

VARIABLE SELECTION AND SUPERVISED
DIMENSION REDUCTION FOR LARGE-SCALE
GENOMIC DATA WITH CENSORED SURVIVAL
OUTCOMES

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fullfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Lauren N. Spirko
Diploma Date August 2017

Examining Committee Members:

Dr. Karthik Devarajan, Dissertation Advisor, Fox Chase Cancer
Center

Dr. Cheng Yong Tang, Advisory Chair, Department of
Statistical Science

Dr. Pallavi Chitturi, Department of Statistical Science

Dr. Yuexiao Dong, Department of Statistical Science

Dr. Zoran Obradovic, Department of Computer and
Information Sciences

©

by

Lauren N. Spirko

August, 2017

All Rights Reserved

ABSTRACT

VARIABLE SELECTION AND SUPERVISED DIMENSION REDUCTION FOR
LARGE-SCALE GENOMIC DATA WITH CENSORED SURVIVAL OUTCOMES

Lauren N. Spirko

DOCTOR OF PHILOSOPHY

Temple University, August 2017

Dr. Karthik Devarajan, Dissertation Advisor

One of the major goals in large-scale genomic studies is to identify genes with a prognostic impact on time-to-event outcomes, providing insight into the disease's process. With the rapid developments in high-throughput genomic technologies in the past two decades, the scientific community is able to monitor the expression levels of thousands of genes and proteins resulting in enormous data sets where the number of genomic variables (covariates) is far greater than the number of subjects. It is also typical for such data sets to have a high proportion of censored observations. Methods based on univariate Cox regression are often used to select genes related to survival outcome. However, the Cox model assumes proportional hazards (PH), which is unlikely to hold for each gene. When applied to genes exhibiting some form of non-proportional hazards (NPH), these methods could lead to an under- or over-estimation of the effects.

In this thesis, we develop methods that will directly address the specific challenges noted above. First, we introduce several variable selection techniques that are able to accommodate various forms of NPH. Our first method

is based on the Kullback-Leibler (KL) divergence measure and Yang-Prentice (YP) model, from which we are able to create a statistical test and gene ranking measure that does not require an estimate of the baseline hazard. It includes the Cox PH and proportional odds (PO) models as special cases. Next, we propose a pseudo- R^2 measure derived from the partial likelihood and score statistic of various models. We show that it has a generalized form with the PH, PO, proportional log odds (PLO) and crossing hazards (CH) models as special cases. This R^2 measure does not require the estimation of model parameters and can be interpreted as the percentage of separability between the event and non-event groups, demonstrating both a computational and intuitive advantage. We also propose alternative R^2 measures, for the PH and PO models, that use the likelihood ratio and KL divergence measure. Lastly, we address the issues of censoring and high-dimensionality by proposing a supervised dimension reduction method for variable selection and survival prediction using continuum power regression (CPR) in conjunction with the generalized F or the semiparametric accelerated failure time (AFT) models. We evaluate the performance of our measures using extensive simulation studies and publicly available large-scale genomic datasets in ovarian and oral. Ultimately, we show that our proposed methods outperform existing methods and successfully address the issues of high-dimensionality, NPH and censoring.

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF FIGURES	ix
LIST OF TABLES	xi
1 INTRODUCTION	1
2 LITERATURE REVIEW	5
2.1 Censored Survival Data: Notation & Framework	5
2.2 An Overview of Survival Models	6
2.2.1 Cox Proportional Hazards (PH) Model	6
2.2.2 Assumption of the Cox PH Model	9
2.2.3 Linear Transformation Model	10
2.2.4 Comparing the Univariate Models	18
2.3 Existing Methods for Variable Selection and Ranking	19
2.3.1 Weighted Cox Regression	19
2.3.2 Concordance Regression (CON)	20
2.3.3 Test for Covariate Effect	22
2.3.4 R^2 -type Measures	23
2.3.5 Comparing the Methods for Variable Selection	26
2.4 Existing Methods for Supervised Dimension Reduction	27
2.4.1 Classic Approach	28
2.4.2 CPR: PCR and PLS-based Methods	29
2.4.3 Regularization and Parameter Shrinkage Procedures	34
2.4.4 Machine Learning Methods	39
2.4.5 Other Methods	41
2.4.6 Comparing Methods	45

