presented in parenthesis.

For the full siblings study, the pairwise distances for mortality could be separated into two groups, group 0 collecting all the pairwise death age distances of full sibling pairs and group 1 for the unrelated pairs. This allows us to compare the difference between the mean of group 1 and the mean of group 0 and construct 95% Bootstrap percentile confidence interval for the test statistic with 10000 replicates. In the case of mortality, the average death age distance of full sibling pairs is 1.571 years less compared to that of two unrelated

variable	mortality	lifestyle	disease
group 0 mean	8.091	1.405	1.119
group 1 mean	9.662	1.654	1.229
difference	1.571	0.249	0.110
95% CI	(0.919, 2.211)	(0.167, 0.331)	(0.020, 0.202)

Table 1.3: Bootstrap percentile confidence intervals for the mean differences in the full siblings study

persons in the cohort. The corresponding 95% Bootstrap percentile confidence interval (CI) for the difference between the mean of group 1 and the mean of group 0 is $(0.919, 2.211)$. We could establish the analysis for the pairwise distances of lifestyle factors and disease variables in the same fashion. The observed test statistics and corresponding confidence intervals are summarized in *Table 1.3*. All the three mean differences between group 1 and group 0 are positive and the confidence intervals do not overlap 0, which means that the full siblings are significantly closer than unrelated people in terms of death age distances, lifestyle factor scores and disease scores.

1.7 Discussion

The Beaver Dam Eye Study, which began collecting data from a population aged 43 and older in 1988, and continues to the present, provides an ideal opportunity to apply some emerging statistical tools to examine questions regarding relationships between various kinds of information collected at the start of the study and mortality. Since the study contains a large number of people with relatives in the study, this provided an ideal opportunity to examine the correlations between familial relationships, lifestyle factors, disease and mortality. The methodological approach we have proposed here is easily adaptable to other studies for exploring relations between attributes of subjects with multiple clusters of observable attributes, simultaneously with other factors for which pairwise dissimilarities are observed. Some caveats with respect to the mortality data here are worth mentioning. The mortality data is censored at both ends, that is, we do not see cohorts of the oldest subjects who have died before the study began, and, at the other end, we have access to death ages only to those in the study who have died by March 2011. The left censoring is, to some extent accounted for in the presence of baseage in the SS-ANOVA model for

deathage–note that there is an interaction term for baseage and education, since it was observed that the oldest cohort in the study clearly had fewer years of formal education than younger members. This study does not use the subjects who would otherwise be included who do not have a recorded death age prior to March 2011. This is, of course a possible source of bias in the conclusions, and we hope to continue following this group as time goes on. Further research concerning residual lifetimes is ongoing, and the results may be able to utilize in addition the partial information contributed by subjects that are known to be alive past a particular time. Other information that is not used here includes attributes collected in the followup examinations. We cannot in this study exclude possible genetic effects behind the lifestyle factors - we only observe that our lifestyle factors significantly run in families, exactly why is beyond the scope of this project. We have shown that pairwise differences in lifestyle factors that run in families correlate well with pairwise differences in death age that also run in families, partially accounting for the familial death age effect. This leads to new questions to be asked about the complex relations between genetics, family structure, lifestyle factors, and other variables. We provide here an overall metholological approach which shows promise to help in answering these questions.

2 USING DISTANCE COVARIANCE FOR IMPROVED VARIABLE SELECTION WITH APPLICATION TO LEARNING GENETIC RISK MODELS

2.1 Introduction

The idea of feature screening came along as high dimensional data were collected in modern technology. It was aimed at dealing with the challenges of computational expediency, statistical accuracy and algorithmic stability due to high dimensionality. Fan and Lv proposed the sure independence screening (SIS), Fan and Lv (2008) and showed that the Pearson correlation ranking procedure possessed a sure screening property for linear regression with Gaussian predictors and responses. A new feature screening procedure for high dimensional data based on distance correlation, Székely et al. (2007), named DC-SIS, was presented in Li et al. (2012). DC-SIS retained the sure screening property of the SIS, and additionally possessed new advantages of handling grouped predictors and multivariate responses by using distance correlation. Moreover, since distance correlation was applicable to arbitrary distributions, DC-SIS could also be used for screening features without specifying a regression model between the response and the predictors, and thus was robust to model mis-specification.

However, both SIS and DC-SIS relied on a user-specified model size d which decided the number of predictors being selected. Let the sample size be n, d was chosen to be multipliers of the integer part of $n/\log n$ in Fan and Lv (2008) and Li et al. (2012) which did not depend on any other characteristics of the data. As pointed out by a referee of Li et al. (2012), the choice of d was of great importance in practical implementation and might influence the screening results significantly. Our study is aimed at fixing this shortcoming by including an automatic stopping criteria for DC-SIS based on the property of distance covariance.

The screening procedures may fail if a feature is marginally uncorrelated, but jointly correlated with the response, or in the reverse situation where a feature is jointly uncorrelated but has higher marginal correlation than some important features. An iterative SIS was proposed in Fan and Lv (2008) to fix this problem. Current research interest involves dealing with this drawback but this work is not related to this quest.

We demonstrate our improved method through two real examples. The small round blue cell tumors (SRBCT) data were relatively easy to classify and had been studied extensively. The Cancer Genome Atlas (TCGA) ovarian cancer data, however, were much more challenging due to the large number of genes and limited sample size. The target

was to identify the important genes that contribute to the sensitive or resistant status after receiving a particular chemotherapy treatment. A substantial fraction of the population was difficult to classify and a “withholding decision" option is allowed in the support vector machine with reject option model to adapt to this fact. A multiple cross validation is used to quantify uncertainty given a humongous number of candidates, and we see a commonly observed dilemma that different variables are selected by using different subsets of the data. Comparison between the results from the original data and those from the data obtained by randomly permuting the response provide further justification on our conclusions. Furthermore, the multiple cross validation on the permuted data discloses the existence of spuriously correlated variables in high dimensional data and thus failure of variable selection and model building based on training data.

2.2 Some Preliminaries

Distance correlation

Székely et al. (2007) proposed distance correlation as a measurement of dependence between two random vectors. The method has been successfully applied to various problem, see Kong et al. (2012) for example. To be specific, the authors defined the distance covariance between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ to be

$$\mathrm{V}^2(X, Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(s,t) - f_X(s) f_Y(t)|^2}{|s|_p^{1+p} |t|_q^{1+q}} dt\, ds$$

where $f_{X,Y}(s,t)$, $f_X(s)$, and $f_Y(t)$ are the characteristic functions of (X, Y), X, and Y respectively, and c_p, c_q are constants chosen to produce scale free and rotation invariant measure that doesn't go to zero for dependent variables. The idea is originated from the property that the joint characteristic function factorizes under independence of the two random vectors. This leads to the remarkable property that $V^2(X, Y) = 0$ if and only if X and Y are independent.

The sample version of distance covariance and distance correlation involves pairwise distances. For a random sample $(X, Y) = \{(X_k, Y_k) : k = 1, ..., n\}$ of n i.i.d random vectors (X, Y) from the joint distribution of random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$, the Euclidean distance matrices $(a_{ij}) = (|X_i - X_j|_p)$ and $(b_{ij}) = (|Y_i - Y_j|_q)$ with $i, j = 1, \ldots, n$ are computed. Define the double centering distance matrices

$$A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, \quad i, j = 1, \ldots, n,$$

where

$$\overline{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \overline{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \quad \overline{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij},$$

similarly for $B_{ij} = b_{ij} - \overline{b}_{i\cdot} - \overline{b}_{\cdot j} + \overline{b}_{\cdot\cdot}$, $i, j = 1, \ldots, n$. Then, the sample distance covariance $V_n(X, Y)$ is defined by

$$V_n^2(X, Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}.$$

The sample distance correlation $R_n(X, Y)$ is defined by

$$R_n^2(X, Y) = \begin{cases} \dfrac{V_n^2(X, Y)}{\sqrt{V_n^2(X)V_n^2(Y)}}, & V_n^2(X)V_n^2(Y) > 0; \\ 0, & V_n^2(X)V_n^2(Y) = 0, \end{cases}$$

where the sample distance variance is defined by

$$V_n^2(X) = V_n^2(X, X) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}^2.$$

Feature screening via distance correlation (DC-SIS)

Fan and Lv (2008) proposed sure independence screening (SIS) procedure based on the Pearson correlation for feature selection. The distance correlation version of this technique (DC-SIS) was studied in Li et al. (2012). With a user-specific model size d, the variables whose distance correlations with the response ranking from 1st to dth in decreasing order were selected. The authors explored the theoretic properties of the DC-SIS and proved that the DC-SIS kept the desired sure screening property established in Fan and Lv (2008). Moreover, due to the property of distance correlation, DC-SIS procedure was robust to model mis-specification, which was demonstrated in their simulations.

2.3 Improving DC-SIS using distance covariance

Theorem 2.1. *Suppose random vectors $X, Z \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, and assume Z is independent of (X, Y), then*

$$V^2(X + Z, Y) \leqslant V^2(X, Y), \tag{2.1}$$

where V is the population distance variance defined in Székely et al. (2007).

Proof.

$$\begin{aligned} V^2(X+Z,Y) =& \| f_{X+Z,Y}(t,s) - f_{X+Z}(t)f_Y(s) \|^2 \\ =& \frac{1}{c_p c_q}\int_{\mathcal{R}^{p+q}} \frac{1}{|t|_p^{1+p}|s|_q^{1+q}} |f_{X+Z,Y}(t,s) - f_{X+Z}(t)f_Y(s)|^2 dtds. \end{aligned}$$

The following fact follows from the definition of characteristic function and independence assumption.

$$\begin{aligned} & |f_{X+Z,Y}(t,s) - f_{X+Z}(t)f_Y(s)|^2 \\ =& |Ee^{it^T(X+Z)+is^TY} - Ee^{it^T(X+Z)}Ee^{is^TY}|^2 \\ =& |Ee^{it^TX+is^TY}Ee^{it^TZ} - Ee^{it^TX}Ee^{it^TZ}Ee^{is^TY}|^2 \\ =& |f_{X,Y}(t,s)f_Z(t) - f_X(t)f_Z(t)f_Y(s)|^2 \\ =& |f_Z(t)|^2|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2, \end{aligned}$$

Since $|f_Z(t)| \leqslant 1$ by the property of characteristic function[1], we have

$$|f_{X+Z,Y}(t,s) - f_{X+Z}(t)f_Y(s)|^2 \leqslant |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2,$$

which implies

$$\begin{aligned} V^2(X+Z,Y) \leqslant & \frac{1}{c_p c_q}\int_{\mathcal{R}^{p+q}} \frac{1}{|t|_p^{1+p}|s|_q^{1+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 dtds \\ =& \| f_{X+Z,Y}(t,s) - f_{X+Z}(t)f_Y(s) \|^2 \\ =& V^2(X,Y). \end{aligned}$$

□

We know that if $E|X|_p < \infty, E|X+Z|_p < \infty$ and $E|Y|_p < \infty$, then almost surely

$$\begin{aligned} \lim_{n\to\infty} V_n^2(X+Z,Y) &= V^2(X+Z,Y), \\ \lim_{n\to\infty} V_n^2(X,Y) &= V^2(X,Y). \end{aligned}$$

Thus, for the sample distance covariance, if n is large enough, we should have

$$V_n^2(X+Z,Y) \leqslant V_n^2(X,Y),$$

[1] In Kosorok (2009), the author obtained equality here which is incorrect.

under the assumption of independence between (X, Y) and Z.

In the case where (X, Z) is of interest, which is the usual situation for variable selection setting, we could use the above theorem by incorporating degenerated random vectors as follows. Suppose $X \in \mathbb{R}^{p_1}$ and $Z \in \mathbb{R}^{p_2}$, then we augment X and Z to be $\tilde{X} = (X, 0_{p_2})$ and $\tilde{Z} = (0_{p_1}, Z)$ respectively. $\tilde{X}$ and $\tilde{Z}$ are therefore of the same dimension and $\tilde{X} + \tilde{Z} = (X, Z)$.

We implemented the above theorem as a check for stopping for DC-SIS. For the original DC-SIS, it required a user-specified model size d, which was always chosen as multipliers of the integer part of $n/\log n$. For our improved screening procedure with distance correlation, we first ranked the importance of $x_i, i = 1, ..., p$ using the marginal distance correlations with the response as DC-SIS did and initialized $\mathcal{S}$ as the singleton including the index of the top one variable. Instead of selecting the top d variables, we kept adding variables to $x_{\mathcal{S}} = \{x_i : i \in \mathcal{S}\}$ according to the ordered list of variables until observing a decrease in the distance covariance between $x_{\mathcal{S}}$ and y. The procedure took the following steps and we denoted the procedure as DCOV method.

1. Calculate marginal distance correlations for $x_i, i = 1, ..., p$ with the response.
2. Rank the variables in decreasing order of the distance correlations. Denote the ordered variables as $x_{(1)}, x_{(2)}, ..., x_{(p)}$. Start with $x_{\mathcal{S}} = \{x_{(1)}\}$.
3. For i from 2 to p, include $x_{(i)}$ to $x_{\mathcal{S}}$, i.e., updating $x_{\mathcal{S}}$ by the concatenated variables $(x_{\mathcal{S}}, x_{(i)})$, if $V_n^2(x_{\mathcal{S}}, y)$ does not decrease. Stop otherwise.

2.4 Real application on SRBCT data

The small round blue cell tumors (SRBCTs) are 4 different childhood tumors named so because of their similar appearance on routine histology, which makes correct clinical diagnosis extremely challenging. However, accurate diagnosis is essential because the treatment options, responses to therapy and prognoses vary widely depending on the diagnosis. They include Ewing's family of tumors (EWS), neuroblastoma (NB), non-Hodgkin lymphoma (in our case Burkitt's lymphoma, BL) and rhabdomyosarcoma (RMS). The SRBCT data being published in Khan et al. (2001) included the expression of 2308 genes measured on 63 samples (23 EWS, 8 BL, 12 NB and 20 RMS). This data are known as an easy-classified example and have been studied by many. Lee et al. (2004) using the multicategory SVM is one of several methods that have excellent classification results on this data set. Hence, we focus more on the selected genes.

We applied our improved feature screening procedure on this dataset and compared our selection of genes with the 96 top genes reported in Khan et al. (2001). This is a multicategory

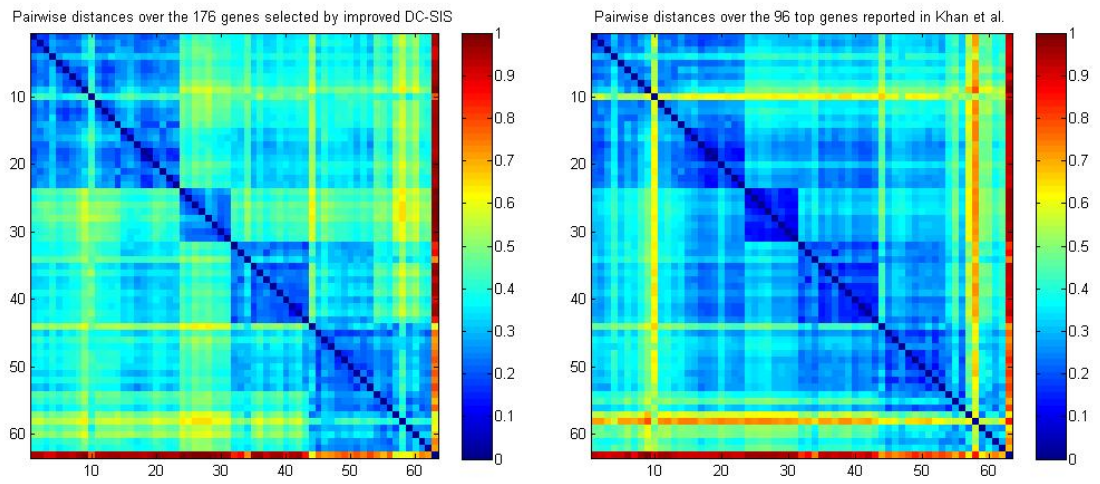

Figure 2.1: Comparison of pairwise distances between the two selections of genes. Left and right panel present the pairwise distances of the 63 samples over the improved DC-SIS selection of 176 genes and the 96 reported genes in Khan et al. (2001) respectively.

classification and the genes were screened in a one-versus-rest fashion. Specifically, for each of the four different types of tumors, we generated a response indicator vector taking value of 0 if the sample came from the current interested type and 1 otherwise. This allowed us to implement our screening procedure and obtained the genes which showed high distance correlation with the current type of tumor. The four groups of selected genes were combined as a whole collection of in total 176 genes. 47 genes turned out to be in common for the DCOV selection and the top 96 genes used in Khan et al. (2001).

We further examined the power of these two groups of genes in differentiating the 4 types of tumors by presenting the pairwise distances of the 63 samples (*Figure 2.1*). As shown in the plot, the samples were arranged in the order of EWS, BL, NB and RMS. The pairwise distances resulted from the two selections of genes were scaled to maximum of 1 respectively so that they shared the same magnitude. Both groups of genes could tell the 4 types of tumors apart. Compared with the 96 genes from Khan et al. (2001), however, the 176 DCOV selected genes show better distinguishability and clearer contrast over the 4 classes. Moreover, the right panel almost missed the samples labeled from 57 to 62 in the class of RMS but the 176 DCOV genes could recognize them with big differences between the in and out class pairwise distances. The dataset were known to be easy for classification and both sets of genes were able to classify the testing set of 20 samples perfectly via k-nearest neighbor method with $k = 3$.

2.5 Real application on TCGA ovarian cancer data

Data description

Ovarian cancer is the fifth-leading cause of cancer death among women in the United States; 22,240 new cases and 14,030 deaths were estimated to have occurred in 2013, see website; Bell et al. (2011). The standard treatment for high-grade serous ovarian cancer is aggressive surgery followed by chemotherapy. Despite treatment, a vast majority of ovarian cancer patients eventually relapse and die of their disease with a major cause of chemotheraphy resistance ,Selvanayagam et al. (2004). Identification and prediction of patients with chemoresistant thus become important for improving the outcome of ovarian cancer.

The Cancer Genome Atlas (TCGA) collected high-quality, high-dimensional, and multi-modal genetic data from women with ovarian cancer. There were 279 samples with explicit chemostatus and gene expression (Affymetrix HT-HGU133a) data in the public set. among which 191 subjects were sensitive to chemotherapy and 88 were chemoresistant. Expression data for 12042 genes after log transformation are used for analysis. The issue is to explore whether there are genes whose expression pattern is strongly correlated with the response indicating chemotherapy status.

DCOV gene selection results based on all the observations

Our feature screening procedure on the gene expression data for the 279 patients selected 82 genes, among which 5 were reported to be related to ovarian cancer in the literature. IGFBP5 ranked as the 5th is one of the six members of insulin-like growth factor-binding protein (IGFBP) family and is known to be important for cell growth control, induction of apoptosis and other IGF-stimulated signaling pathways. IGFBP5 expression is shown to prevent tumor growth and inhibited tumor vascularity in a xenograft model of human ovarian cancer and is suggested that IGFBP5 plays a role as tumor suppressor by inhibiting angiogenesis, Rho et al. (2008). GPR3, the 7th, is a member of a family of G-protein couple receptors whose activation of PKA and subsequent increase of cyclic AMP level, promotes meiotic arrest in the oocyte, Mehlmann et al. (2004). Mice deficient in GPR3 display premature ovarian aging and loss of fertility Ledent et al. (2005). MAPK4, the 18th, is a member of MAPK signaling pathway. MAPK signal transduction cascade is dysregulated in a majority of human tumors, Basu et al. (2009). It is suggested playing an important role in molecular diagnostics and molecular therapeutics for lowgrade ovarian cancer, Bast and Mills (2010). FZD5 ranked as the 22th encodes Frizzled-5 protein, which is believed to be

the receptor for the Wnt5A ligand, Thiele et al. (2011). The Wnt5A ligand plays a context-dependent role in human cancers. It has been demonstrated that Wnt5a is expressed at significantly lower levels in human Epithelial ovarian cancer (EOC) cell lines and in primary human EOCs compared with either normal ovarian surface epithelium or fallopian tube epithelium, Bitler et al. (2011). FGF22, the 56th, is a member of Fibroblast Growth Factors (FGFs) family, whose members possess broad mitogenic and cell survival activities, and are involved in a variety of biological processes, including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion. The inhibition of FGFR2, which is a member of this family, has been found to increase cisplatin sensitivity in ovarian cancer, Cole et al. (2010).