3	EXAMPLES: MODEL FITTING AND VARIABLE SELECTION	47
3.1	Model Fitting	48
3.2	Subset Analysis Using a Weighted Average of Gene Expressions	52
3.3	Graphical Illustration of Time-Varying Effects in Individual Gene Expression	58
4	PROPOSED METHODS FOR VARIABLE SELECTION AND RANKING	64
4.1	Test for Gene Effect	65
4.1.1	General Framework: YP Model	66
4.1.2	Special Cases: PH and PO	71
4.1.3	\hat{I}_{PO} and \hat{I}_{YP} as Ranking Measures	73
4.1.4	R^2 Measures Using I_{PH} and I_{PO}	74
4.2	Pseudo- R^2 Measures	77
4.2.1	Generalized Pseudo- R^2 Measure	78
4.2.2	Computational Correction for R_{CH}^2 and R_{PLO}^2	84
4.2.3	Likelihood Ratio (LR) Based R^2 Measures for the PO Model	85
4.3	Simulations and Examples	87
4.3.1	Simulation Schemes	87
4.3.2	Simulation Results	92
4.3.3	Examples	99
5	PROPOSED METHODS FOR SUPERVISED DIMENSION REDUCTION	114
5.1	(A)CPRAFT	116
5.1.1	CPRAFT Algorithm	120
5.1.2	ACPRAFT Algorithm	120
5.2	A Flexible Parametric Approach to (A)CPRAFT	122
5.2.1	GF-based (A)CPRAFT Algorithm	122
5.3	A Semi-Parametric Approach to (A)CPRAFT	124
5.4	(A)CPRAFT as a Ranking Method	125
5.4.1	Simulation Results	127
5.4.2	Examples	138
5.5	Development of a Survival Prediction Model using (A)CPRAFT	150
5.5.1	(A)CPRAFT and Survival Prediction	151
5.5.2	Measures of Prediction Accuracy	154

5.5.3	Examples	154
6	CONCLUSION AND FUTURE WORK	161
6.1	Future Work	163
6.1.1	Applications	163
6.1.2	Censoring Variable and Covariate Dependence	164
6.1.3	Supervised Dimension Reduction Methods	164
	BIBLIOGRAPHY	166

LIST OF FIGURES

3.1	Venn Diagrams for Subset B and B-CNF. (a)-(c) Oral, (d)-(e) Ovarian	53
3.2	Survival Curves for Subset B, Ovarian	56
3.3	Survival Curves for Individual Genes: Oral	59
3.4	Survival Curves for Individual Genes: Ovarian	60
3.5	Survival Curves for 2 Gene Interactions	63
4.1	Top 500 Selected Genes, R^2 PO-based Measures	103
4.2	Top 500 Selected Genes, Other R^2 Measures	103
4.3	Top 500 Selected Genes, I_1, I_2 and Congreg Measures	106
4.4	Survival Curves for Dichotomized Weighted Average using I_{PO} , I_{YP} and I_{PH}	108
4.5	Top 500 Selected Genes, PO Measures	110
4.6	Individual Genes Selected by PO Measures	112
4.7	Two Gene Interaction	113
5.1	ROC Curves for Simulation Scheme 1	130
5.2	ROC Curves for Simulation Scheme 2, VIP vs. PLS Coeff for GF131	
5.3	ROC Curves for Simulation Scheme 2, GF vs. PLS_{unadj} using VIP	132
5.4	Survival Curves for Dichotomized VIP Weighted Average (a)-(c) Oral, (d)-(f) Ovarian	141
5.5	Selected Genes using $VIP > 1$	142
5.6	Survival Curves for Dichotomized CPR Coeff Weighted Average, Ovarian	144
5.7	Predicted Survival Curves for Weighted Averages using CPR Coefficients	145
5.8	Ovarian: Survival Curves for Dichotomized PLS Components .	146
5.9	Ovarian: Odds Curves for Dichotomized PLS Components . .	147
5.10	Oral: Survival Curves for Dichotomized PLS Components . . .	148