39 pathways were found to be associated with the 82 genes, among which 3 pathways are known to be important for ovarian cancer. MAPK signaling pathway is suggested playing an important role in molecular diagnostics and molecular therapeutics for lowgrade ovarian cancer, Bast and Mills (2010). Wnt signaling pathway is best known for its role in tumorigenesis. Bast and Mills (2010) demonstrated the difference in Wnt signaling pathway between normal ovarian and cancer cell lines and between benign tissue and ovarian cancer. They also pointed out that those differences implicate that Wnt signaling leads to ovarian cancer development despite the fact that gene mutations are uncommon. Yin et al. (2011) suggested that genetic variants in TGF-β signaling pathway are associated with ovarian cancer risk and may facilitate the identification of high-risk subgroups in the general population.

Support vector machine with reject option

We estimated the probabilities of being chemosensitive or chemoresistant by a penalized Bernoulli likelihood main effect spline model using the `R` package `gss` in Gu (2007). Aside from the additive expression effects of the selected 82 genes, we also included two more covariates, namely the cancer grade and cancer stage of the subjects. Cancer grade is an indicator for grade 2 and grade 3. Cancer stage indicates whether the subject is in stages IIIC and IV or not. As shown in *Figure 2.2*, the estimated probabilities have high density around small and large values for sensitive and resistant patients respectively, with overlapping in the middle values. This suggested that we were less confident about the chemostatus for the patients in the middle range and so we sought a principled approach which withholds decision for such cases.

Wegkamp et al. (2011); Bartlett and Wegkamp (2008) investigated the support vector machines with a built-in reject option for binary classification where the results of classifi-

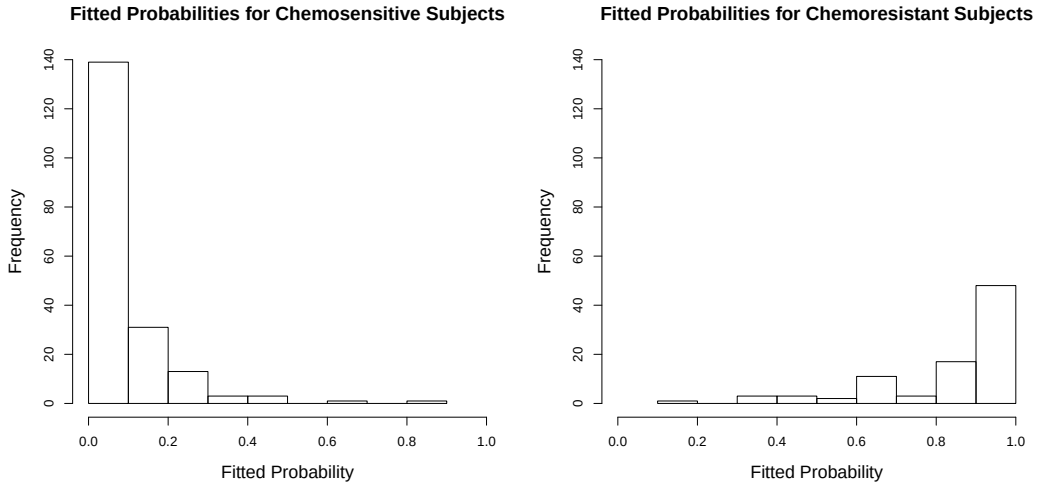

Figure 2.2: Fitted probabilities by penalized Bernoulli likelihood model with the 82 genes.

cation could be $-1, +1$ or withhold decision. Given a discriminant function $f : \mathcal{X} \to \mathbb{R}$, the method only reports $sgn(f(x)) \in \{-1, 1\}$ if $|f(x)| > \delta$ and withholds decision if $|f(x)| \leqslant \delta$. Suppose that the cost of making a wrong decision is 1 and that of rejecting to make a decision is $d \in [0, \frac{1}{2}]$, then an proper risk function is

$$L_{d,\delta}(f) = E l_{d,\delta}(Yf(X)) = P\{Yf(X) < -\delta\} + dP(|Yf(X)| \leqslant \delta)$$

with the discontinuous loss function

$$l_{d,\delta}(z) = \begin{cases} 1, & \text{if } z < -\delta; \\ d, & \text{if } |z| \leqslant \delta; \\ 0, & \text{otherwise.} \end{cases}$$

The classifier associated with the discriminant function

$$f_d^*(x) = \begin{cases} -1, & \text{if } \eta(x) < d; \\ 0, & \text{if } d \leqslant \eta(x) \leqslant 1 - d; \\ +1, & \text{if } \eta(x) > 1 - d, \end{cases}$$

with $\eta(x) = P\{Y = +1 | X = x\}$ minimizes the risk $L_{d,\delta}(f)$ with

$$L_d^* = L_{d,\delta}(f_d^*) = E \min\{\eta(X), 1 - \eta(X), d\}.$$

To avoid working with the discontinuous loss $l_{d,\delta}$, Wegkamp et al. (2011); Bartlett and Wegkamp (2008) proposed a convex surrogate loss, which is the generalized hinge loss,

$$\phi_d(z) = \begin{cases} 1 - az, & \text{if } z < 0; \\ 1 - z & \text{if } 0 \leqslant z < 1; \\ 0, & \text{otherwise,} \end{cases}$$

where $a = (1-d)/d \geqslant 1$. It followed that f_d^* also minimizes the risk associated with ϕ_d over all measurable $f : \mathcal{X} \to \mathbb{R}$.

The discriminant functions f took the form $f_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x)$ based on a set of known functions $f_j : \mathcal{X} \to \mathbb{R}$ and coefficients $\lambda_j \in \mathbb{R}, 1 \leqslant j \leqslant M$. The coefficients were chosen to minimize the empirical risk

$$\hat{R}_\phi(f_\lambda) = \frac{1}{n} \sum_{i=1}^n \phi(Y_i f_\lambda(X_i)).$$

To reflect the preference for sparse solutions, which is desirable when M is large compared to the sample size n, an l_1 type restriction $\|\lambda\|_1 = \sum_{j=1}^M |\lambda_j|$ was incorporated in Wegkamp et al. (2011) and f_λ is estimated by $f_{\hat{\lambda}_r}$, where

$$\hat{\lambda}(r) = \arg\min_{\lambda \in \mathbb{R}^M} \hat{R}_\phi(f_\lambda) + r\|\lambda\|_1 \tag{2.2}$$

and $r > 0$ is a tuning parameter. We followed Wegkamp et al. (2011) to call this model support vector machines with reject option(SVM-R).

The authors in Bartlett and Wegkamp (2008) also showed that the choice of $\delta = 1/2$ gives the optimal bound established by the excess risk of ϕ_d on the excess risk $L_{d,\delta} - L_d^*$ for any fixed $d \in [0, 1/2)$ and measurable function f. For this reason, we fixed $\delta = 1/2$ for our practical use of the method. Furthermore, we took the set of known functions $f_j : \mathcal{X} \to \mathbb{R}$ to be linear functions of the log transformation on the 12024 genes. The optimization problem (2) was formulated into a linear programming task and solved using MATLAB.

Figure 2.3 presents the 82 genes for the 279 subjects in groups according to the SVM-R classification results. The results correspond to the particular choice of $d = 1/4$ and $r = 4$ to illustrate the benefits of SVM-R. As shown in the plot, there is a big difference in the gene expression between the subjects assigned to be resistant and sensitive. The behavior of the 82 genes for those without a certain decision tends to be in-between.

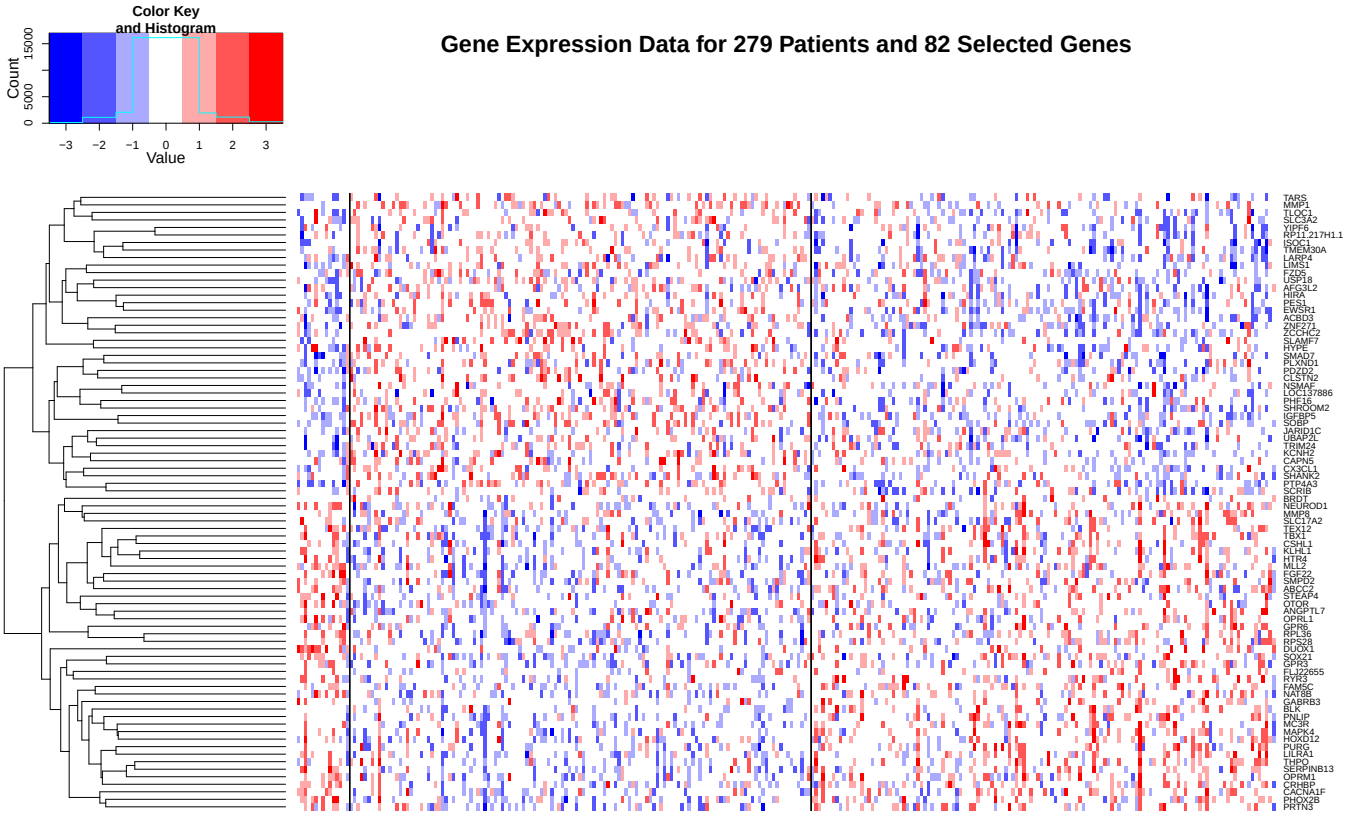

Figure 2.3: Gene expression data for the 82 selected genes and 279 subjects with SVM-R classification for $d = 1/4$ and $r = 4$. The subjects are grouped according to their assigned decisions by the SVM with a reject option. The left group involves 15 patients (1 sensitive and 14 resistant) classified to be resistant. The middle group is assigned to be sensitive and contains 123 sensitive and 8 resistant subjects. 67 sensitive patients and 66 resistant patients with a withhold decision are shown in the right group.

Five fold cross validation

In order to choose the tuning parameter in SVM-R, we need to hold aside a tuning set before selecting the genes. Leaving out different observations leads to different gene selection results. Here we applied a five fold cross validation analysis to examine the variations of selections of genes and SVM-R model fitting results across different partitions of the dataset. The implementation followed the steps below.

1. Randomly partition the 279 subjects into 5 non-overlapping folds.

2. Select genes from the 12024 genes based on 4 folds as the training set.

3. Build SVM-R model with the selected genes and the two cancer status variables based on the training set.

	S_1	S_2	S_3	S_4	S_5
S_1	53				
S_2	16	77			
S_3	23	21	87		
S_4	18	16	15	33	
S_5	27	30	31	21	94
82 genes	38	38	44	28	50

Table 2.1: Pairwise intersections of $S_1, \ldots, S_5$ and the 82 genes. The diagonal numbers are the numbers of selected genes in each S_i.

4. Use the leaving-out fold as the tuning set to choose the tuning parameter for SVM-R with mean I—loss, defined below, as the criteria.

5. Repeat $2. - 4.$ for the 5 folds.

The I—loss for a subject is 1 if a misclassification occurs, d if a withholding decision is made and 0 otherwise. The mean I—loss is the average over the I—losses for all the subjects in a given set of data. We looked for tuning parameter values minimizing the mean I—loss.

The above procedure produced 5 selections of genes before SVM-R, namely $S_1, \ldots, S_5$. In addition, we also have the 82 genes selected from all the subjects previously. Table 2.1 presents the pairwise intersections of these 6 sets with each other. The union of the 5 selections includes 211 genes, which covers 77 genes in the 82 genes. 73 out of 211 genes have frequency more than 1 where 63 of them appear in the 82 genes. After implementing SVM-R, the union of genes is reduced to 98 genes. *Figure 2.4* displays the histogram of these 211 genes colored by the frequency after SVM-R runs for $d = 1/5$. The pink region denotes the parts further eliminated by SVM-R, which is consistent with the DCOV selection in that SVM-R further rules out the genes with low frequency in the union.

Multiple cross validation

In order to consider uncertainty in variable selection and model building due to different partitions of the dataset, we further extended the five fold cross validation to multiple cross validation (MCV) and assessed the prediction power through the following procedure. The results were summarized in the upper part of Table 2.2.

1. Randomly partition 279 samples into a $1/5$ tuning set, a $2/3 \times 4/5 = 8/15$ training set and a $1/3 \times 4/5 = 4/15$ testing set.

2. Select genes from the 12024 genes using the proposed method on the training set.

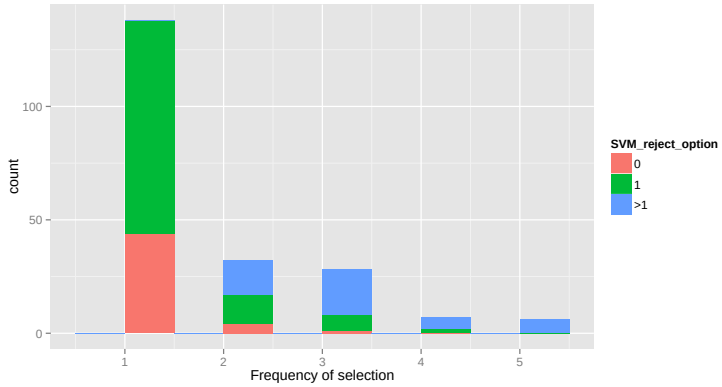

Figure 2.4: Frequency for the union of $S1, \cdots, S5$, colored by frequencies after SVM-R for $d = 1/5$.

3. Build SVM-R model using the selected genes and the two cancer status variables based on the training set.

4. Use the tuning set to choose the tuning parameter for SVM-R with mean l—loss as the criteria.

5. Use the model with chosen parameter to predict labels for the testing set.

6. Repeat $1. - 5.$ for 50 times.

To understand more about the 50 models, we further explored the prediction results for $d = 1/5$. The prediction labels from the 50 models were aggregated, following the idea of ensemble methods. The result for each individual was recorded in a vector of three frequencies, namely the frequency of being classified as sensitive subjects, the frequency of obtaining a withholding and the frequency of being assigned to be resistant out of the 50 models. Let (s_i, w_i, r_i) be the vector for the ith patient.

A finer analysis was conducted by looking at the strength of being sensitive or resistant according to (s_i, w_i, r_i). A voting score v_i was defined as $(s_i - r_i)/w_i$. Hence, a positive v_i indicated a tendency of being sensitive whereas a negative v_i suggested more possibility of being resistant.

To understand the meaning of v_i, we first divided the voting scores into 4 consecutive intervals, each covering about $1/4$ of the population. The last interval was further partitioned into two to identify a subgroup of patients with extremely large v_i's and high homogeneity in the class label. Table 2.3 (upper part) presents the 5 intervals and describes the distribution of v_i's as well as the proportion of sensitive subjects within each range of v_i, compared to the overall proportion of sensitive patients, i.e. $191/279$, in the population. It

	d	num of reps with decision	mean training accuracy(std)	mean testing accuracy(std)	mean num training with decision(std)	mean num test with decision(std)
original	1/3	50	0.8319(0.0914)	0.6943(0.0544)	101.9400(27.8043)	49.1800(15.4295)
	1/4	43	0.9371(0.0336)	0.7807(0.1250)	43.0698(27.0338)	20.1860(12.9638)
	1/5	37	0.9420(0.0358)	0.8215(0.1460)	24.4595(21.3874)	11.4865(10.0626)
permute	1/3	49	0.7984(0.0078)	0.6910(0.0426)	112.1837(26.4352)	55.5918(13.9954)
	1/4	28	0.9225(0.0023)	0.6867(0.0810)	56.9643(21.4346)	25.2857(10.4132)
	1/5	9	0.9686(0.0015)	0.7071(0.1322)	53.4444(28.3333)	24.5556(13.0682)

Table 2.2: Results for the 50 individual replications for $d = 1/3, 1/4$ and $1/5$. The upper and lower part are results for the original and permuted data respectively. The third column shows the number of replications out of 50 with at least one definite decision made on the testing set. The fourth and fifth columns of the table conclude the mean training and testing accuracies with standard deviation in the parenthesis respectively restricted to the repetitions with decision made. The last two columns display the mean and standard deviation for the number of patients assigned decisions for the training and testing sets respectively given the replications with decision made.

turned out that the trend of being sensitive weakened monotonically as the voting score decreased. The stratification specified a subgroup of 15 patients, who possessed the greatest voting scores, with very high accuracy to be chemosensitive. The next highest voting score subgroup of 47 subjects also showed strong confidence of being sensitive compared to the sample proportion. The conclusion from partitioning the voting scores was conservative but led to more convincing and steady classification results.

		voting score				
		$(-0.1, 0]$	$(0, 0.1]$	$(0.1, 0.2]$	$(0.2, 0.4]$	$(0.4, 1.5]$
original	frequency	76	74	67	47	15
	proportion	0.5658	0.6486	0.7164	0.8085	0.9333
permuted	frequency	145	67	43	24	0
	proportion	0.6759	0.6866	0.7209	0.6667	NA

Table 2.3: Frequency of voting score v_i's and proportion of sensitive subjects in each subinterval for $d = 1/5$. The upper and lower parts correspond to the original and permuted data respectively.

Each replication of the 50 multiple cross validations gave rise to a different collection of selected genes. This issue is common to selecting variables from a humongous number of candidates, in the not-low-hanging-fruit situation. The union of the 50 gene selections before SVM-R consisted of 1245 genes and included all the 82 genes discussed previously. 34 out of 1245 genes were chosen at least 10 times, where 33 of them appeared in the 82 genes, but very few appeared in more than 25 runs. The l_1 penalty provided additional elimination, and for $d = 1/5$, 498 out of 1245 genes remained after SVM-R runs. *Figure 2.5* displays the histogram for the 1245 genes before SVM-R. We distinguished their frequency after SVM-R by different colors. It is shown that a large number of genes with low frequency are further deleted by SVM-R model, i.e. pink color.

Permutation of the response

Our method involved several components, including variable selection, SVM-R, MCV and the voting score, which were interacting with each other, and led to 15 patients with over 93% accuracy to be sensitive for $d = 1/5$. To further understand the mechanism and to demonstrate that the outcomes were not produced by noises, we randomly permuted the response and went through the whole procedure to compare the results with those for the original data.

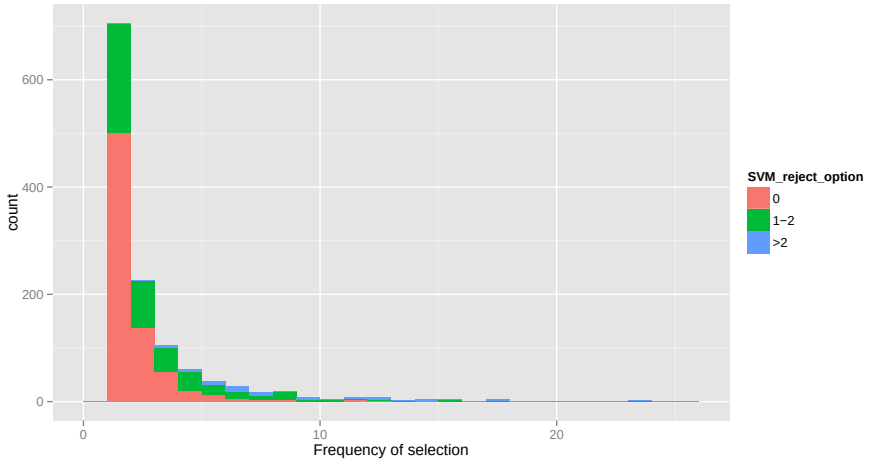

Figure 2.5: Frequency for 1245 genes being selected by DCOV method, colored by frequencies after SVM-R for $d = 1/5$.

It followed that the DCOV method selected genes spuriously correlated with the permuted response based on the training data in each replication of the 50 MCVs. The maximum distance correlation value of the selected genes in each repetition was very close to that for the original data. The highly correlated genes appeared due to the high dimensionality of over 12000 genes and less than 200 training samples.

However, the MCV step played the role of a safeguard against the fake signals. As Table 2.2 depicted, the mean training accuracies for the original and permuted data showed similar behavior for the original and permuted data, meaning that the selected genes were indeed important for the training data. Thus, the chosen genes in the permutation set should provide little prediction power for the tuning and testing data. Hence, the validation sets selected large tuning parameter values driving all the patients with no decision for many of the 50 replications for $d = 1/4$ and $1/5$. This did not happen for $d = 1/3$ since the sample ratio $191/279$ is slightly greater than $1/3$. For the rest of the replications with decision making, the mean testing accuracies for the permuted data remained at the level of the sample proportion of sensitive subjects for all three values of d, which deviated much from the increasing pattern in the original data.

These suggested that the MCV procedure was able to provide double fail-secure for fake signals. On one hand, SVM-R placed a cap on the conditional probability of misclassification and eliminated the replications where the selected genes could not produce results achieving the specified confidence on the validation set. On the other hand, the mean testing accuracies on the replications passing through the safeguard of tuning sets would be no better than assigning everyone to the sensitive class when there was no real signal.

Furthermore, the poor prediction performance of the 50 individual models ended up with unsurprisingly disappointing voting score results for the permuted data, as shown in the lower part of Table 2.3. Many of the patients obtained a relatively small value of the voting score and nobody got a score in the range where the original data had the highest accuracy, meaning that the confidence was quite low. Moreover, the stratified ratios of sensitive subjects for different ranges of the voting scores did not show anything insightful other than being around the sample proportion.

2.6 Summary of procedures in TCGA Ovarian Cancer data

Several components were included in the TCGA Ovarian cancer data. Here we summarize the pieces in the algorithm below so that potential users are able to follow the procedure with new datasets.

1. Set replication number N for MCV and $p_{train}, p_{test} \in (0,1)$ with $p_{train} + p_{test} \leqslant 1$ for the proportion of training and testing set.

2. For i runs from 1 to N

   a) Randomly partition the data into a p_{train} training set, a p_{test} testing set and a $1 - p_{train} - p_{test}$ tuning set. (If the model in mind does not need parameter tuning, one can omit the tuning set with $p_{train} + p_{test} = 1$.)

   b) Select variables using DCOV on the training set:

      - Calculate marginal distance correlations for $x_j, j = 1, \cdots, p$ with the response on the training set.
      - Rank the variables in decreasing order of the distance correlations. Denote the ordered variables as $x_{(1)}, x_{(2)}, \cdots, x_{(p)}$. Start with $x_{\mathbb{S}} = \{x_{(1)}\}$.
      - For j from 2 to p, include $x_{(j)}$ to $x_{\mathbb{S}}$, i.e., updating $x_{\mathbb{S}}$ by the concatenated variables $(x_{\mathbb{S}}, x_{(i)})$, if $V_n^2(x_{\mathbb{S}}, y)$ on the training set does not decrease. Stop otherwise.

   c) Build models, for example SVM-R in the TCGA Ovarian case, using the selected variables on the training set.

   d) Use the tuning set to select parameters for the model.

   e) Use the model with chosen parameter to predict the response value for the testing set.

3. Aggregate the predicted results for N replications:

- For classification task, one may use majority vote to obtain the final labels. The voting scores for the distinct labels can be used to evaluate the strength of being classified to each category for every observation.
- For regression task, one may take the average of the N predicted values as the final prediction.

2.7 Discussion

The paper introduced a new variable selection procedure based on the property of distance covariance and demonstrated the application through two examples. The small round blue cell tumors data played a role of a toy example to show that the performance of the proposed method worked well in easy cases. The TCGA ovarian cancer data, however, were much more challenging to deal with due to the humongous number of variables and very limited sample size. The uncertainty of variable selection was discussed through gene selection results using random subsets of the data. The support vector machine with reject option was used to withhold decision for subjects who were difficult to classify. An ensemble method of combining models built on random subsets of the data was implemented to assess the prediction performance. Although we had applied these tools (DCOV, SVM-R, MCV) to biomedical data in the paper, we argue that they are quite portable across disciplines.

As shown in Table 2.3, a small portion of the model building population got classified for $d = 1/5$. Is it worthwhile to attempt the classification in such cases? It depends on the application, for example differential costs of two types of misclassification, and subjective considerations including quality of life influenced by the treatment, therapy expense and expected survival time.

Both the analysis of five fold cross validation and multiple cross validation showed the uncertainty of gene selection results based on different subsets of the data. The large number of variables that appeared only in a small number of runs suggested noises in the data and the difficulty caused by limited training sample size in the high dimensional scenario. It could also suggest the conundrum that the "true" model consists of a large number of variables with modest effects of which different subsets gives rise to roughly equal prediction ability. Options for further study in this and other difficult problems include allowing the DCOV stopping criteria to be modified by some amount ϵ, and allowing the greedy variable selection algorithm to be doubly greedy by testing the next best m of the remaining variables rather than just the next variable. It remains to obtain theoretical results to guide exploration in alternate scenarios.

The analysis of random permutation on the response served as both a validation of our results and a discussion of what one is likely to obtain without any true signal. If someone started with an entirely different data set having the same proportions for the two classes with that in the original data but no real signal at all, as what one might get from scrambling, and went through every step, and finally obtained a subgroup of patients with large voting score values, the result was no better than just guessing that everyone was sensitive. This experiment was also a cautionary tale that if one had not held out validation sets, the analyst could be easily fooled by spurious correlated variables and perfect training accuracy. Our proposed multiple cross validation and analysis through the voting scores provided protection against finding fake signals.

3 BACKWARD MULTIPLE IMPUTATION ESTIMATION OF THE CONDITIONAL LIFETIME EXPECTANCY FUNCTION WITH APPLICATION TO CENSORED HUMAN LONGEVITY DATA

3.1 Introduction

Survival analysis has been focusing on the popular hazard function for decades, and one of most famous models is the Cox's proportional hazard model, Cox (1972). However, the hazard function, defined as the risk of immediate failure, can be conceptually difficult to understand. The expected or remaining lifetimes are intuitively more attractive because of the easy interpretation and turns out to be a more relevant metric under many circumstances. For example, it is more transparent to patients if the doctor explains it as "on average, the lifetime is expected to be 80 years if one also at 70 with similar demographic and healthy background like you takes this treatment" rather than in the language of "the average hazard is expected to decrease by 25% among the treated patients similar to you." Furthermore, in the analysis of reliability and actuarial data, a life insurance company may care more about the life expectancy of a person, and an engineering firm might want to know the expected remaining lifetime of a system given survival past certain time. These motivates us to put more attention to directly estimate key summary measures regarding to remaining lifetimes. This paper targets at lifetime expectancy function and the mean residual life function.

The lifetime expectancy function(LEF) of a survival time T (with $T > 0$), denoted by $e(t)$, is defined as

$$e(t) = E(T|T > t) = t + \int_t^{\tau_T} \frac{S(u)}{S(t)} du,$$

where $S(t) = P(T > t)$ is the survival function and $\tau_T = \inf\{t : S(t) = 0\}$. Denote $m(t)$ the mean residual life function(MRLF) which is the expected remaining lifetime given survival up to time t and

$$m(t) = e(t) - t = E(T - t|T > t).$$

Both $e(t)$ and $m(t)$ uniquely determine $S(t)$ as the following equation showsHall and Wellner.

$$S(t) = \frac{e(0)}{e(t) - t} \exp\left\{-\int_0^t [e(u) - u]^{-1} du\right\} = \frac{m(0)}{m(t)} \exp\left\{-\int_0^t m(u)^{-1} du\right\}.$$

Hall and Wellner provides necessary and sufficient conditions, such that $m(t)$ is a proper

MRLF (or that $e(t)$ is a proper LEF). That is, $F(t) = P(T \leqslant t)$ is a proper continuous distribution function if and only if $m(t)$ satisfies:

1. $m(t) \geqslant 0$ for all $t \geqslant 0$;
2. $e(t) = m(t) + t$ is nondecreasing in t;
3. if there exists a τ such that $m(\tau) = 0$ then $m(t) = 0$ for all $t \geqslant \tau$, otherwise, $\int_0^\infty m(t)^{-1} dt = \infty$;
4. $m(t)$ is a right continuous function and has a left limit with positive increments at discontinuities.

In practice, real data always contain additional information besides the survival time itself and researchers are interested in how the variables contribute to lifetimes. This is when the conditionality of LEF or MRLF plays a role. For example, in the context of mobile devices, modeling the conditional LEF that the users keep active with certain Apps or games after installation helps the providers target and stratify their customers, and offers insights about the effectiveness of different features related to the product. In the situation of property purchase, it is of the interest to both sell and buy sides to know how long it takes the house to be sold after being listed for sale by a certain agent or on a real estate website considering the size, building year, location and estimated price of the houses. Moreover, as we will depict in our real data analysis, lifestyle factors such as smoking and socioeconomic status, disease and healthy metrics are all informative towards one's expected lifespan.

In this paper, we propose a framework for estimating the conditional LEF $e(t|x) = E(T|T > t, X = x)$ when covariates X information is available and the survival times are subject to right censoring. Following the same idea with the Buckley-James estimator in Buckley and James (1979) to address censoring by imputation, our method replaces the censored survival times in backward order with a heuristic guess of a fitted LEF using a user-specific base model and the covariates. One is then able to model LEF with a completely imputed dataset. We provide variance estimation and confidence interval for the estimated LEF based on the idea of multiple imputation in Rubin (2004). When there is no covariate, our estimator is proven to be the same as the one derived by inverting the Kaplan-Meier estimator for the survival function, Kaplan and Meier (1958). Considerable research has been done on estimation of the conditional MRLF. Chen and Cheng (2006); Sun and Zhang (2009) and Oakes and Dasu (1990) discussed different semiparametric conditional MRLF estimations. McLain and Ghosh (2011) covered nonparametric estimation for MRLF with covariates and we show that this method is equivalent to our framework by choosing kernel

regression as the base model. We investigate the behaviors of our proposed estimator in practical settings via different simulation studies. Finally, we demonstrate our method to model human lifetimes with the Beaver Dam Eye Study data, Klein et al. (1991), where survival information and a number of useful variables, from demographic records to medical measurements, are included.

3.2 Semiparametric and nonparametric estimation of conditional MRL function

There are several papers in the literature discussing how to estimate MRL function $m(t|x)$ with right censoring conditional on $x = (x_1, \cdots, x_p)^T$ which is the $p-$dimentional vector of explanatory variables. It is easy to obtain the corresponding lifetime expectancy function $e(t|x)$ by $t + m(t|x)$. First, Oakes and Dasu (1990) considered the semiparametric proportional MRL model

$$m^p(t|x) = m_0^p(t) \exp(\beta^T x),$$

where $m_0^p(t)$ is a baseline MRL function and β is a $p-$dimentional vector of regression coefficients. Chen and Cheng (2006) proposed to estimate $m(t|x)$ as an additive expectancy regression model. The model takes a semiparametric form of

$$m^a(t|x) = m_0^a(t) + \gamma^T x,$$

where $m_0^a(t)$ is a baseline MRL function and γ is a $p-$dimentional vector of regression coefficients. Sun and Zhang (2009) framed the general family of semiparametric transformation models

$$m^g(t|x) = g\{m_0(t) + \beta^T x\}$$

that include the previous proportional and additive models as special cases.

As discussed in McLain and Ghosh (2011), the nondecreasing property of $e(t|x)$ may be violated under the existing semiparametric models. The authors in McLain and Ghosh (2011) considered taking a different perspective to satisfy this natural constraint. They first calculate the nonparametric estimation $\hat{S}_P(t|x)$ of the conditional survival function using generalized Kaplan-Meier estimator according to Dabrowska et al. (1989); Gonzalez-Manteiga and Cadarso-Suarez (1994), and then take the inversion to obtain the nonparametric estimator $\hat{m}_P(t|x)$ for the conditional MRL function. A smoothed estimation of MRL is available by inverting the smoothed $\hat{S}_P(t|x)$ based on Bernstein polynomials. It is straightforward that $\hat{m}_P(t|x)$ is a valid MRL function since $\hat{S}_P(t|x)$ is a well defined survival

function.

3.3 Backward multiple imputation framework for estimating LEF

Backward imputation without covariates

Idea and method

Let's first consider the cases without any covariate to intuitively understand the idea. Let T be a continuous nonnegative random variable and C be the censoring variable. We assume that T and C are independent. The observed data set consists of n independent and identically distributed replicates of $(Y_i, \delta_i), i = 1, \cdots, n$ where $Y_i = \min(T_i, C_i)$, and $\delta_i = I_{\{T_i \leqslant C_i\}}$ is the censoring indicator. Let $y_{(1)} < \cdots < y_{(M)}$ be the distinct ordered values of the n observations and $n_1, \cdots, n_M$ be the corresponding number of observations taking each specific values of $y_{(i)}, i = 1, \cdots, M$. Denote $t_{(1)} < \cdots < t_{(K)}$ and $c_{(1)} < \cdots < c_{(J)}$ the distinct ordered event times and censored times respectively. The notation $n(t_{(k)})$ or $n(c_{(j)})$ takes the number of observations at $t_{(k)}$ or $c_{(j)}$.

The $c_{(j)}$'s are right censored and we know that the true values should be greater than the censored times $c_{(j)}$. One reasonable guess for the true values is the lifetime expectancy at $c_{(j)}$, i.e. $e(c_{(j)}) = E(T|T > c_{(j)})$. This is the same idea as the single imputation of Little and Rubin, Little and Rubin (2014) and as the Buckley-James estimator, Buckley and James (1979). We could use the sample lifetime expectancy, which is the mean of the observations greater than $c_{(j)}$, as an estimate for $e(c_{(j)})$. However, this does not work if there still exists censored data to the right of the targeted $c_{(j)}$. We can address this problem by processing our guessing regime for $c_{(j)}$'s backwardly from J to 1. After imputing the censored values, it is easy to obtain sample lifetime expectancy at any time point t. The detailed steps are as follows:

Algorithm 1. ***Backward imputation without covariates***

1. *We do nothing if $c_{(J)}$ is the largest value in the dataset, i.e. $c_{(J)} = y_{(M)}$. Otherwise, we estimate $e(c_{(J)})$ by the sample mean of the observations beyond $c_{(J)}$, i.e.*

$$\hat{e}_B(c_{(J)}) = \frac{\sum_{i=1}^{n} y_i I_{\{y_i > c_{(J)}\}}}{\sum_{i=1}^{n} I_{\{y_i > c_{(J)}\}}} = \frac{\sum_{i=1}^{M} y_{(i)} n_i I_{\{y_{(i)} > c_{(J)}\}}}{\sum_{i=1}^{M} n_i I_{\{y_{(i)} > c_{(J)}\}}} = \frac{\sum_{k=1}^{K} t_{(k)} n(t_{(k)}) I_{\{t_{(k)} > c_{(J)}\}}}{\sum_{k=1}^{K} n(t_{(k)}) I_{\{t_{(k)} > c_{(J)}\}}}.$$

   Replace $c_{(J)}$ by $\hat{e}_B(c_{(J)})$ and treat it as observed.

2. *Repeat the above procedure backwardly for $j = J-1, \cdots, 1$ to replace $c_{(j)}$ by $\hat{e}_B(c_{(j)})$ which is the sample mean of the observations beyond $c_{(j)}$ in the imputed data. Since the process runs for j from J to 1, we will have imputed all the censored values greater than $c_{(j)}$ and there is no missingness to estimate $e(c_{(j)})$ by the sample mean of the observations larger than $c_{(j)}$.*

3. *Let $\tilde{y}_1, \cdots, \tilde{y}_n$ be the data after backward imputation procedure. If y_i is observed or it is the largest observation and is censored, then $\tilde{y}_i = y_i$. Otherwise, y_i is one of the censored time and $\tilde{y}_i = \hat{e}_B(y_i)$. The backward procedure only obtains estimates of $e(t)$ at the censored times. In general, we estimate $e(t)$ for $t \geqslant 0$ by the following formula*

$$\hat{e}_B(t) = \frac{\sum_{i=1}^{n} \tilde{y}_i I\{y_i > t\}}{\sum_{i=1}^{n} I\{y_i > t\}}.$$

Relationship with Kaplan-Meier estimator

Another way to obtain an estimator for $e(t)$ is by inverting an estimator for $S(t)$. We know that Kaplan-Meier estimator $\hat{S}_{KM}(t)$ is the MLE for $S(t)$ w.r.t the empirical likelihood. Denote $\hat{e}_{KM}(t)$ the estimate for $e(t)$ by inverting $\hat{S}_{KM}(t)$. The following theorem proves the equivalence between $\hat{e}_B(t)$ and $\hat{e}_{KM}(t)$. This also demonstrates the equivalence between the spirit of backward imputation and the idea of redistribution-to-the-right to estimate survival function established by Efron (1967).

Theorem 3.1. *Let T be a continuous nonnegative random variable which is independent of the censoring variable C. We observe n i.i.d replicates of $(Y_i, \delta_i), i = 1, \cdots, n$ where $Y_i = \min(T_i, C_i)$, and $\delta_i = I_{\{T_i \leqslant C_i\}}$. Denote $\hat{e}_B(t)$ the backward imputation estimator for $e(t)$ as described in Algorithm 1 and $\hat{e}_{KM}(t)$ the inverted Kaplan-Meier estimator for $e(t)$ which takes the following explicit form:*

$$\hat{e}_{KM}(t) = \begin{cases} t_{(k)} + \dfrac{1}{\hat{S}_{KM}(t_{(k-1)})} \displaystyle\sum_{l=k+1}^{K} (t_{(l)} - t_{(l-1)}) \hat{S}_{KM}(t_{(l-1)}), & t_{(k-1)} < t < t_{(k)} \\ t_{(k)} + \dfrac{1}{\hat{S}_{KM}(t_{(k)})} \displaystyle\sum_{l=k+1}^{K} (t_{(l)} - t_{(l-1)}) \hat{S}_{KM}(t_{(l-1)}), & t = t_{(k)}, k = 1, \cdots, K-1 \\ 0, & t \geqslant t_{(K)}. \end{cases}$$

Then $\hat{e}_B(t) = \hat{e}_{KM}(t)$ for $t \geqslant 0$.

Backward imputation with covariates

We want to make use of the covariates information for estimating life expectancy function. We assume the censoring to be conditionally independently of of the survival time given

the covariates $X = x$. Now our observations are n i.i.d samples $(Y_i, \delta_i, x_i), i = 1, \cdots, n$, where $Y_i = \min(T_i, C_i)$, and $\delta_i = I_{\{T_i \leqslant C_i\}}$. Suppose we have a base regression model

$$f(x) = E(T|X = x)$$

that uses the covariates information to predict the mean survival times when there is no censoring. We substitute the sample mean in the previous backward imputation procedure by the base regression model. This means that we treat the estimate for $e(c_{(j)}|x) = E(T|T > c_{(j)}, X = x)$ as our guess for the censored case $c_{(j)}$ with its covariates x. The following algorithm illustrates the detailed steps.