5.11 Predicted Survival Curves for Weighted Averages using PLS Components	149
--	-----

LIST OF TABLES

3.1	Univariate Analysis Results	49
3.2	GOF p -values for Weighted Averages, Subsets A	55
3.3	Ovarian GOF p -values for Weighted Averages Subset B-CNF	57
3.4	Oral GOF p -values for Weighted Averages Subset B and B-CNF	58
4.1	R^2 Measure: Special Cases	84
4.2	Scheme 1 Results, AUCs	94
4.3	Scheme 2 Results, AUCs	97
4.4	Youden Index, Scheme 1 & 2	99
4.5	R^2 Examples in the Oral Dataset	101
4.6	R^2 Examples in the Ovarian Dataset	101
4.7	R^2 Ranges in the Oral and Ovarian Datasets	102
4.8	GOF for Genes Selected using I Measures	109
5.1	GF Model: Special Cases	119
5.2	Simulation Scheme 1	131
5.3	AUC using VIP, Simulation Scheme 1	133
5.4	AUC using PLS Coeff, Simulation Scheme 1	133
5.5	Youden Index, Simulation Scheme 1	134
5.6	AUC using VIP, Simulation Scheme 2	135
5.7	AUC using PLS Coeff, Simulation Scheme 2	136
5.8	Youden Index, Simulation Scheme 2	137
5.9	GOF for Genes Selected using $VIP > 1$	143
5.10	Ovarian: Measures of Prediction Accuracy, Median Results	157
5.11	Oral: Measures of Prediction Accuracy, Median Results	158
5.12	RNA-Seq: Measures of Prediction Accuracy, Median Results	159

CHAPTER 1

INTRODUCTION

Significant advances have been made in genomics in the past two decades due to developments in high-throughput technologies. These include microarrays, SNP and methylation arrays, next-generation sequencing, proteomics and metabolomics. These technological developments have enabled the scientific community to monitor the expression levels of tens of thousands of genes and proteins, thus facilitating the generation of enormous amounts of data requiring analysis and interpretation (Elston and Spence 2006). With this new information, researchers can now attempt to understand and estimate the effects of specific genes on various diseases and characteristics associated with those diseases. In particular, one specific area of interest is studying the relationship between gene expression and time to death or recurrence of some disease. Being able to identify genes that are in some way predictive or related to a specific event of interest could help provide insight into treatment

and prognosis.

There are explicit challenges to the application of standard statistical methods to these high-dimensional data sets. Firstly, when dealing with time to death or recurrence of some disease, we often have censored survival times. Secondly, these data sets typically contain the expression levels of tens of thousands of genes, resulting in far more covariates than observations (i.e. $p \gg n$). A large number of covariates in conjunction with censored survival outcomes may cause issues. One of the most commonly used models in survival analysis is the Cox proportional hazards (PH) regression model which assumes that the hazard ratio is constant over time. In the univariate setting, it is unlikely that all genes will follow this PH assumption. Moreover, in the multivariate setting, dimension reduction techniques would need to be applied to reduce the large number of genes into a small, manageable number of components. These challenges separately are not new to the statistics community. There has been literature addressing censoring, dimension reduction, and alternatives to the Cox PH model. However, the analysis is unique in the sense that these challenges are now being combined in one problem.

The goal of this dissertation is to discuss various models and methods that can be used to link the gene expression to censored survival outcomes. In Chapter 2, we perform a literature review, which begins by describing the commonly-used PH model and the disadvantages associated with using it on these specific data sets. Then, we introduce models and variable selection methods found in the literature that have attempted to address the issues found with the PH model. We also discuss some dimension reduction tech-

niques to handle the high-dimensional $p \gg n$ scenario. In Chapter 3, we will apply the models discussed in Chapter 2 to two large-scale genomic data sets in order to illustrate the lack of suitability of the PH model and to demonstrate the need for more flexible models that can handle time-varying gene effects. Through this analysis, we will also identify specific models that fit a large proportion of the genomic variables and allow for non-proportional hazards (NPH). We then propose methods that address the issues of NPH, censoring and dimension reduction. In Chapter 4, we develop two types of variable selection methods, a test for gene effect and various pseudo- R^2 measures. The test for gene effect is based on Kullback-Leibler divergence using the Yang-Prentice (YP) model, a general model with the Proportional Odds (PO) and PH models as special cases. Using this approach, we develop statistical tests for gene effect as well as gene ranking measures. The pseudo- R^2 measures proposed in this chapter are derived using both the score statistics and likelihood ratios from the PO and Proportional Log Odds (PLO) models. The benefit of these methods is that they do not rely on the PH assumption and, instead, use the PO, PLO and YP models that allow for varying types of NPH. In Chapter 5, we discuss a supervised dimension reduction technique that has two-stage dimension reduction and model-fitting that can be used for both variable selection and survival prediction. This method incorporates the versatile dimension reduction technique known as continuum power regression (CPR), as well as a model fitting approach using a flexible parametric and semiparametric version of the accelerated failure time (AFT) model. In both chapters, we evaluate the performance of our measures using extensive sim-

ulation studies and publicly available large-scale genomic datasets. Lastly in Chapter 6, we conclude and discuss potential future work.