Algorithm 2. ***Backward imputation with covariates***

1. *We do nothing if $c_{(J)}$ is the largest response value in the dataset, i.e. $c_{(J)} = y_{(M)}$. Otherwise, we obtain the fitted model $\hat{f}$ using the observations $\{(y_i, x_i)|y_i > c_{(J)}\}$. Note that all the observations with $y_i > c_{(J)}$ should be uncensored in this step by the definition of $c_{(J)}$. Replace $c_{(J)}$ by $\hat{f}(x_0)$ where x_0 represent the observed covariates values for $c_{J)}$, and treat it as observed.*

2. *Repeat the above procedure backwardly for $j = J-1, \cdots, 1$ with the imputed data.*

3. *Let $\tilde{y}_1, \cdots, \tilde{y}_n$ be the data after backward imputation procedure. Obtain the fitted base model $\hat{f}$ using the data $\{(\tilde{y}_i, x_i)|y_i > t\}$ and we estimate $e(t|x)$ by $\hat{f}(x)$.*

There are several advantages of this framework. One is that it allows time-varying effects of the covariates since we obtain the fitted $e(t|x)$ restricting to the subset of the data with the original censored survival time greater than the time point t. Another flexibility about this procedure is the freedom to choose the base model f that describes the data the best. The following result states that using kernel regression as the base model in backward imputation procedure is equivalent to the nonparametric estimator $\hat{e}_P(t|x)$ proposed by McLain and Ghosh, McLain and Ghosh (2011). This also implies that $\hat{e}_P(t|x)$ shares the similar pros and cons to kernel regression. For example, one has to take care of the choice of kernel, the contamination in the distance due to irrelevant variables and curse of dimensionality. One is able to address these issues by applying more appropriate base models in backward imputation method to accommodate different datasets.

Theorem 3.2. *Let $K : \mathbb{R}^p \to \mathbb{R}$ be the p-dimensional kernel function and h_n denotes the bandwidth. Let $\hat{e}_B(t|x)$ be the estimator for $e(t|x)$ from backward imputation using kernel regression with K and h_n, and $\hat{e}_P(t|x)$ be the nonparametric estimator of the conditional LEF proposed in McLain and Ghosh (2011) using the same K and h_n. Then $\hat{e}_B(t|x) = \hat{e}_P(t|x)$ for $t \geqslant 0$ given x.*

Variance estimation with multiple imputation

The methods illustrated above are in the fashion of single imputation, which does not take into account the uncertainty about the predictions of the unknown censored values. It is likely that the variance estimation for $\hat{e}_B(t|x)$ is biased toward zero. We incorporate the idea of multiple imputation procedure, Rubin (2004), in our proposed method. Instead of filling in the conditional expected values for each censored value as described above, we replace by a random sample drawn from the posterior predictive distribution under the base model each time. It introduces randomness that represent the uncertainty about the right value to impute. We repeat the backward multiple imputation for a number of times and the results are combined finally to obtain a valid variance estimation and confidence interval for the estimate of conditional life expectancy function. The procedures are shown below.

Algorithm 3. ***Backward multiple imputation with covariates***

1. *Set up the number of multiple imputation m. For each replication, repeat 2 - 4.*

2. *We do nothing if $c_{(J)}$ is the largest response value in the dataset, i.e. $c_{(J)} = y_{(M)}$. Otherwise, we obtain the fitted model $\hat{f}$ using the observations $\{(y_i, x_i)|y_i > c_{(J)}\}$. Note that all the observations with $y_i > c_{(J)}$ should be uncensored in this step by the definition of $c_{(J)}$. Replace $c_{(J)}$ by a random sample from the posterior predictive distribution of the fitted model at x_0 where x_0 represent the observed covariates values for $c_{(J)}$, and treat it as observed.*

3. *Repeat the above procedure backwardly for $j = J-1, \cdots, 1$ with the imputed data.*

4. *Let $\tilde{y}_1, \cdots, \tilde{y}_n$ be the data after backward imputation procedure. Obtain the fitted base model $\hat{f}$ using the data $\{(\tilde{y}_i, x_i)|y_i > t\}$ and we estimate $e(t|x)$ by $\hat{f}(x)$. Moreover, keep record of the estimated variance for $\hat{f}(x)$.*

5. *With m imputations, one collects m different sets of the point and variance estimates for $e(t|x)$. Let $\hat{Q}_i$ and $\hat{U}_i$ be the point and variance estimates of $e(t|x)$ from the ith imputed data set, $i = 1, \cdots, m$. Note that $\hat{Q}_i$ and $\hat{U}_i$ are functions of x and we eliminates the dependency on x in the notation for simplicity.*

6. *The point estimate for $e(t|x)$ from multiple imputations is*

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i.$$

7. *Let $\bar{U}$ be the within-imputation variance, i.e.*

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} \hat{U}_i$$

and B be the between-imputation variance

$$B = \frac{1}{m-1}\sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2.$$

The the variance estimation for the estimated $e(t|x)$ is the total variance

$$T = \bar{U} + (1 + \frac{1}{m})B.$$

8. *The statistic $(Q - \bar{Q})T^{-1/2}$ is approximately distributed as a t-distribution with degrees of freedom*

$$v_m = (m-1)[1 + \frac{\bar{U}}{(1+m^{-1})B}]^2.$$

When v_m is large, one may approximate by a normal distribution. Confidence interval for $e(t|x)$ can be derived accordingly.

3.4 Simulation

In this section, we present results from three simulation studies that were run to evaluate the performance of the backward multiple imputation method. There were three settings to generate the data. The first corresponds to an additive model with $e(t|x) = t + \exp(-t) + \beta_1 x_1 + \beta_2 x_2$. The second is a proportional model with $e(t|x) = t + \exp(-t + \beta_1 x_1 + \beta_2 x_2)$. The third is a hybrid model with $e(t|x) = t + \beta_1 x_1 + \beta_2 x_2 + \exp(-t + \beta_3 x_3 + \beta_4 x_4)$. We generated the censoring variable C from exponential distribution with rate parameter that resulted in 30% right censoring rate. The sample size was $n = 300$ for the first two settings and was 400 for the third one. The covariates are mutually independent. x_1 and x_3 were drawn from Bernoulli(1/2) and x_2 and x_4 were drawn from Uniform(0, 2) with $\beta_1 = 1, \beta_2 = 0.5, \beta_3 = -1$ and $\beta_4 = -0.5$. The base model was linear regression for simulation 1 and was SS-ANOVA for the remaining two settings. We assumed the errors followed a normal distribution for all three cases. This means our posterior predictive distribution from which we drew samples to impute the censored cases is also a normal distribution. The multiple imputation number was set to be 20 and we repeated for 300 times

	target point	$Q_{0.1}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.9}$
Additive	mean bias	-0.0095	-0.0160	-0.0167	-0.0257	-0.0423
	mean std.	0.1656	0.1770	0.2194	0.3346	0.6199
	coverage prob	0.946	0.948	0.948	0.950	0.950
Proportional	mean bias	0.0061	0.0030	-0.0047	-0.0181	-0.0235
	mean std.	0.0346	0.0369	0.0430	0.0566	0.0813
	coverage prob	0.973	0.963	0.983	0.970	0.987
Hybrid	mean bias	0.0207	0.0122	0.0047	-0.0225	-0.0578
	mean std.	0.1893	0.2057	0.2489	0.3490	0.5504
	coverage prob	0.963	0.963	0.943	0.947	0.940

Table 3.1: Summary of results for estimated life expectancy function using backward imputation method with three different settings.

for all three settings. The life expectancy functions were estimated for 0.1, 0.25, 0.5, 0.75 and 0.9 quantiles of the censored survival time Y given all covariates values fixed at 1. We examined the estimated bias, standard deviation and empirical coverage of 95% confidence intervals. (We used normal approximation since v_m's calculated according to algorithm 3 are large.) *Table 3.1* summarized the results and it showed that the backward multiple imputation framework with tailored base model performed pretty well for estimating $e(t|x)$. Moreover, it also gave the desired coverages for the true $e(t|x)$ using the 95% normal confidence interval with the variance estimation based on the multiple imputation idea.

3.5 Application to Beaver Dam Eye Study data

Data description

The Beaver Dam Eye Study (BDES), Klein et al. (1991), is an ongoing population-based study of age-related ocular disorders with five, ten, fifteen and twenty year follow-ups. Subjects at baseline, examined between 1988 and 1990, were a group of 4926 people aged 43-86 years from Beaver Dam, WI. The survival statuses, including ages at death, for this population were updated by 12/31/2013 with 2014 individuals who were alive. BDES provides us an excellent opportunity to study the lifetime expectancy with our proposed methods.

A number of variables, including measurements on individual health and lifestyles, were recorded in the study. We took advantages of a couple of the most important ones which were used in Kong et al. (2012) to examine the association with human mortality. To maintain the largest sample sizes, we focused on the baseline data. *Table 3.2* lists the

variable	units	description
lastage	years	censored age at death
survflag	yes/no	survival indicator
baseage	years	age at baseline
gender	F/M	gender
edu	years	highest year school/college completed
bmi	kg/m^2	body mass index
smoke	yes/no	history of smoking
inc	yes/no	household personal income $>$ 20K
diabetes	yes/no	history of diabetes
cancer	yes/no	history of cancer
heart	yes/no	history of cardiovascular disease
kidney	yes/no	history of chronic kidney disease

Table 3.2: Variable description in the SS-ANOVA model

description of all the variables involved in the study. A number of variables with weak signals for longevity are discussed in the supplementary material.

Model fitting

SS-ANOVA modelWahba (1990); Gu (2013); Wang (2011) has a successful history in modeling BDES dataLu et al. (2005); Kong et al. (2012). Our base model is an SS-ANOVA model with the following form.

$$\begin{aligned}(\mathit{imputed})\ \mathit{lastage} = {} & \mu + f_1(\mathit{baseage}) + \beta_{\mathit{gender}} I_{\{\mathit{gender}=F\}} + \\ & f_2(\mathit{edu}) + f_{12}(\mathit{baseage} : \mathit{edu}) + f_3(\mathit{bmi}) + \\ & \beta_{\mathit{smoke}} I_{\{\mathit{smoke}=no\}} + \beta_{\mathit{inc}} I_{\{\mathit{inc}>20K\}} + \\ & \beta_{\mathit{diabetes}} I_{\{\mathit{diabetes}=no\}} + \beta_{\mathit{cancer}} I_{\{\mathit{cancer}=no\}} + \\ & \beta_{\mathit{heart}} I_{\{\mathit{heart}=no\}} + \beta_{\mathit{kidney}} I_{\{\mathit{kidney}=no\}}\end{aligned} \qquad (*)$$

Functions f_1, f_2 and f_3 are cubic splines and f_{12} uses the tensor product construction. The remaining covariates are unpenalized and modeled as linear terms with $I_{\{\cdot\}}$ as indicator functions. We incorporated this base model in *Algorithm 3* with multiple imputation replications $m = 200$ to estimate the conditional lifetime expectancy function in the population

of BDES. One adjustment we applied for *Algorithm 3* was that we used model $(*)$ if the sample size involved in step 2 of *Algorithm 3* was greater than 100 otherwise we simply used sample mean of ages of death among all the samples involved in this step. We observed that both the estimations for LEF and the variance estimation became stable after 20 multiple imputations.

Results for cohort of baseline age 70 in BDES

Figure 3.1-3.4 display the predicted conditional lifetime expectancy functions for the cohort with baseline age of 70. In *Figure 3.1*, we present how the expected lifetime changes with BMI, education and gender for nonsmoking rich and healthy individuals with baseage of 70. The natural constraint of monotonic nondecreasing over t for $e(t|x)$ is very well satisfied. The plots suggest that females tend to have longer lifespans compared with males. Higher education and mid-valued BMI are protective for longevity. The covariates effects fade out as t gets large with several possible reasons. First, the sample size is limited when restricting to subjects over 85. Second, it is likely that those long-lived individuals have survived from the hist risk factors so that we could not find the significance for the covariates.

In *Figure 3.2*, we examine the effects of smoking, cardiovascular disease and income for the subgroup of females with mid-valued BMI, education and no other disease. From the plots, it appears that smoking and having a history of heart disease have negative influences on longevity in this population. Higher household income slightly protects longevity. *Figure 3.3* discovers how diabetes and chronic kidney disease change expected survival given the rest of covariates. It turns out that diabetes is a strong risk factor that reduces human longevity. Chronic kidney disease, though not as harmful as diabetes, also exerts a negative effects on survival times among this subgroup of people.

Figure 3.4 takes a different perspective from the previous 3 plots and focuses on the two continuous variables BMI and education for a cohort of rich and healthy female nonsmokers who entered the study when they were 70. The five surfaces correspond to 5 time points, $t = 70, 75, 80, 85$ and 90. Each surface represent the estimated expected lifetime across different values of BMI and education. When t is small, we observe the quadratic influence of BMI where very low BMI values are very harmful and the optimal value happens at around 26 or 27 and tails down slowly for higher values. Note that Beaver Dam is a small town in the Midwest and may not be representative of some population groups in other areas of the country. The education displays a monotonic increasing effect on lifetime in this cohort. The higher the completed education is, the longer the expected lifetime

is. When t gets large, we again find that the influences of BMI and education disappear. More results about some other weakly related variables and other baseline age cohorts are discussed in supplementary material.

Results for cohort of baseline age 50 in BDES

We examined the fitted LEF for the group with baseline age of 70 in BDES data previously. Let's further look at another cohort with baseline age of 50, which contains around 80% of censoring compared to a censoring rate of about 20% for the cohort of baseline age at 70. *Figure 3.5* presents the effects of BMI, education and gender for the subgroup of rich and healthy nonsmokers of baseline age 50. In comparison with *Figure 3.1*, it is obvious that the confidence intervals from the multiple imputation method are much wider, resulting in the insignificance of gender although the estimated LEF for males and females are still apart given the rest of the covariates. In addition, the percentage of observed death ages with education greater than 20 is about 10% lower than that the average in this cohort, which further explains why the widths of the confidence intervals correspond to education of 20 are much more inflated in *Figure 3.5*. Similar behaviors are also observed in *Figure 3.6* which explores the effects of income, smoking and heart disease. The heavy censoring rate among subjects with baseline age of 50 gives rise to the statistical insignificance of these important covariates.

Additional variables in BDES

We have discussed the most important variables associated with longevity in BDES. There are a couple of medical measurements weakly correlated with the survival times in additional to the ones already in the model $(*)$. The variables we further took into accounts are listed in *Table 3.4*, and the new SS-ANOVA model are shown in $(**)$. Functions f_1 to f_8 are cubic splines, and f_{12} and f_{17} use the tensor product construction.

$$
\begin{aligned}
(\mathit{imputed})\ \mathit{lastage} = {} & \mu + f_1(\mathit{baseage}) + \beta_{\mathit{gender}} I_{\{\mathit{gender}=F\}} + \\
& f_2(\mathit{edu}) + f_{12}(\mathit{baseage} : \mathit{edu}) + f_3(\mathit{bmi}) + \\
& \beta_{\mathit{smoke}} I_{\{\mathit{smoke}=no\}} + \beta_{\mathit{inc}} I_{\{\mathit{inc}>20K\}} + \\
& \beta_{\mathit{diabetes}} I_{\{\mathit{diabetes}=no\}} + \beta_{\mathit{cancer}} I_{\{\mathit{cancer}=no\}} + \\
& \beta_{\mathit{heart}} I_{\{\mathit{heart}=no\}} + \beta_{\mathit{kidney}} I_{\{\mathit{kidney}=no\}} + \\
& f_4(\mathit{hdl}) + f_5(\mathit{hgb}) + I_{\{\mathit{gender}=F\}} f_6(\mathit{hgb}) + \\
& f_7(\mathit{glucose}) + f_{17}(\mathit{baseage} : \mathit{glucose}) + f_8(\mathit{crp}).
\end{aligned}
\qquad (**)
$$

variable	units	description
hdl	mg/dL	high-density lipoprotein cholesterol (serum)
hgb	g/dL	hemoglobin (blood)
glucose	mg/dL	glucose (serum)
crp	mg/L	C-reactive protein

Table 3.3: Additional variables in the SS-ANOVA model

Again, we use the cohorts with baseline age of 70 to demonstrate the results, illustrated in *Figure 3.7-3.8*. In *Figure 3.5*, we observed the positive influence of HDL as well as the harm of high glucose on longevity given all the rest of covariates. Hgb has a quadratic effects when time point t is small with larger sample size, and is interactive with gender as captured in *Figure 3.7*. The optimal hgb range is from 12 to 14 and 14 to 18 for females and males respectively. High level of C-reactive protein is a sign in response to inflammation and turns out to be decrease survival as expected. The changes in the corresponding estimated life expectancy are small when the values of these 4 variables run from the best to the worst scenarios, demonstrating their weak effects in addition to the important variables appeared in $(*)$.

Validation using bootstrapped samples

This is an observational study and the true $e(t|x)$ is unknown. We used bootstrap method to get the empirical distributions of $e(t|x)$ for different values of t and x to check if the results coming from backward multiple imputation matches the mean and standard deviation of the empirical distribution. The following steps cover the bootstrap details.

1. Obtain bootstrap samples by resampling with replacement.
2. Use backward imputation, i.e. *Algorithm 2*, with SSANOVA on the bootstrapped samples.
3. Estimate $e(t|x)$ with the imputed bootstrap data for the combinations of t and x used to generate *Figure 3.1-3.3*.
4. Repeat steps 1-3 for 1000 times to get empirical distribution of $e(t|x)$ for each combination of t and covariates values.

From the above bootstrap procedure, we obtained estimated mean and standard error of $e(t|x)$ from the empirical distributions, denoted as $\hat{e}_{BOOT}(t|x)$ and $\widehat{std}\{\hat{e}_{BOOT}(t|x)\}$. We

quantiles of the ratios	$Q_{0.1}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.9}$
$\frac{\hat{e}_{BM}(t\|x)}{\hat{e}_{BOOT}(t\|x)}$	0.9971	0.9987	1.0006	1.0031	1.0053
$\frac{\widehat{std}\{\hat{e}_{BM}(t\|x)\}}{\widehat{std}\{\hat{e}_{BOOT}(t\|x)\}}$	0.7561	0.8694	0.9897	1.0929	1.1959

Table 3.4: Comparison of the estimates of $e(t|x)$ and its estimated standard deviation by bootstrap and backward multiple imputation.

could compare those with the ones derived from the previous backward multiple imputation, denoted as $\hat{e}_{BM}(t|x)$ and $\widehat{std}\{\hat{e}_{BM}(t|x)\}$. For baseline age of 70, there were 6400 combinations of t and covariates values where t runs from 70 to 93. *Table 3.3* summarizes the differences between the ratios of the two estimators and the ratios of the two corresponding estimated standard deviations. The ratios between the estimators are closely centered at 1. The ratios between the two standard deviations spread out a little bit but still concentrated around 1, meaning that $\hat{e}_{BM}(t|x)$ and $\widehat{std}\{\hat{e}_{BM}(t|x)\}$ match the empirical distribution pretty well. Furthermore, *Figure 3.9* randomly picks 8 cases for the bootstrapped distributions of $\hat{e}_{BOOT}(t|x)$ and displays the histograms. It turns out the bootstrapped distributions are all alike normal distributions. It means that a normal confidence interval derived from the bootstrapped distribution is satisfactory. Therefore, the normal confidence intervals by the multiple imputation procedure are justified since $\hat{e}_{BM}(t|x)$ and $\widehat{std}\{\hat{e}_{BM}(t|x)\}$ are close to $\hat{e}_{BOOT}(t|x)$ and $\widehat{std}\{\hat{e}_{BOOT}(t|x)\}$ as shown in *Table 3.3*.

3.6 Discussion

In this article, we presented our backward multiple imputation framework for estimating the conditional lifetime expectancy function. In the case without covariates, our estimator is proven to be equivalent to the estimation for LEF by inverting the Kaplan-Meier survival function estimator. In the case with covariates, one is free to select a base model that best capture the data. One is able to recover the nonparametric estimator for conditional LEF proposed in McLain and Ghosh (2011) based on the generalized Kaplan-Meier estimator by using kernel regression in our framework. The simulation studies demonstrated the performance of our methods and validated the use of multiple imputation for variance estimation under three different settings. The application to the Beaver Dam Eye Study data illustrated the use of SS-ANOVA model together with our backward multiple imputation method. We presented the fitted results for the cohorts with baseline age of 70 where a

number of variables, including gender, smoking, education level, BMI values and several diseases, were shown to be significantly associated with the human longevity.

There are a couple of issues that we will consider as our future direction. First, as pointed out by McLain and Ghosh (2011), lots of existing models for estimating MRLF or LEF do not satisfy the nondecreasing property of $e(t|x)$. We know that kernel regression ensures the validation of this condition. The real application results in BDES data also seemed to be validated with nondecreasing curves. However, it is of our practical and theoretical interest to explore what base models guarantee this property as well. In this paper, we discussed the use of multiple imputation to obtain the variance estimation for the estimated LEF. Still more research about the other ways to construct the variance estimator under certain base models, including asymptotic distributions of the LEF estimator, is needed in order to reduce the burden in computation.

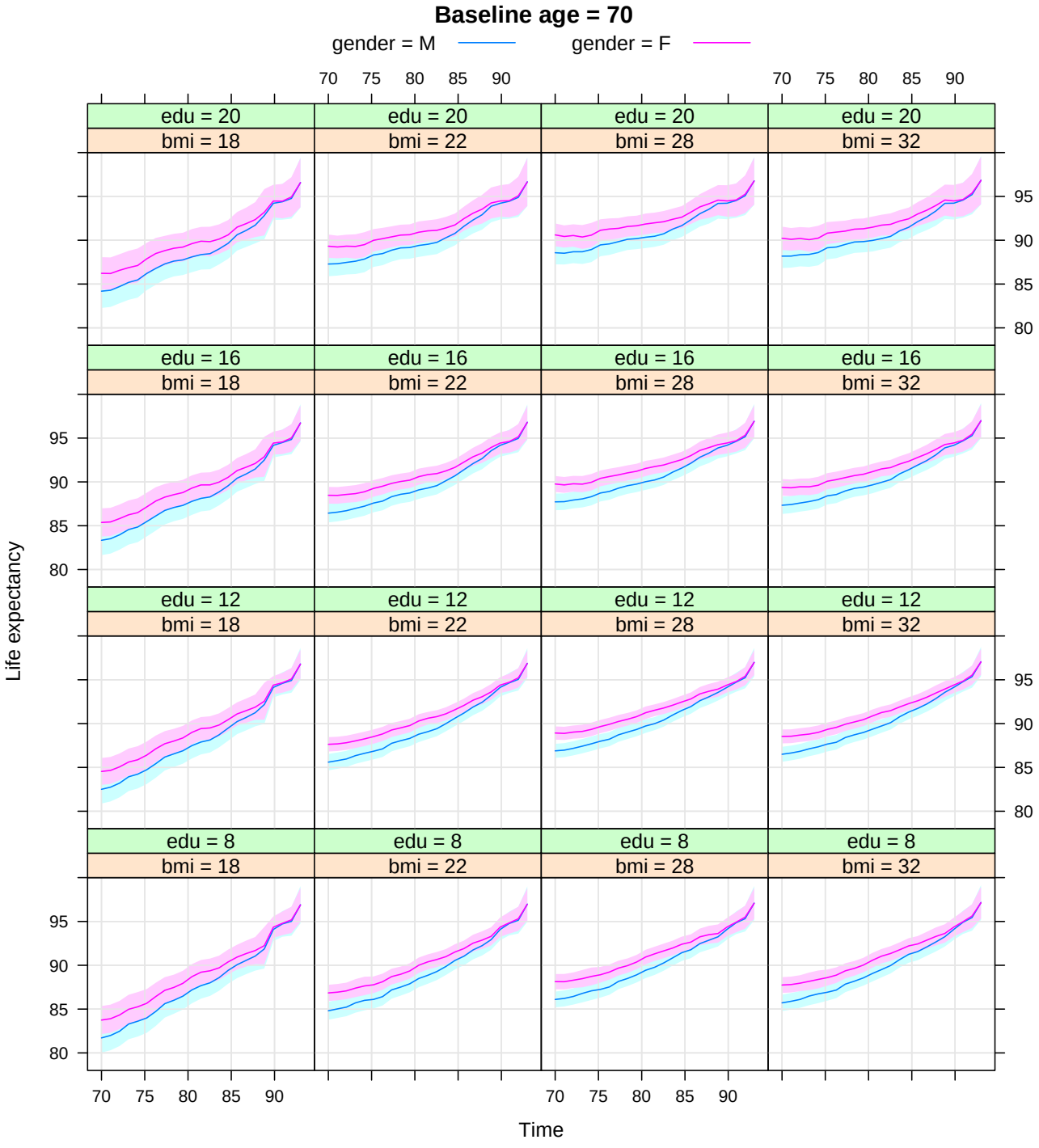

Figure 3.1: Lifetime expectancy function estimation by *bmi*, *edu*, and *gender* for the subgroup with *baseage* = *70*, *smoke* = *no*, *income* $\geqslant$ *20K* and no disease. The x-axis is time t from 70 to 93. The y-axis is $\hat{e}(t|X = x)$. The shaded area presents 95% normal confidence intervals.

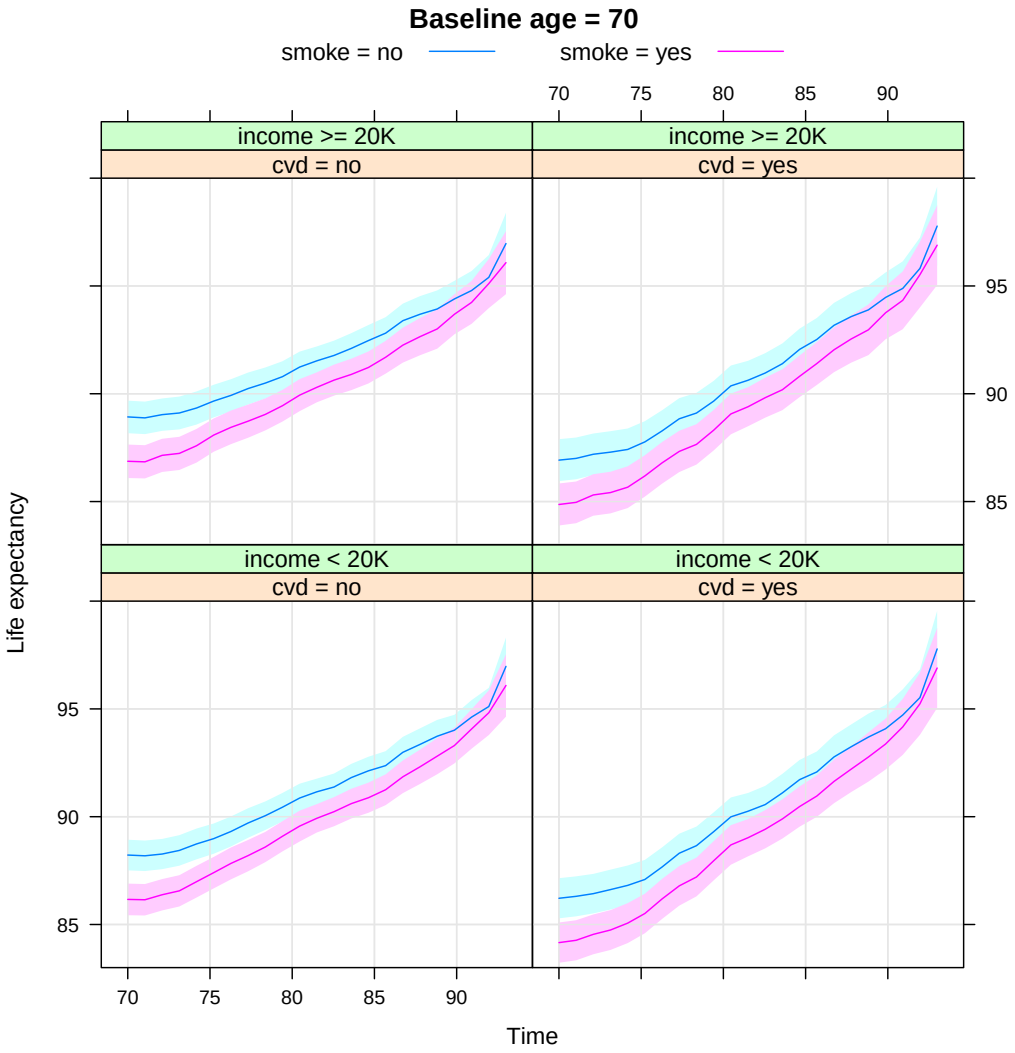

Figure 3.2: Lifetime expectancy function estimation by *smoking*, *heart disease*, and *income* for the group with *baseage = 70, gender = F, bmi = 28(median of the population), edu = 12(median of the population)* and no other disease. The x-axis is time t from 70 to 93. The y-axis is $\hat{e}(t|X = x)$. The shaded area presents 95% normal confidence intervals.

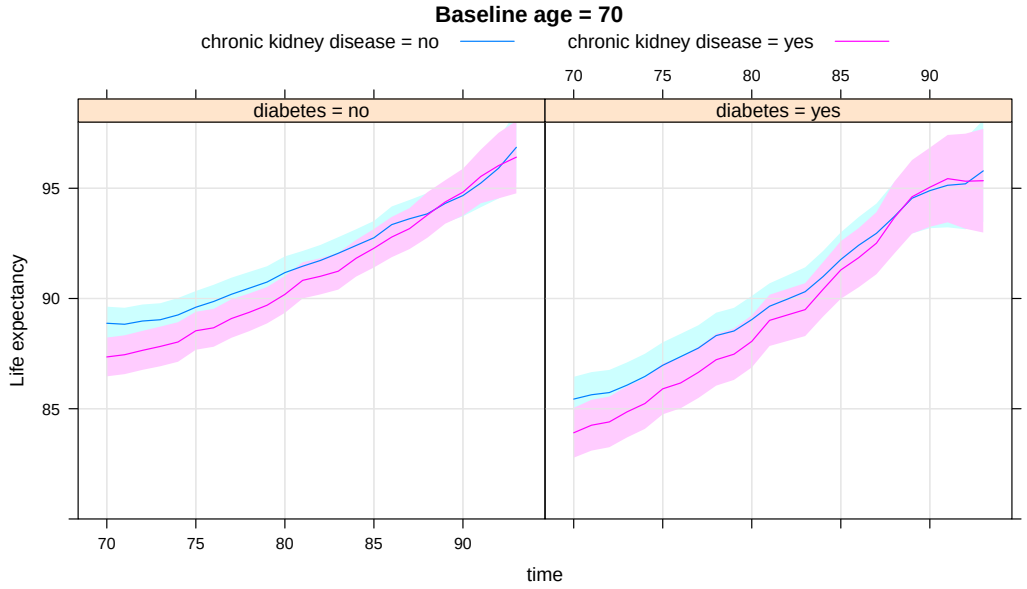

Figure 3.3: Lifetime expectancy function estimation by *diabetes* and *chronic kidney disease* for subjects with *baseage = 70, gender = F, smoke = no, income* $\geqslant$ *20K, bmi = 28, edu = 12* and no heart disease, cancer or stroke. The x-axis is time t from 70 to 93. The y-axis is $\hat{e}(t|X = x)$. The shaded area presents 95% normal confidence intervals.

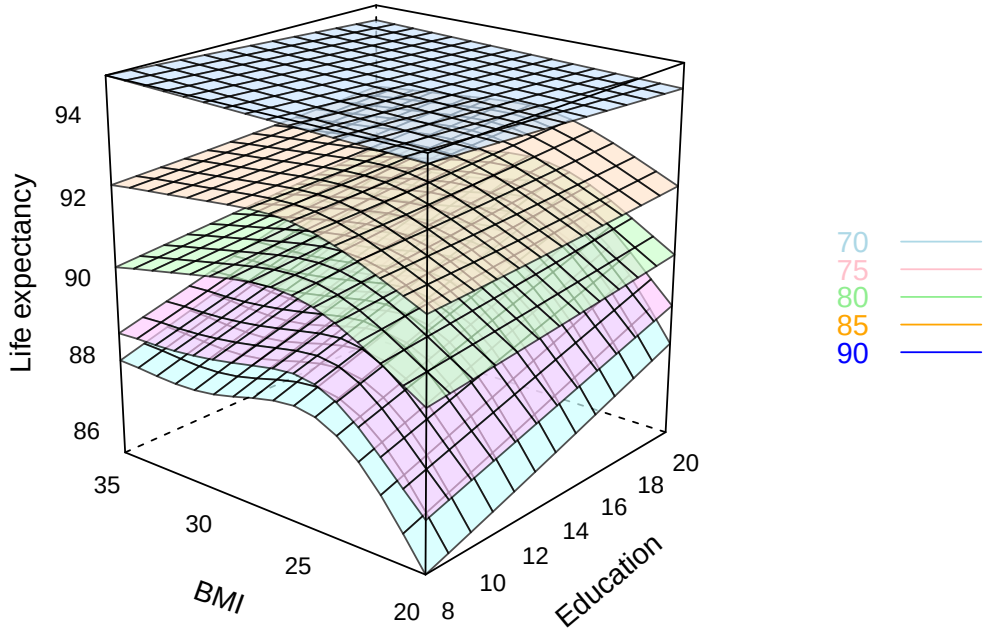

Figure 3.4: *BMI* and *edu* effects on expected lifetime for *baseage* = *70*, *gender* = *F*, *smoke* = *no*, *income* $\geqslant$ *20K* and no disease with $t = 70, 75, 80, 85$ and 90.

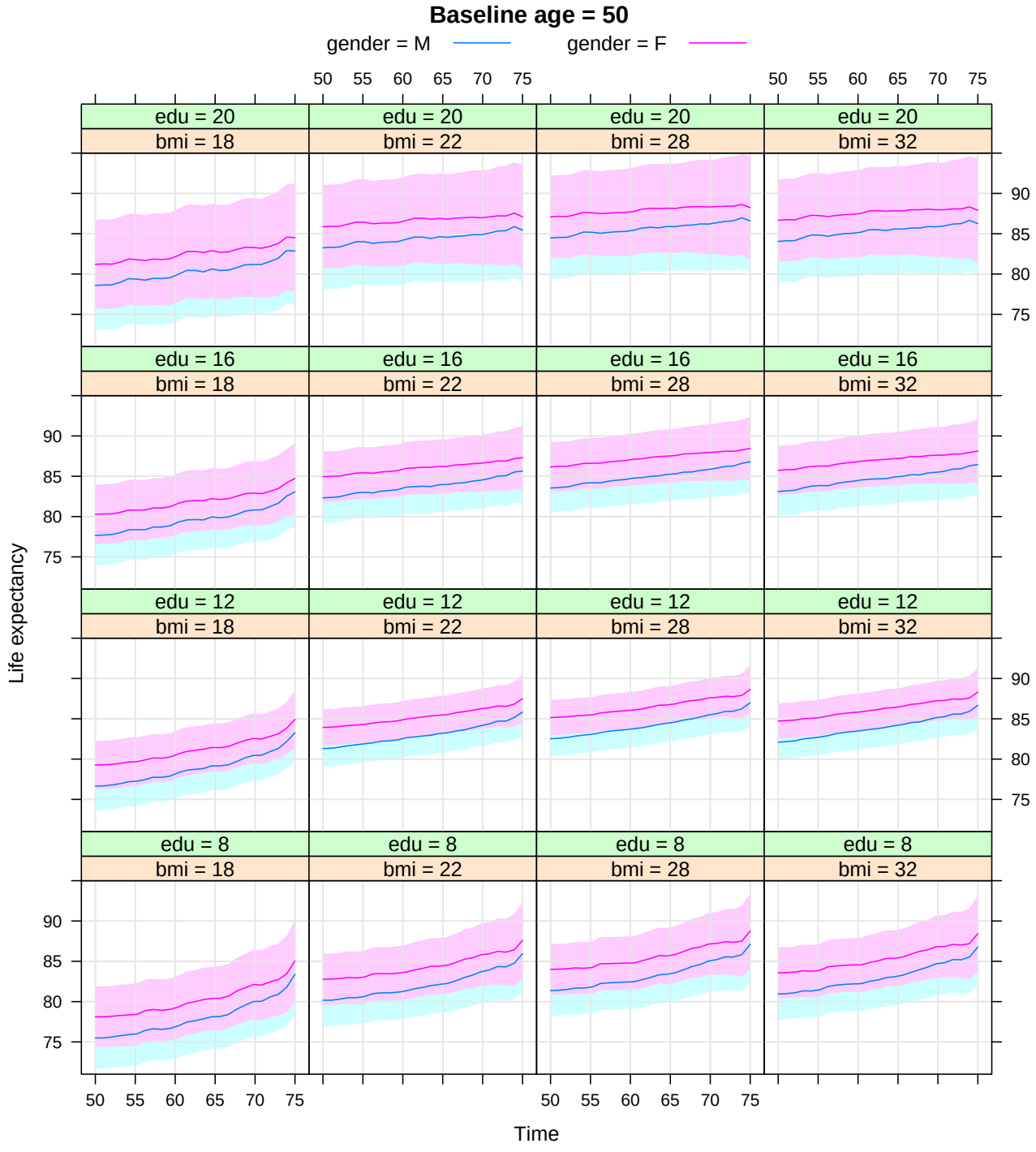

Figure 3.5: Lifetime expectancy function estimation by *bmi*, *edu*, and *gender* for the subgroup with *baseage* = *50*, *smoke* = *no*, *income* $\geqslant$ *20K* and no disease. The x-axis is time t from 50 to 74. The y-axis is $\hat{e}(t|X = x)$. The shaded area presents 95% normal confidence intervals.

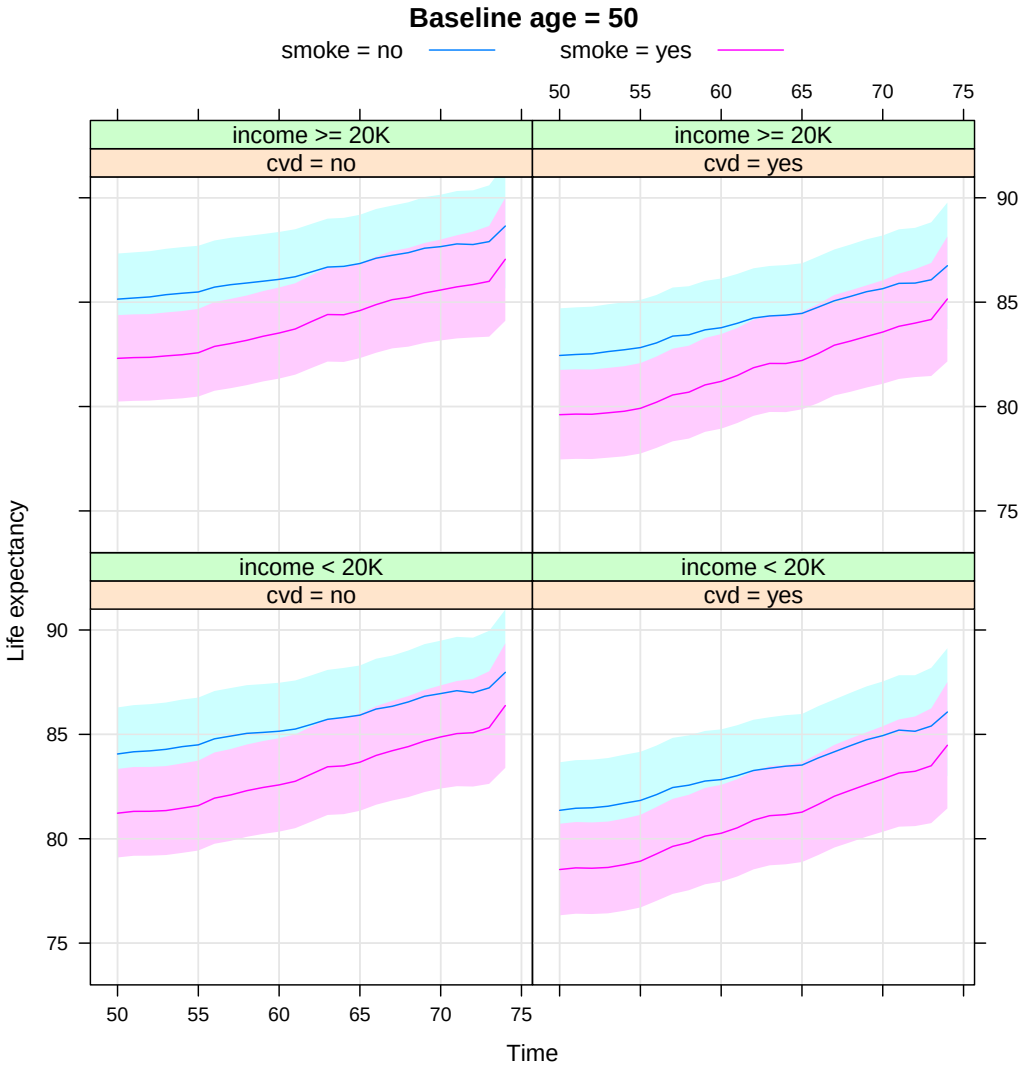

Figure 3.6: Lifetime expectancy function estimation by *smoking*, *heart disease*, and *income* for the group with *baseage = 50, gender = F, bmi = 28(median of the population), edu = 12(median of the population)* and no other disease. The x-axis is time t from 50 to 74. The y-axis is $\hat{e}(t|X = x)$. The shaded area presents 95% normal confidence intervals.