The ultimate goal of this dissertation is to develop variable selection and dimension reduction techniques that will specifically address the issues of censoring and NPH for large-scale genomic data.

CHAPTER 2

LITERATURE REVIEW

2.1 Censored Survival Data: Notation & Framework

In this section, we define the basic quantities, parameters, and notation used in modeling survival data. Let T be the time until some specific event, where the event may be death, recurrence or development of a disease, appearance of a tumor and so on. Common reasons for censoring could be that a subject does not experience the event before the study ends or withdraws from the study. Let C be the censoring time, where the event is observed only if it occurs prior to this time, and $\delta = I(T \leq C)$ be the indicator for whether the event has been observed. Here, $\delta = 0$ means the observation is censored and $\delta = 1$ means the event occurred. Because of censoring, it is often impossible to observe all true survival times, so we let $Y = \min(T, C)$ be the observed survival time. This observed value will give an estimate for the true

survival time, T . Three functions characterize the distribution of T :

Survival Function, $S(t) = P(T > t)$: probability of surviving beyond time t

Probability Density Function, $f(t) = \frac{-dS(t)}{dt}$: unconditional probability of event occurring at time t

Hazard Function, $\lambda(t) = \frac{f(t)}{S(t)}$: probability that an individual at time t experiences the event in the next instant.

Note, $S(t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function. Another useful quantity is the cumulative hazard function, denoted by $\Lambda(t) = \int_0^t \lambda(u)du = -\ln[S(t)]$. Later, $\lambda(t|\mathbf{z})$ will be referred to as the conditional hazard given covariate \mathbf{z} .

Typically, there are p covariates measured for each of n subjects in a study. Let \mathbf{Z} be the $n \times p$ matrix of covariates. In the following models, the general notation \mathbf{z} represents a p -vector of covariates. This notation and the functions described above will be used in the rest of this study.

2.2 An Overview of Survival Models

2.2.1 Cox Proportional Hazards (PH) Model

The Cox proportional hazards (PH) model is a semi-parametric regression model proposed by Cox (1972). The hazard rate, $\lambda(t|\mathbf{z})$, is defined as the instantaneous risk of an event at time t for covariate vector \mathbf{z} . The model

is given by

$$\lambda(t|\mathbf{z}) = \lambda_o(t) \exp(\boldsymbol{\beta}'\mathbf{z}), \quad (2.2.1)$$

where $t > 0$, $\lambda_o(t)$ is the baseline hazard function, and $\boldsymbol{\beta}$ is a vector of regression coefficients. The baseline hazard can be interpreted as the hazard rate of an individual who has a value of zero for all covariates in the model. The baseline hazard rate is the non-parametric part of the model, as it is unspecified and no assumptions are made about its form. Alternatively, a parametric form is assumed for the effect of the predictors on the hazard rate.

The survival function for the Cox PH model is

$$S(t|\mathbf{z}) = S_0(t)^{\exp \boldsymbol{\beta}'\mathbf{z}}, \quad (2.2.2)$$

and the probability density function has the form

$$f(t|\mathbf{z}) = \lambda_o(t) \exp(\boldsymbol{\beta}'\mathbf{z}) S_0(t)^{\exp \boldsymbol{\beta}'\mathbf{z}}. \quad (2.2.3)$$

Estimation for the coefficient $\boldsymbol{\beta}$ can be done by maximizing the log partial likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i \left\{ \mathbf{z}'_i \boldsymbol{\beta} - \log \left[\sum_{j \in R(t_i)} \exp(\mathbf{z}'_j \boldsymbol{\beta}) \right] \right\}, \quad (2.2.4)$$

where t_i is the survival time for subject i , δ_i is the censoring indicator, and $R(t_i)$ is the risk set at time t_i .

The Cox PH model is the most commonly used model in survival anal-

ysis. This stems from its intuitive and simplistic structure, and it provides a convenient method for summarizing gene effects by estimating the model parameters with log partial likelihood maximizations. In fact, the Cox PH model has been applied to many high-dimensional genomic data sets with the purpose of relating gene expression to censored survival time. For example, Beer *et al.* (2002) was one of the first to combine gene expression with survival where they applied a univariate Cox PH model to determine which of the 4,966 genes could be used to predict patient survival in early stage lung adenocarcinoma, a form of lung cancer. Van der Net *et al.* (2008) and Tothill *et al.* (2008) also applied the Cox PH model to their large genomic data sets to determine the relationship between gene expression and time to coronary heart disease and ovarian cancer survival, respectively. Goeman *et al.* (2005) also used the PH model to test the association of a genomic pathway with survival.