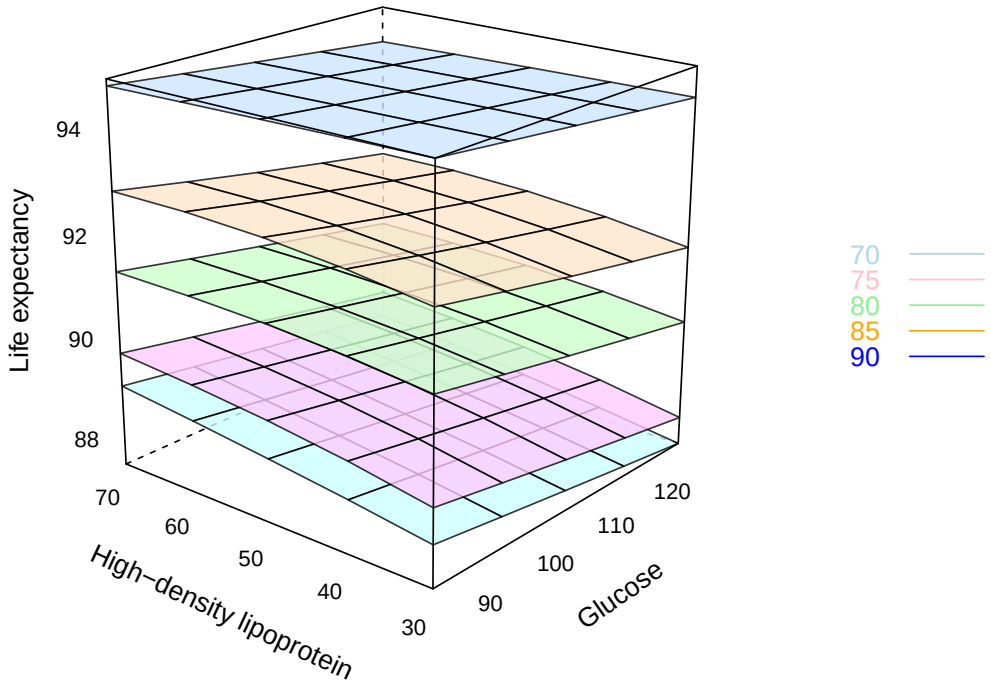

Figure 3.7: *HDL* and *Glucose* effects on expected lifetime for *baseage = 70, gender = F, smoke = no, edu = 12, bmi = 28, income ⩾ 20K, hgb = 14(median), crp = 2(median)* and no disease with $t = 70, 75, 80, 85$ and 90.

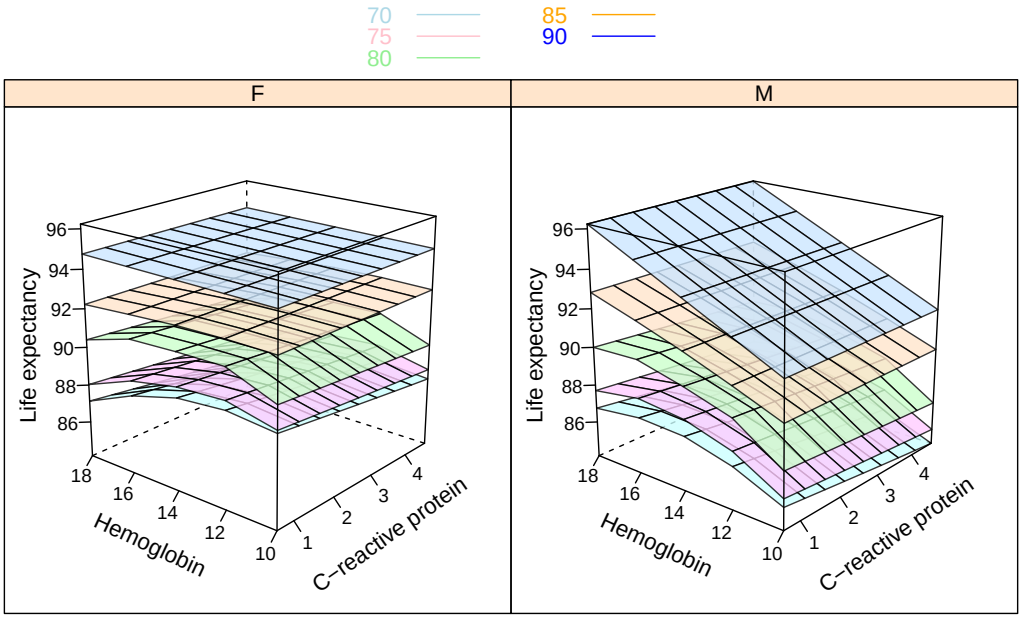

Figure 3.8: *HGB* and *C-reactive protein* effects on expected lifetime for *baseage = 70, gender = F, smoke = no, edu = 12, bmi = 28, income* $\geqslant$ *20K, hdl = 50(median), glucose = 95(median)* and no disease with $t = 70, 75, 80, 85$ and 90.

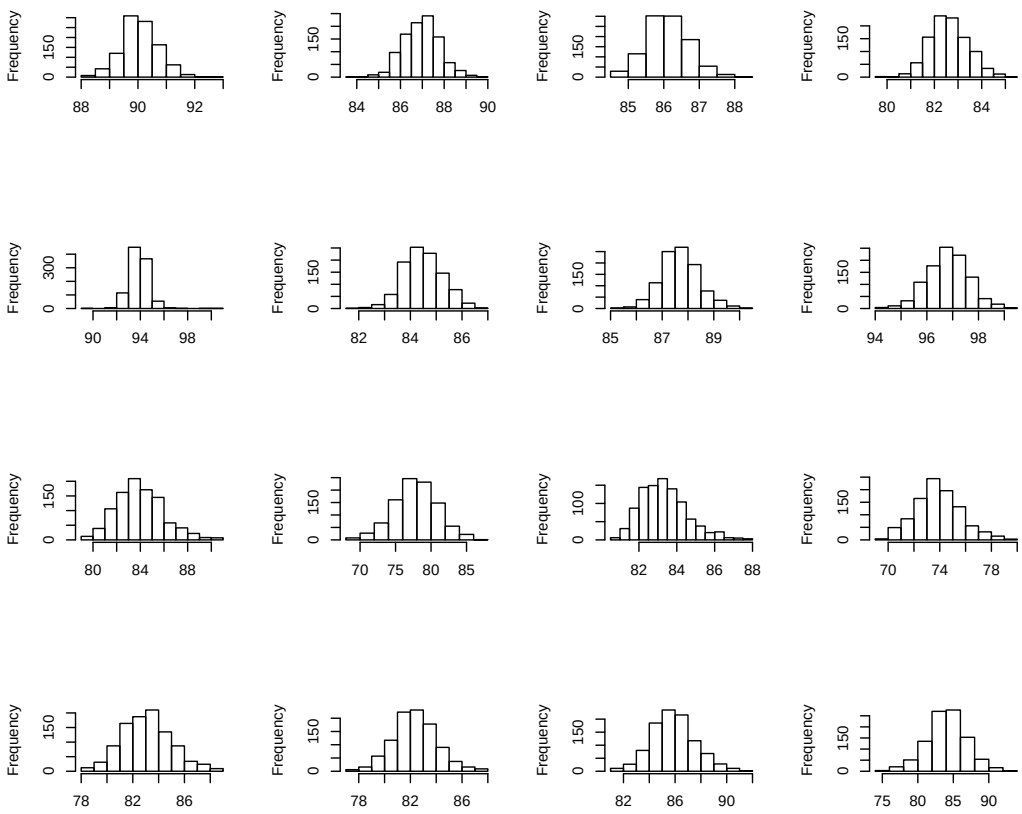

Figure 3.9: Bootstrapped distributions for randomly selected $\hat{e}_{BOOT}(t|x)$ out of 6400 combinations of t and x. The top 2 rows are 8 random selections for baseline age of 70. The bottom 2 rows are 8 random selections for baseline age of 50.

4 COMPARISON BETWEEN BACKWARD IMPUTATION METHOD AND BUCKLEY-JAMES METHOD

4.1 Introduction

When it comes to survival analysis, Cox proportional hazard model dominates statistical modeling. While there are many good reasons for using it, other models exist and can be a better choice under some circumstances. For example, Cox's model is seldom used if there is no censoring in the data in which case linear regression prevails. Buckley-James estimator, Buckley and James (1979) is a popular alternative to Cox's model as the usual least square regression adapted to censored data, which also tries to impute the censored times as the backward imputation method introduced in Chapter 3. This section delivers some preliminary results in simulation and real data comparing the backward imputation method and Buckley-James estimator. More investigation on the relationship between the two methods is being conducted and will be included as the results come up.

4.2 Buckley-James Method

Let T_i denote the transformed survival time, for instance, the logarithm of the failure time corresponding to the accelerated failure time model. Assuming that the true survival time is linearly related with the covariates X_i, then we can describe the relationship through the model

$$T_i = \alpha + X_i^T \beta + \epsilon_i, i = 1, \cdots, n, \tag{4.1}$$

where ϵ_i's are i.i.d random errors and are independent from the covariates. When T_i is subject to right censoring, what we observe is (Y_i, δ_i, X_i), where $Y_i = \min(T_i, C_i)$, C_i is the transformed censoring time by the same transformation for T_i, and $\delta_i = I_{\{T_i \leqslant C_i\}}$ is the censoring indicator.

The above model reduces to linear regression model when no censoring exists and one can estimate the parameters by least-square method. When right censoring exists, Buckley-James methods tries to make a reasonable guess for the true values of the censored times T_i. It turns out it carries the same idea as the widely-used single imputation in missing data in Little and Rubin (2014). Putting together the intercept α and the error ϵ_i, we define the new error term

$$\xi_i = \alpha + \epsilon_i = T_i - X_i^T \beta,$$

with the true coefficients β. When T_i is right censored, we observe Y_i and know that the true survival time T_i must be greater than Y_i. Hence, an intuitive imputation for Y_i is

$$E(T_i|T_i > Y_i, X_i) = X_i^T\beta + E(\xi_i|\xi_i > Y_i - X_i^T\beta) \tag{4.2}$$

$$= X_i^T\beta + \int_{Y_i - X_i^T\beta}^{\infty} \frac{tdF(t)}{1 - F(Y_i - X_i^T\beta)}, \tag{4.3}$$

where F is the c.d.f. of ξ. Buckley-James method obtains an estimation $\hat{F}$ of F by the Kaplan-Meier estimator for the corresponding survival function. Then one could plug in $\hat{F}$ in (4.3) to have an imputed value of Y_i given a value of β. We could rewrite the imputed response as

$$Y_i^* = \delta_i Y_i + (1 - \delta_i)E(T_i|T_i > Y_i, X_i), \tag{4.4}$$

and model (4.1) becomes

$$Y_i^* = \alpha + X_i^T\beta + \epsilon_i^*, i = 1, \cdots, n, \tag{4.5}$$

where ϵ_i^*'s are independent with mean zero. Least-square methods can be applied to (4.5) for estimating β. Buckley-James method iteratively solves for β and estimates α after the iteration converges.

4.3 Requirements of Buckley-James method

There are two main assumptions behind model (4.1) for Buckley-James estimator, namely linearity and homogeneity. Linearity comes in since model (4.1) explicitly assumes linear relationship between the covariates and the survival times. Homogeneity is the reason for dropping of X_i in the conditional expectations $E(\xi_i|\xi_i > Y_i - X_i^T\beta)$ in (4.2). These assumptions are easily to be violated in real data and are hard to detect due to censoring.

Here we simulated cases when linearity and homogeneity are not satisfied and display the data to tell the audiences that checking these assumptions under censoring is difficult and sometimes impossible. The first simulation violated homoscedasticity with the true model as $T_i = 1 + 2X_i + N(0, (\frac{1+X_i}{2})^2)$ and X_i's were randomly drawn from $U(0, 2)$. The censoring variable was independently drawn from uniform distribution resulting in about 50% of censoring. *Figure 4.1* presents two cases, namely sample size $n = 50$ and 300. It turns out that the inflated variance with increasing X is hardly observed even with a moderate sample size of $n = 300$ when the information of survival times is partly censored.

Simulations 2 and 3 have both nonlinearity and heteroscedasticity. In these two cases,

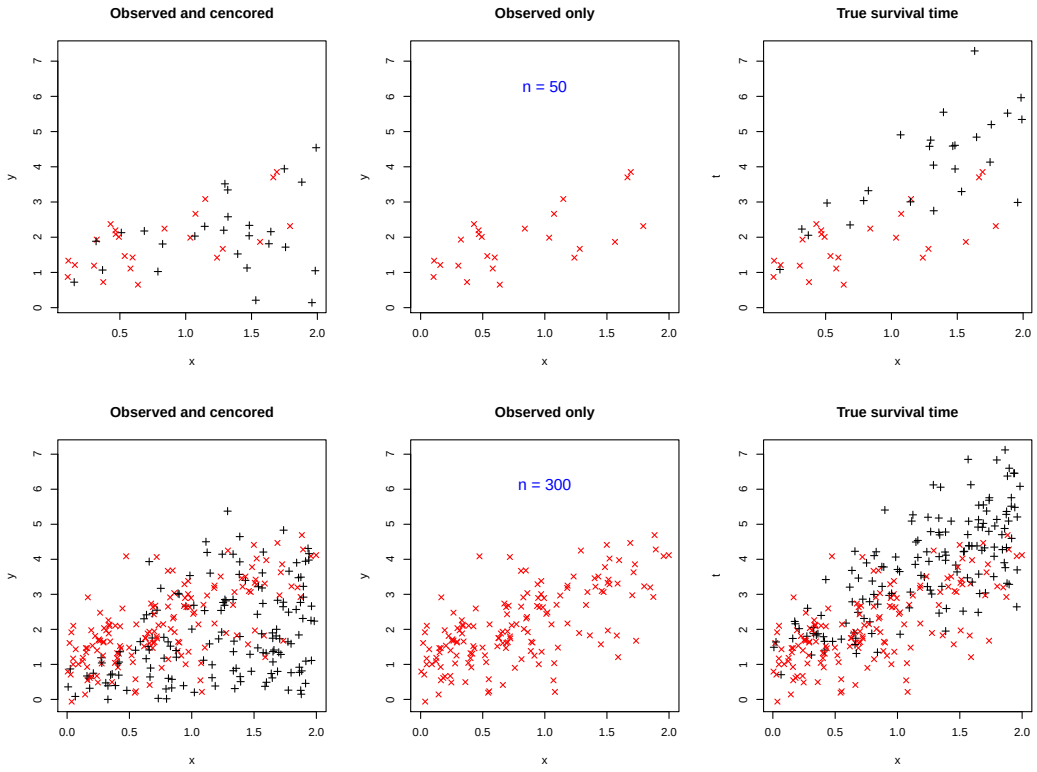

Figure 4.1: Simulated data from $T_i = 1 + 2X_i + N(0, (\frac{1+X_i}{2})^2)$ with about 50% right censoring. The plots in the top row correspond to sample size of 50. The bottom row is for sample size of 300. Red crosses and black pluses are for observed survival times and censored times respectively.

we generated survival times from $T_i = 1 - 2X_i^2 + N(0, (\frac{1+X_i}{2})^2)$ and $T_i = 1 + 4X_i \sin(\pi X_i) + N(0, (\frac{1+X_i}{2})^2)$ with X_i's independently from $U(0, 2)$ and sample size of 50. The censoring variables were independently drawn from uniform distribution resulting in about 50% of censoring. With the existence of censoring, the true nonlinear relationship between T and X is not clearly captured from the plots, shown in *Figure 4.2-4.3*. So is the hidden heterogeneity.

4.4 Comparing Buckley-James method and backward imputation method in simulation studies

We further investigated the performances of Buckley-James method and backward imputation method in the three simulation settings above. We included a parameter ρ controlling the degree of heterogeneity in the variance of the error so that the errors were drawn

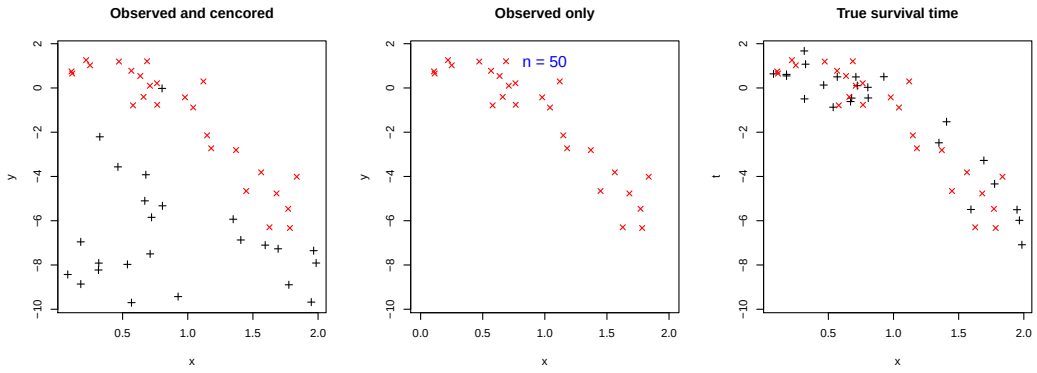

Figure 4.2: Simulated data from $T_i = 1 - 2X_i^2 + N(0, (\frac{1+X_i}{2})^2)$ with about 50% right censoring and sample size of 50. Red crosses and black pluses are for observed survival times and censored times respectively.

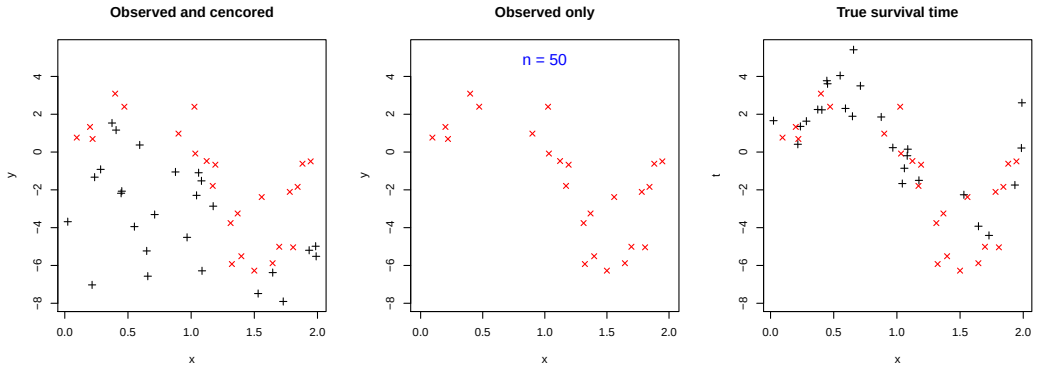

Figure 4.3: Simulated data from $T_i = 1 + 4X_i \sin(\pi X_i) + N(0, (\frac{1+X_i}{2})^2)$ with about 50% right censoring and sample size of 50. Red crosses and black pluses are for observed survival times and censored times respectively.

from $N(0, (1/2 + \rho X)^2)$ with $rho = 0, 0.2, 0.4, \cdots, 1$. For each of the three simulations, we implemented four imputation methods, namely Buckley-James with linear model, Buckley-James with SS-ANOVA model, backward imputation with linear model and backward imputation with SS-ANOVA model. In addition, we had the sample sizes $n = 50$ or 300 and the censoring rate to be approximately 30%, 50% and 70%. For every combination of the sample size and censoring rate, we took the average of mean squared error(MSE) of the imputed times and true survival times for the censored cases in 500 replications to evaluate the performance.

Figure 4.4 summarizes the simulation results for the first scenario when only homogeneity is violated. MSE increases almost linearly in the size of heteroscedasticity. When ρ is

close to 0, i.e., no or minor violation of homogeneity, there is almost no difference among the 4 methods in term of MSE of the imputed censored times. When ρ gets larger, the two Buckley-James methods result in worse imputation than the two backward imputation methods. Moreover, backward imputation methods seem to be less affected by higher censoring rate and smaller sample size. When the model is no longer linear, Buckley-James with linear model turns out to be more biased compared with the other three methods as depicted in *Figure 4.5-4.6*. The differences between Buckley-James and backward imputation under SS-ANOVA are exaggerated for censoring rate at 70% with heterogeneity in the variance of the error.

Based on the observations from the simulations, we found that the original Buckley-James linear model is very fragile with model mis-specification. Buckley-James method using SS-ANOVA model usually succeeds in imputation when nonlinearity exists. Backward imputation method with SS-ANOVA model is more robust to both nonlinearity and heteroscedasticity especially under circumstances with high censoring rate and small sample sizes compared to Buckley-James methods.

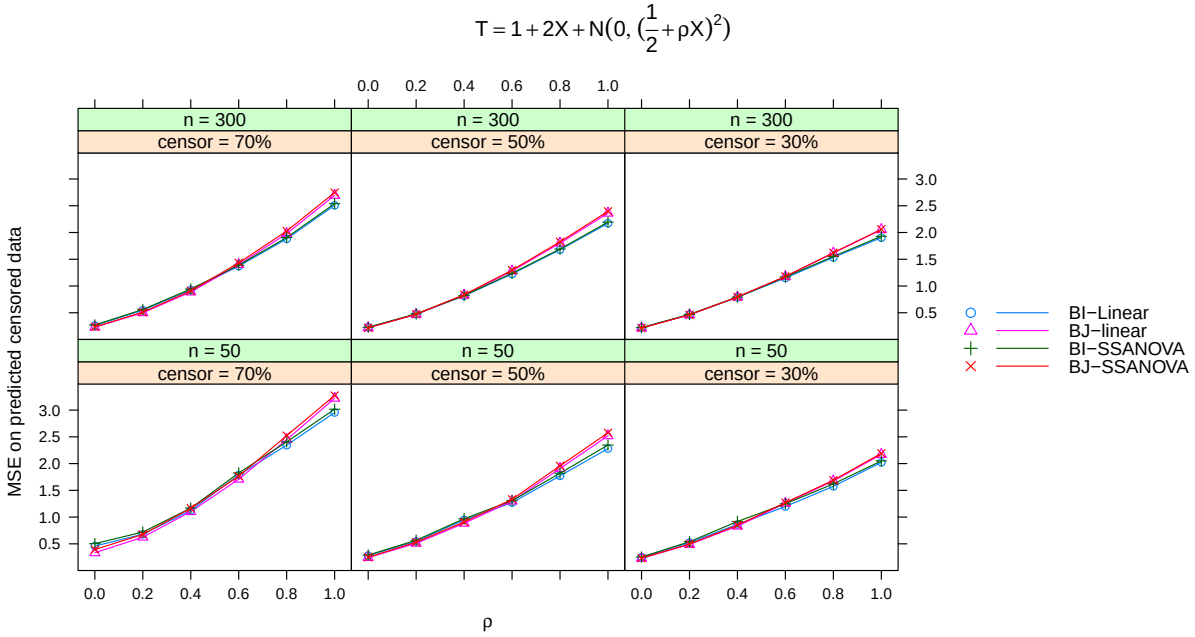

Figure 4.4: Summary of results for model $T_i = 1 + 2X_i + N(0, (\frac{1}{2} + \rho X_i)^2)$ with $\rho = 0, 0.2, 0.4, \cdots, 1$.

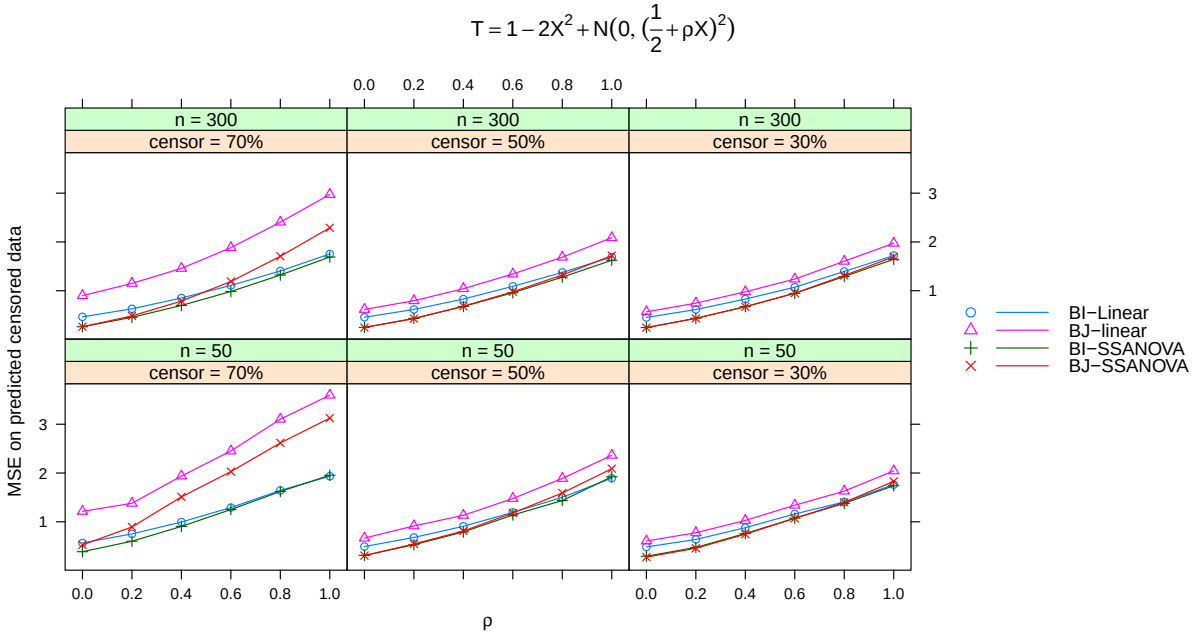

Figure 4.5: Summary of results for model $T_i = 1 - 2X_i^2 + N(0, (\frac{1}{2} + \rho X_i)^2)$ with $\rho = 0, 0.2, 0.4, \cdots, 1$.

4.5 Comparing Buckley-James method and backward imputation method with real data

Stanford heart transplantation data

In this section, we implemented the two methods in two well-known datasets, namely the Stanford heart transplantation data in Miller and Halpern (1982) and the veteran's administration lung cancer trial data in Prentice (1973), to explore the behaviors of the two in real data. Let's first focus on the heart transplantation data. The time-to-event outcome of this dataset is the lifetime since first heart transplantation between October 1967 and February 1980. 55 out of 157 patients were censored with two covariates, age at the time of first transplant and the T5 mismatch score which measures the degree of tissue incompatibility between the hearts of the initial donor and recipient with respect to HLA antigens. *Figure 4.7* describes the relationship between the logarithm of lifetimes versus the the variables. Although the censoring hides the true relationship, we don't observe obvious nonlinear patterns. It does seem that the variance is slightly inflated when T5 mismatch score is large and age is small.

Random samples were drawn from the uncensored subjects as test set and the rest data were treated as training data. We implemented Buckley-James with linear and SS-ANOVA

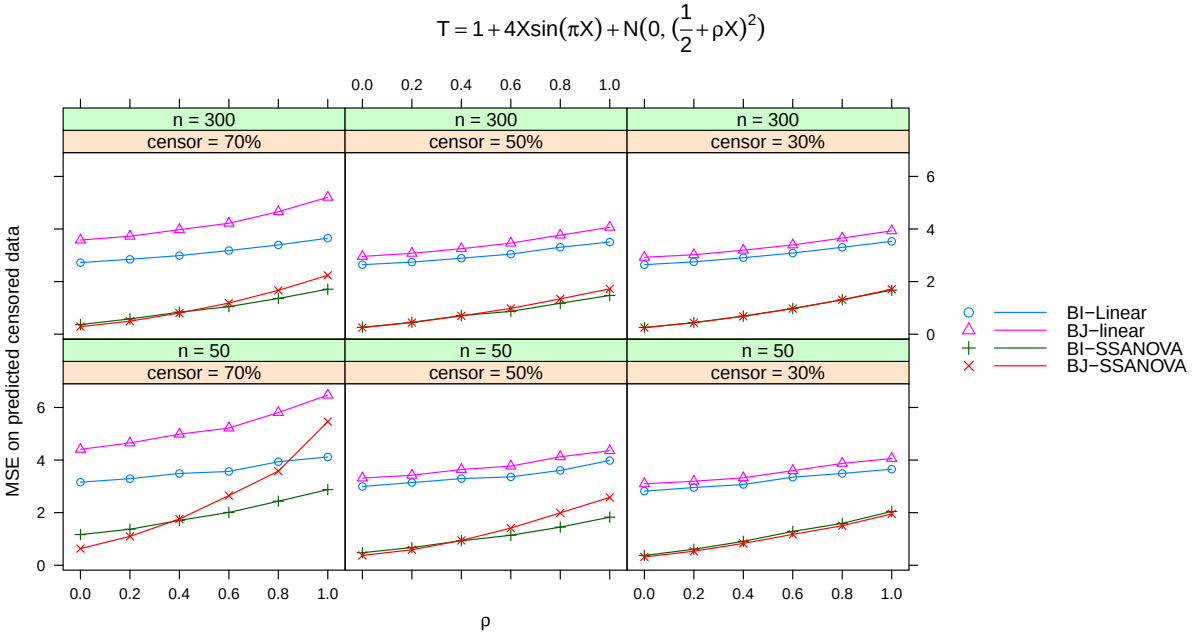

Figure 4.6: Summary of results for model $T_i = 1 + 4X_i \sin(\pi X_i) + N(0, (\frac{1}{2} + \rho X_i)^2)$ with $\rho = 0, 0.2, 0.4, \cdots, 1$.

models and backward imputation with SS-ANOVA model on the training data with log survival times where main effects for T5 mismatch score and age were fitted.(Backward imputation with linear model for comparison was not included here since the previous simulation studies indicate that SS-ANOVA is probably a better choice combining with backward imputation.) With the censored times being imputed in the training data, models were fitted on the three imputed training data to predict the lifetimes in the test set. Then, the mean squared error of the predicted lifetimes in the test set can be used as a measurement of performance. Moreover, we also fitted linear regressions to the three imputed training data to see how far away the estimated coefficients were from the ones out of an accelerated failure time(AFT) model on the whole dataset. If the coefficients of the AFT model are assumed to be the ground truth, then the l_2 norm between the coefficients from the imputed training data and the whole data evaluates the success of the imputation procedure as well. In order to create different censoring rate, we generated test sets of sizes 30, 50 and 70 with 500 replications for each test sample size. *Table 4.1* displays the averages of MSE and l_2 norm for the three methods under different test sample sizes. The fact that Buckley-James with linear model outperforms Buckley-James with SS-ANOVA model confirms the linear relationships between the two covariates and the survival. With all three test sample sizes, backward imputation is in the leading place under the two evaluation metrics. The use of

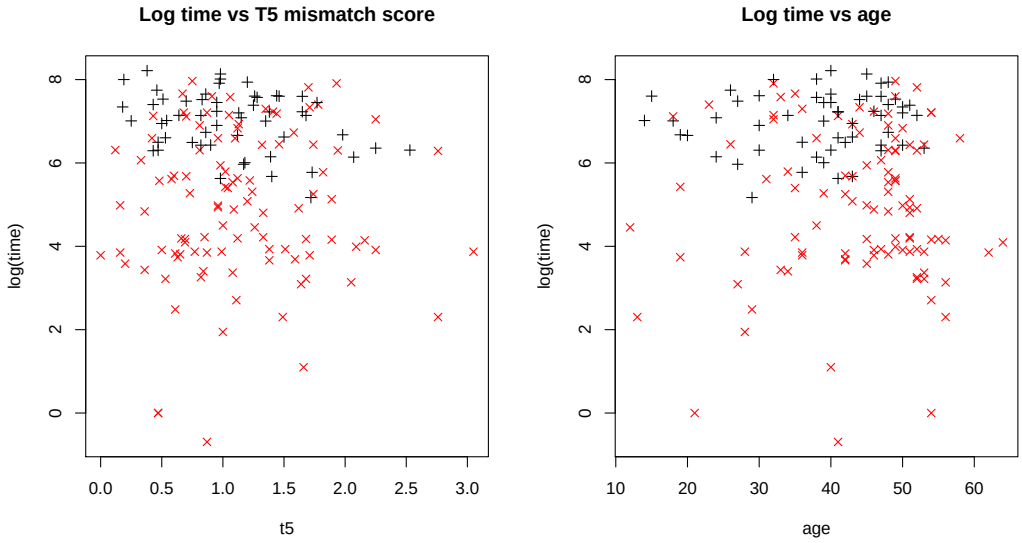

Figure 4.7: T5 mismatch score and age versus log survival times for Stanford heart transplantation data. Red crosses and black pluses are for observed survival times and censored lifetimes respectively.

testsize	MSE BI-S	MSE BJ-L	MSE BJ-S	l_2 BI-S	l_2 BJ-L	l_2 BJ-S
30	4.945	5.309	5.411	0.556	0.614	0.606
50	5.582	6.123	6.437	0.552	1.103	1.029
70	6.778	8.041	8.890	0.815	2.229	2.135

Table 4.1: Comparisons of Buckley-James method and backward imputation method for Stanford heart transplantation data.

backward imputation further improves the imputed result if the censoring rate is relatively high (testsize of 70 leads to about 2/3 of censoring).

Veteran's administration lung cancer data

The veteran's administration lung cancer data include 137 patients with 9 censored survival times. Information of age, treatment type (standard or test), tumor type (4 types), prior therapy (yes or no) and Karnofsky score was collected on each patient at the time of entry into the study. Karnofsky score is a measure between 10 to 100 of general health with 10 indicating that the patient is completely hospitalized and 100 suggests the patient is able to take care of him or herself. *Figure 4.8* displays age and Karnofsky score versus the logarithm of survival times. Both variables seem to be nonlinearly related with the survival

times and variance of the response declines as the age and Karnofsky score increase.

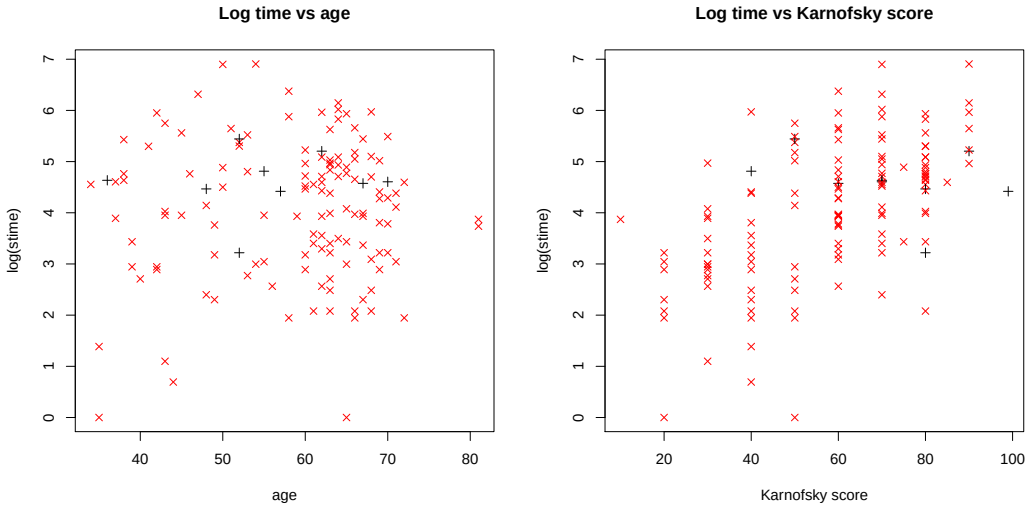

Figure 4.8: Age and Karnofsky score versus log survival times for veteran's administration lung cancer data. Red crosses and black pluses are for observed survival times and censored lifetimes respectively.

Only 9 patients were censored, thus one has to leave out most part of the data as testing set if the analysis in the previous transplant data is desired. Here we created pseudo censoring variables from $U(0, V)$ with $V = 100, 300$ or 500 to generate different censoring rate with 500 replication for each value of V. We always kept the 9 originally censored patients as censored. Again, we compared Buckley-James with linear and SS-ANOVA models and backward imputation with SS-ANOVA model on the pseudo data with log scaled censored survival times. All three methods fitted main effects for the 5 covariates with an additional interaction term between prior therapy and Karnofsky score as discovered in McLain and Ghosh (2011). Two measurements were used to evaluate the performance of the three methods. First, we calculated the MSE for the imputed censored times for pseudo censored cases. Second, linear regression models were ran on the imputed data sets out of the three methods and the coefficients were compared to the ones estimated from AFT on the original dataset with 9 censored subjects by l_2 norm of the difference. *Table 4.2* takes averages of the 500 replications for each values of V. As we expected by looking at the scatterplots, the nonlinearity influences the results of Buckley-James with linear model. Backward imputation with SS-ANOVA model performs the best under all three censoring rate and is more robust towards nonlinearity and heterogeneity.

V	MSE BI-S	MSE BJ-L	MSE BJ-S	l_2 BI-S	l_2 BJ-L	l_2 BJ-S
500	0.447	0.587	0.452	0.108	0.241	0.184
300	0.480	0.618	0.498	0.165	0.301	0.238
100	0.681	0.979	0.839	0.343	0.616	0.542

Table 4.2: Comparisons of Buckley-James method and backward imputation method for veteran's administration lung cancer data.

5 CONCLUDING REMARKS

This thesis contains 4 individual parts studying different problems. The first piece of work focused on the question how familial relationships, the sharing lifestyles and diseases intertwine with each other to influence longevity. The Beaver Dam eye study contains a large number of people with relatives in the study, this provided an ideal opportunity to examine the pairwise associations among the different components. We used the recently developed statistical tool distance correlation to fulfill this goal since the familial relationships can be treated as pairwise distances which is appropriate for distance correlation but not other kinds of correlations. We have shown that pairwise differences in lifestyle factors and diseases that run in families correlate well with pairwise differences in death age that also run in families, partially accounting for the familial death age effect. The signal of running in families gets stronger if the relationship among people are closer as we observed by restricting the analysis to full siblings in the study.

The second work introduced a new variable selection procedure based on the property of distance covariance and demonstrated the application through two examples. The procedure first ranks the importance of the predictors by distance correlation between individual predictors and the response. Then it decides if a new predictor should be accepted by looking at whether adding it to the calculation of distance covariance increases the value of distance covariance sequentially. The small round blue cell tumors data played a role of a toy example to show that the performance of the proposed method worked well in easy cases. The TCGA ovarian cancer data, however, were much more challenging to deal with due to the humongous number of variables and very limited sample size. The uncertainty of variable selection was discussed through gene selection results using random subsets of the data. The support vector machine with reject option was used to withhold decision for subjects who were difficult to classify. An ensemble method of combining models built on random subsets of the data was implemented to assess the prediction performance.

In the third part, we presented our backward multiple imputation framework for estimating the conditional lifetime expectancy function. In the case without covariates, our estimator is proven to be equivalent to the estimation for LEF by inverting the Kaplan-Meier survival function estimator. In the case with covariates, one is free to select a base model that best captures the data. One is able to recover the nonparametric estimator for conditional LEF proposed in McLain and Ghosh (2011) based on the generalized Kaplan-Meier estimator by using kernel regression in our framework. The simulation studies demonstrated the performance of our methods and validated the use of multiple

imputation for variance estimation under three different settings. The application to the Beaver Dam Eye Study data illustrated the use of SS-ANOVA model together with our backward multiple imputation method. We showed the fitted results for the cohorts with baseline age of 70 where a number of variables, including gender, smoking, education level, BMI values and several diseases, were shown to be significantly associated with the human longevity. Moreover, the effects of the covariates on expected lifetime seem to diminish as the conditional time one already survives increases.

The last part targets at comparing our backward imputation method and the well-known Buckley-James method for imputing right censored survival data. First, we discussed that the linear Buckley-James method fails when nonlinearity and heterogeneity exist through simulated data. The results suggest that backward imputation with SS-ANOVA is less biased for nonlinear and heterogenous data, especially with small sample size and high censoring rate. The comparison through two real examples, namely the Stanford heart transplantation data and the veteran's administration lung cancer data, both convince the better performance of backward imputation with SS-ANOVA over Buckley-James method. More theoretical research between the two methods are being conducted and will be included.

A.1 Proof of Theorem 1 for Chapter 3

Proof. It is easy to see that both $\hat{e}_B(t)$ and $\hat{e}_{KM}(t)$ are step functions. So we only need to prove that both functions jump at the same t with the same value. From the explicit expression of $\hat{e}_{KM}(t)$, we know that $\hat{e}_{KM}(t)$ is left continuous and is discrete at $t_{(1)}, \cdots, t_{(K)}$. For $\hat{e}_B(t)$ and a particular value t^*:

1. If $t^* \notin \{y_1, \cdots, y_n\}$, then $I_{\{y_i > t^*\}}$ is continuous in the neighborhood of t^* for all i and hence $\hat{e}_B(t)$ is continuous around t^*.