There have also been several papers that explore the relationship between gene expression and survival in oral cancer. Saintigny *et al.* (2011) studied the gene expression profiles of 86 patients and applied a univariate Cox regression model on each gene to identify which genes had a significant relationship with oral cancer development. They also performed a multivariate analysis using an algorithm based on boosting, which constructs a prognostic model by maximizing the partial log likelihood function that imposes a penalty for each non-zero coefficients utilized in the model. Similarly, Mendez *et al.* (2009) used multivariate Cox regression to determine the genes associated with oral cancer, and Xu *et al.* (2010) also used the model to aid in their research dealing with the relationship between gene expression and the

initiation of lymphotropism, an important prognostic factor of 5 year survival in oral cancer. Rosenwald *et al.* (2002) also used the PH model in their study on subjects with diffuse large-B-cell lymphoma, and Geisler *et al.* (2002) used it to estimate hazard ratios for two specific proteins in subjects with head and neck cancer.

2.2.2 Assumption of the Cox PH Model

The PH model has a very important assumption that is often overlooked when the model is being used. Simply, it assumes proportional hazards. In other words, it assumes that the hazard ratio is constant over time. The hazard ratio corresponding to covariate values \mathbf{z} and \mathbf{z}^* is

$$\frac{\lambda(t|\mathbf{z})}{\lambda(t|\mathbf{z}^*)} = \frac{\lambda_o(t) \exp(\boldsymbol{\beta}'\mathbf{z})}{\lambda_o(t) \exp(\boldsymbol{\beta}'\mathbf{z}^*)} = \frac{\exp\left[\sum_{i=1}^p \beta_i z_i\right]}{\exp\left[\sum_{i=1}^p \beta_i z_i^*\right]} = \exp\left[\sum_{i=1}^p \beta_i (z_i - z_i^*)\right]. \quad (2.2.5)$$

We can clearly see that the above ratio is a constant, only depending on the covariate \mathbf{z} and not on time.

As stated above, this assumption is not validated, which is especially a concern in large-scale genomic data sets, where it is unlikely that the proportional hazards assumption will hold for each of the thousands of genes. Instead, we would expect some of the genes to express a time-dependent effect on survival, meaning that their effect changes over time. We may see genes that exhibit different forms of non-proportional hazards (NPH), such as converging, diverging or crossing hazards. In fact, Geisler *et al.* (2002) who

used the PH model to estimate the relationship between two specific genes and time to mortality and locoregional recurrence in neck cancer, found that the PH assumption did not hold for one of their genes. Thus, the results related to that gene were not reliable.

The PH model has no mechanism to deal with NPH, and applying the PH model to data that does not follow the PH assumption can lead to inaccurate conclusions and an either over or under-estimate of the effects. In other words, some significant genes will be missed, while other genes will be falsely identified as significant. Using the PH model will not allow for the identification of genes that exhibit different forms of hazards, whether proportional or non-proportional.

The issue of NPH has received very little attention in the context of large-scale genomic data (Xu *et al* 2005; Dunkler *et al.* 2010; Rouam *et al.* 2010). This dissertation attempts to develop an argument for the importance of research in this area. It will first discuss some alternatives to the Cox PH model already found in the literature. It will then focus on models that can be used when the Cox PH model does not fit and describe methods for handling all types of hazards. It will also use real data examples to validate the need for considering the possibility of genes showing NPH.

2.2.3 Linear Transformation Model

This section describes multiple alternatives to the Cox PH model falling under the general class of linear transformation model. While these linear transformation models have been explored in a few previous studies,

most of the focus has been on the PH and PO models and do not use an application involving high-dimensional genomic data. Cheng et al. (1995) described a class of these models and derived inference procedures to examine the effects of the covariate with censored observations. Cheng et al. (1997) and Chen et al. (2002) also discuss procedures for estimation in the linear transformation model and apply their method to lung cancer data sets. Xu et al. (2005) were one of the first to investigate the model for microarray data.

In this sub-section, we will first describe the linear transformation model and briefly show that the Cox PH model is a special case. We will then describe specific individual cases, where the models have an advantage over the Cox PH model because they do not require the PH assumption.