2. If $t^* \in \{c_{(1)}, \cdots, c_{(J)}\} \backslash \{t_{(1)}, \cdots, t_{(K)}\}$ and denote the number of censored data points at t^* by n^*, then $I_{\{y_i > t^*\}} = I_{\{y_i > t^* +\}}$ for all i. Hence,

$$\hat{e}_B(t^*) = \hat{e}_B(t^* +).$$

   In addition, since $t^* \in \{c_{(1)}, \cdots, c_{(J)}\} \backslash \{t_{(1)}, \cdots, t_{(K)}\}$, we have

$$\hat{e}_B(t^*) = \frac{\sum_{i=1}^n \tilde{y}_i I_{\{y_i > t^*\}}}{\sum_{i=1}^n I_{\{y_i > t^*\}}}$$

   and

$$\begin{aligned}
\hat{e}_B(t^* -) &= \frac{\sum_{i=1}^n \tilde{y}_i I_{\{y_i > t^* -\}}}{\sum_{i=1}^n I_{\{y_i > t^* -\}}} \\
&= \frac{\hat{e}_B(t^*) n^* + \sum_{i=1}^n \tilde{y}_i I_{\{y_i > t^*\}}}{n^* + \sum_{i=1}^n I_{\{y_i > t^*\}}} \\
&= \frac{\hat{e}_B(t^*) n^* + \hat{e}_B(t^*) \sum_{i=1}^n I_{\{y_i > t^*\}}}{n^* + \sum_{i=1}^n I_{\{y_i > t^*\}}} \\
&= \hat{e}_B(t^*).
\end{aligned}$$

   Hence, $\hat{e}_B(t)$ is continuous at times when only censoring occurs.

3. If $t^* \in \{t_{(1)}, \cdots, t_{(K)}\}$ and suppose $t^* = t_{(k)}$ for some $k = 1, \cdots, K$. If $k = K$, it is

obvious that $\hat{e}_B(t_{(K)}) = \hat{e}_{KM}(t_{(K)}) = 0$. For $k = 1, \cdots, K-1$,

$$\begin{aligned}\hat{e}_B(t_{(k)}) &= \frac{\sum_{i=1}^n \tilde{y}_i I_{\{y_i > t_{(k)}\}}}{\sum_{i=1}^n I_{\{y_i > t_{(k)}\}}} \\ &= \frac{\sum_{i=1}^n \tilde{y}_i I_{\{y_i \geqslant t_{(k+1)}\}} + \sum_{i=1}^n \tilde{y}_i I_{\{t_{(k)} < y_i < t_{(k+1)}\}}}{\sum_{i=1}^n I_{\{y_i > t_{(k)}\}}},\end{aligned}$$

Note that within $(t_{(k)}, t_{(k+1)})$, one can only have censored observations. By the backward imputation procedure and analysis on censored times above, we know that

$$\tilde{y}_i = \hat{e}_B(y_i) = \hat{e}_B(t_{(k+1)}-)$$

for $t_{(k)} < y_i < t_{(k+1)}$. Therefore,

$$\begin{aligned}\hat{e}_B(t_{(k)}) &= \frac{\hat{e}_B(t_{(k+1)}-) \sum_{i=1}^n I_{\{y_i \geqslant t_{(k+1)}\}} + \hat{e}_B(t_{(k+1)}-) \sum_{i=1}^n I_{\{t_{(k)} < y_i < t_{(k+1)}\}}}{\sum_{i=1}^n I_{\{y_i > t_{(k)}\}}} \\ &= \hat{e}_B(t_{(k+1)}-).\end{aligned}$$

The above result shows that $\hat{e}_B(t)$ jumps only at $\{t_{(1)}, \cdots, t_{(K)}\}$ and is left continuous. Suppose there exists an l such that $c_{(l)} = t_{(k)}$, i.e. there are both events and censoring happened at $t = t_{(k)}$ (if $t_{(k)}$ is a pure event time point, the following analysis still applies by deleting all the terms related with $c_{(l)}$), then

$$\begin{aligned}\hat{e}_B(t_{(k)}-) &= \frac{\sum_{i=1}^n \tilde{y}_i I_{\{y_i \geqslant t_{(k)}\}}}{\sum_{i=1}^n I_{\{y_i \geqslant t_{(k)}\}}} \\ &= \frac{\sum_{i=1}^n \tilde{y}_i I_{\{y_i > t_{(k)}\}} + n(t_{(k)}) t_{(k)} + n(c_{(l)}) \hat{e}_B(c_{(l)})}{\sum_{i=1}^n I_{\{y_i \geqslant t_{(k)}\}}} \\ &= \frac{\hat{e}_B(t_{(k)}) \sum_{i=1}^n I_{\{y_i > t_{(k)}\}} + n(t_{(k)}) t_{(k)} + n(c_{(l)}) \hat{e}_B(t_{(k)})}{\sum_{i=1}^n I_{\{y_i > t_{(k)}\}} + n(t_{(k)}) + n(c_{(l)})}. \qquad \text{(A.1)}\end{aligned}$$

Now, let's look at $\hat{e}_{KM(t)}$. By the explicit formula, we know that for $k = 1, \cdots, K-1$,

$$\begin{aligned}\hat{e}_{KM}(t_{(k)-}) &= t_{(k-1)} + \frac{1}{\hat{S}_{KM}(t_{(k-1)})} \sum_{l=k}^{K} (t_{(l)} - t_{(l-1)}) \hat{S}_{KM}(t_{(l-1)}) \\ \hat{e}_{KM}(t_{(k)}) &= t_{(k)} + \frac{1}{\hat{S}_{KM}(t_{(k)})} \sum_{l=k+1}^{K} (t_{(l)} - t_{(l-1)}) \hat{S}_{KM}(t_{(l-1)}),\end{aligned}$$

Thus, it is easy to see that

$$\hat{e}_{KM}(t_{(k)-}) = \hat{e}_{KM}(t_{(k)})\frac{\hat{S}_{KM}(t_{(k)})}{\hat{S}_{KM}(t_{(k-1)})} + t_{(k)}\left[1 - \frac{\hat{S}_{KM}(t_{(k)})}{\hat{S}_{KM}(t_{(k-1)})}\right]. \tag{A.2}$$

By the definition of Kaplan-Meier estimator, we have

$$\begin{aligned}\frac{\hat{S}_{KM}(t_{(k)})}{\hat{S}_{KM}(t_{(k-1)})} &= \frac{\prod_{t_{(l)} \leqslant t_{(k)}}\left[1 - \frac{n(t_{(l)})}{\sum_{i=1}^{n} I_{\{y_i \geqslant t_{(l)}\}}}\right]}{\prod_{t_{(l)} \leqslant t_{(k-1)}}\left[1 - \frac{n(t_{(l)})}{\sum_{i=1}^{n} I_{\{y_i \geqslant t_{(l)}\}}}\right]} \\ &= 1 - \frac{n(t_{(k)})}{\sum_{i=1}^{n} I_{\{y_i \geqslant t_{(k)}\}}} \\ &= \frac{\sum_{i=1}^{n} I_{\{y_i > t_{(k)}\}} + n(c_{(l)})}{\sum_{i=1}^{n} I_{\{y_i \geqslant t_{(k)}\}}}\end{aligned} \tag{A.3}$$

Putting (2) and (3) together yields the following results:

$$\hat{e}_{KM}(t_{(k)-}) = \frac{\hat{e}_{KM}(t_{(k)})\sum_{i=1}^{M} n_i I_{\{y_{(i)} > t_{(k)}\}} + n(t_{(k)})t_{(k)} + n(c_{(l)})\hat{e}_{KM}(t_{(k)})}{\sum_{i=1}^{M} n_i I_{\{y_{(i)} > t_{(j)}\}} + n(t_{(k)}) + n(c_{(l)})}. \tag{A.4}$$

By the fact that $\hat{e}_B(t_{(K)}) = \hat{e}_{KM}(t_{(K)})$ together with (1) and (4), we know that $\hat{e}_B(t_{(k)}) = \hat{e}_{KM}(t_{(k)})$ for $k = 1, \cdots, K-1$. With the fact that both functions are left continuous step functions with same jump locations, we conclude that $\hat{e}_B(t) = \hat{e}_{KM}(t)$ for all $t \geqslant 0$.

□

A.2 Proof of Theorem 2 for Chapter 3

Before we jump into the proof, we present the nonparametric estimation $\hat{e}_P(t|x)$ defined in McLain and Ghosh (2011). Let $K : \mathbb{R}^p \to \mathbb{R}$ be the p-dimensional kernel function and h_n denotes the bandwidth and

$$W_{ni}(x|h_n) = \frac{K(\frac{x - x_i}{h_n})}{\sum_{j=1}^{n} K(\frac{x - x_j}{h_n})}$$

which is the weight of the ith observation according to the closeness to target point x. Then the generalized Kaplan-Meier estimatorDabrowska et al. (1989); Gonzalez-Manteiga and

Cadarso-Suarez (1994) takes the form of

$$\hat{S}_P(t|x) = I_{\{t \leqslant Y_{(n)}\}} \prod_{\{i: Y_{(i)} \leqslant t\}} \left\{ \frac{\sum_{j=i+1}^{n} W_{n(j)}(x|h_n)}{\sum_{j=i}^{n} W_{n(j)}(x|h_n)} \right\}^{\delta_{(i)}},$$

where $Y_{(i)}$ denotes the ith order statistic, and $\{\delta_{(i)}, W_{n(i)}(x|h_n)\}$ denote the corresponding censoring indicator and weight of the ith observation.

Let $t_{(1)} < \cdots < t_{(K)}$ be the distinct ordered values of the observed survival time. Then $\hat{e}_P(t|x)$ by inverting $S_P(t|x, h_n)$ is given by

$$\hat{e}_P(t|x) = \begin{cases} t_{(k)} + \dfrac{1}{\hat{S}_P(t_{(k-1)}|x)} \displaystyle\sum_{l=k+1}^{K} (t_{(l)} - t_{(l-1)}) \hat{S}_P(t_{(l-1)}|x), & t_{(k-1)} < t < t_{(k)} \\ t_{(k)} + \dfrac{1}{\hat{S}_P(t_{(k)}|x)} \displaystyle\sum_{l=k+1}^{K} (t_{(l)} - t_{(l-1)}) \hat{S}_P(t_{(l-1)}|x), & t = t_{(k)}, k = 1, \cdots, K-1 \\ 0, & t \geqslant t_{(K)}. \end{cases}$$

Proof. For a given covariates value x, $\hat{e}_B(t|x)$ is derived by using Kernel regression with K and h_n as the base model in *Algorithm 2*. This is equivalent to implement a weighted version of *Algorithm 1* with $W_{ni}(x|h_n)$ as weights for the n data points. Notice that $\hat{S}_P(t|x)$ differs from $\hat{S}_{KM}(t)$ by imposing weights $W_{ni}(x|h_n)$ to the n cases. Therefore, it is straightforward to prove Theorem 2 by using the proof of Theorem 1 by introducing weights $W_{ni}(x|h_n)$ instead of the uniform weights. □

REFERENCES

Bartlett, Peter L, and Marten H Wegkamp. 2008. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research* 9:1823–1840.

Bast, Robert C, and Gordon B Mills. 2010. Personalizing therapy for ovarian cancer: Brcaness and beyond. *Journal of Clinical Oncology* 28(22):3545–3548.

Basu, Sudipta, Rania Harfouche, Shivani Soni, Geetanjali Chimote, Raghunath A Mashelkar, and Shiladitya Sengupta. 2009. Nanoparticle-mediated targeting of mapk signaling predisposes tumor to chemotherapy. *Proceedings of the National Academy of Sciences* 106(19):7957–7961.

Bell, D, A Berchuck, M Birrer, J Chien, DW Cramer, F Dao, R Dhir, P DiSaia, H Gabra, P Glenn, et al. 2011. Integrated genomic analyses of ovarian carcinoma.

Bitler, Benjamin G, Jasmine P Nicodemus, Hua Li, Qi Cai, Hong Wu, Xiang Hua, Tianyu Li, Michael J Birrer, Andrew K Godwin, Paul Cairns, et al. 2011. Wnt5a suppresses epithelial ovarian cancer by promoting cellular senescence. *Cancer research* 71(19):6184–6194.

Bravo, Héctor Corrada, Kristine E Lee, Barbara EK Klein, Ronald Klein, Sudha K Iyengar, and Grace Wahba. 2009. Examining the relative influence of familial, genetic, and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences* 106(20):8128–8133.

Buckley, Jonathan, and Ian James. 1979. Linear regression with censored data. *Biometrika* 66(3):429–436.

Chen, Ying Qing, and S Cheng. 2006. Linear life expectancy regression with censored data. *Biometrika* 93(2):303–313.

Cole, Claire, Sin Lau, Alison Backen, Andrew Clamp, Graham Rushton, Caroline Dive, Cassandra Hodgkinson, Rhona McVey, Henry Kitchener, and Gordon C Jayson. 2010. Inhibition of fgfr2 and fgfr1 increases cisplatin sensitivity in ovarian cancer. *Cancer biology & therapy* 10(5):495–504.

Cox, D. R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2):187–220.

Craven, Peter, and Grace Wahba. 1977. *Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation*. Department of Statistics, University of Wisconsin.

Dabrowska, Dorota M, et al. 1989. Uniform consistency of the kernel conditional kaplan-meier estimate. *The Annals of Statistics* 17(3):1157–1167.

Efron, Bradley. 1967. The two sample problem with censored data. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability*, vol. 4, 831–853. Prentice-Hall Engewood Cliffs, NJ.

Fan, Jianqing, and Jinchi Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5): 849–911.

Gao, Fangyu, Grace Wahba, Ronald Klein, and Barbara Klein. 2001. Smoothing spline anova for multivariate bernoulli observations with application to ophthalmology data. *Journal of the American Statistical Association* 96(453):127–160.

Golub, Gene H, Michael Heath, and Grace Wahba. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2):215–223.

Gonzalez-Manteiga, W, and C Cadarso-Suarez. 1994. Asymptotic properties of a generalized kaplan-meier estimator with some applications. *Communications in Statistics-Theory and Methods* 4(1):65–78.

Gu, Chong. 2007. gss: General smoothing splines. *R package version* 1–0.

———. 2013. *Smoothing spline anova models*, vol. 297. Springer Science & Business Media.

Hall, W. J., and J. Wellner. Mea residual life. In *Proceedings of the internaltional symposium on statistics and related topics*, 169–184. Amsterdam, North-Holland.

Kaplan, Edward L, and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* 53(282):457–481.

Khan, J., J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, et al. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine* 7(6):673–679.

Khoshgnauz, Ehsan. 2012. Learning markov network structure using brownian distance covariance. *arXiv preprint arXiv:1206.6361*.

Kimeldorf, George, and Grace Wahba. 1971. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications* 33(1):82–95.

Klein, Ronald, Barbara EK Klein, Kathryn LP Linton, and David L De Mets. 1991. The beaver dam eye study: visual acuity. *Ophthalmology* 98(8):1310–1315.

Kong, Jing, Barbara EK Klein, Ronald Klein, Kristine E Lee, and Grace Wahba. 2012. Using distance correlation and ss-anova to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proceedings of the National Academy of Sciences* 109(50): 20352–20357.

Kong, Jing, Sijian Wang, and Grace Wahba. 2015. Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in medicine* 34(10):1708–1720.

Kosorok, M.R. 2009. Discussion of: brownian distance covariance. *The Annals of Applied Statistics* 3(4):1270–1278.

Ledent, Catherine, Isabelle Demeestere, David Blum, Julien Petermans, Tuula Hämäläinen, Guillaume Smits, and Gilbert Vassart. 2005. Premature ovarian aging in mice deficient for gpr3. *Proceedings of the National Academy of Sciences of the United States of America* 102(25): 8922–8926.

Lee, Yoonkyung, Yi Lin, and Grace Wahba. 2004. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* 99(465):67–81.

Li, Runze, Wei Zhong, and Liping Zhu. 2012. Feature screening via distance correlation learning. *Journal of the American Statistical Association* 107(499):1129–1139.

Little, Roderick JA, and Donald B Rubin. 2014. *Statistical analysis with missing data*. John Wiley & Sons.

Lu, Fan, Sündüz Keleş, Stephen J Wright, and Grace Wahba. 2005. Framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences of the United States of America* 102(35):12332–12337.

Lyons, Russell, et al. 2013. Distance covariance in metric spaces. *The Annals of Probability* 41(5):3284–3305.

Malécot, Gustave, et al. 1948. Mathematics of heredity. *Les mathematiques de l'heredite.*

McLain, Alexander C, and Sujit K Ghosh. 2011. Nonparametric estimation of the conditional mean residual life function with censored data. *Lifetime data analysis* 17(4):514–532.

Mehlmann, Lisa M, Yoshinaga Saeki, Shigeru Tanaka, Thomas J Brennan, Alexei V Evsikov, Frank L Pendola, Barbara B Knowles, John J Eppig, and Laurinda A Jaffe. 2004. The gs-linked receptor gpr3 maintains meiotic arrest in mammalian oocytes. *Science* 306(5703): 1947–1950.

Miller, Rupert, and Jerry Halpern. 1982. Regression with censored data. *Biometrika* 69(3): 521–531.

Oakes, David, and Tamraparni Dasu. 1990. A note on residual life. *Biometrika* 77(2): 409–410.

Prentice, Ross L. 1973. Exponential survivals with censoring and explanatory variables. *Biometrika* 60(2):279–288.

Rho, Seung Bae, Seung Myung Dong, Sokbom Kang, Sang-Soo Seo, Chong Woo Yoo, Dong Ock Lee, Jong Soo Woo, and Sang-Yoon Park. 2008. Insulin-like growth factor-binding protein-5 (igfbp-5) acts as a tumor suppressor by inhibiting angiogenesis. *Carcinogenesis* 29(11):2106–2111.

Rubin, Donald B. 2004. *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons.

Selvanayagam, Zachariah E, Tak Hong Cheung, Nien Wei, Ragini Vittal, Keith Wing Kit Lo, Winnie Yeo, Tsunekazu Kita, Roald Ravatn, Tony Kwok Hung Chung, Yick Fu Wong, et al. 2004. Prediction of chemotherapeutic response in ovarian cancer with dna microarray expression profiling. *Cancer genetics and cytogenetics* 154(1):63–66.

Sun, Liuquan, and Zhigang Zhang. 2009. A class of transformed mean residual life models with censored survival data. *Journal of the American Statistical Association* 104(486):803–815.

Székely, Gábor J, Maria L Rizzo, Nail K Bakirov, et al. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6):2769–2794.

Székely, Gábor J, Maria L Rizzo, et al. 2009. Brownian distance covariance. *The annals of applied statistics* 3(4):1236–1265.

Thiele, Sylvia, Martina Rauner, Claudia Goettsch, Tilman D Rachner, Peggy Benad, Susanne Fuessel, Kati Erdmann, Christine Hamann, Gustavo B Baretton, Manfred P Wirth, et al. 2011. Expression profile of wnt molecules in prostate cancer and its regulation by aminobisphosphonates. *Journal of cellular biochemistry* 112(6):1593–1600.

Tran, Minh-Ngoc, David J Nott, Robert Kohn, et al. 2012. Simultaneous variable selection and component selection for regression density estimation with mixtures of heteroscedastic experts. *Electronic Journal of Statistics* 6:1170–1199.

Wahba, Grace. 1990. *Spline models for observational data*, vol. 59. Siam.

Wahba, Grace, Yuedong Wang, Chong Gu, Ronald Klein, Barbara Klein, et al. 1995. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy: the 1994 neyman memorial lecture. *The Annals of Statistics* 23(6):1865–1895.

Wang, Yuedong. 2011. *Smoothing splines: methods and applications*. CRC Press.

website, National Cancer Institute. http://www.cancer.gov/cancertopics/types/ovarian. *Accessed Aug 28, 2013*.

Wegkamp, Marten, Ming Yuan, et al. 2011. Support vector machines with a reject option. *Bernoulli* 17(4):1368–1385.

Yin, Jikai, Karen Lu, Jie Lin, Lei Wu, Michelle AT Hildebrandt, David W Chang, Larissa Meyer, Xifeng Wu, and Dong Liang. 2011. Genetic variants in tgf-β pathway are associated with ovarian cancer risk. *PloS one* 6(9):e25559.