General Linear Transformation Model The transformation model is a general model with many benefits and special cases. Let T be the survival time and \mathbf{Z} be the $n \times p$ covariate matrix. Then, the transformation model relating T to covariate vector \mathbf{z} is defined as

$$h(T) = -\boldsymbol{\beta}'\mathbf{z} + \epsilon, \tag{2.2.6}$$

where $h(T)$ is a monotone transformation function, $\boldsymbol{\beta}$ is a p -vector of regression coefficients, and ϵ is the error term with some known distribution. The benefit of this framework lies in its flexibility. Choosing different distributions for ϵ can create a variety of models. Some of these special cases are described below.

Cox PH Model Recall that in the Cox PH model $S(t|\mathbf{z}) = S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{z})}$. This model can be written as $\log[-\log S(t|\mathbf{z})] = \log[\Lambda_0(t)] = h(t) + \boldsymbol{\beta}'\mathbf{z}$, which has the generalization

$$g\{S(t|\mathbf{z})\} = h(t) + \boldsymbol{\beta}'\mathbf{z}, \quad (2.2.7)$$

where $h(t)$ is an unspecified strictly increasing function and $g\{\cdot\}$ is some decreasing function. Note, this is equivalent to the linear transformation model in 2.2.6. Here, the random error ϵ has distribution function $F = 1 - g^{-1}$, where F is the unit extreme value distribution. As noted in the previous section, this model is widely used because of its simple form, but it does not allow for NPH.

Proportional Odds Model The PO model is another special case of the linear transformation model, where $g\{\cdot\} = -\text{logit}(\cdot)$ in 2.2.7, where $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$. To demonstrate this, the PO model must be described first.

This model does not assume PH and, instead, allows for NPH. However, it assumes that the effect of the covariates will proportionately increase or decrease the odds of dying or recurrence at time t . The PO model is given by

$$\frac{F(t|\mathbf{z})}{1 - F(t|\mathbf{z})} = \frac{F_o(t)}{1 - F_o(t)} \exp(\boldsymbol{\beta}'\mathbf{z}) = \frac{1 - S_o(t)}{S_o(t)} \exp(\boldsymbol{\beta}'\mathbf{z}) = \frac{1 - S(t|\mathbf{z})}{S(t|\mathbf{z})}, \quad (2.2.8)$$

where $F_o(t)$ and $S_o(t)$ are some baseline cumulative and survival distributions. The value $\exp(\boldsymbol{\beta}'\mathbf{z})$ is the multiplier that quantifies the amount of proportionate increase or decrease in the odds associated with the covariate \mathbf{z} . From

2.2.8, the survival function can be calculated as

$$S(t|\mathbf{Z}) = \frac{S_0(t)}{S_0(t) - S_0(t) \exp(\boldsymbol{\beta}'\mathbf{z}) + \exp(\boldsymbol{\beta}'\mathbf{z})}. \quad (2.2.9)$$

From this, it can be seen that $-\text{logit}\{S(t|\mathbf{z})\} = h(t) + \boldsymbol{\beta}'\mathbf{z}$, and thus the PO model is also a special case of the linear transformation model. Here, the random error ϵ has distribution function $F = 1 - g^{-1}$, where F is the standard logistic distribution.

Proportional Log Odds Model The Proportional Log Odds (PLO) model introduces another parameter, γ , into the PO model and has the form

$$\frac{F(t|\mathbf{z})}{1 - F(t|\mathbf{z})} = \left[\frac{F_0(t)}{1 - F_0(t)} \right]^{\exp(\boldsymbol{\gamma}'\mathbf{z})} \exp(\boldsymbol{\beta}'\mathbf{z}) = \frac{1 - S(t|\mathbf{z})}{S(t|\mathbf{z})} = \left[\frac{1 - S_0(t)}{S_0(t)} \right]^{\exp(\boldsymbol{\gamma}'\mathbf{z})} \exp(\boldsymbol{\beta}'\mathbf{z}), \quad (2.2.10)$$

where $F_0(t)$ and $S_0(t)$ are some baseline cumulative and survival distributions. Note that, similar to the PO model, this can be expressed as $-\text{logit}\{S(t|\mathbf{z})\} = \exp(\boldsymbol{\gamma}'\mathbf{z})h(t) + \boldsymbol{\beta}'\mathbf{z}$, and thus the PLO model is related to the linear transformation model. This model also allows for NPH. From 2.2.10, the survival function can be calculated as

$$S(t|\mathbf{z}) = \frac{1}{\exp(\boldsymbol{\beta}'\mathbf{z}) \left[\frac{1 - S_0(t)}{S_0(t)} \right]^{\exp(\boldsymbol{\gamma}'\mathbf{z})} + 1}. \quad (2.2.11)$$

Yang-Prentice Model Next, we look at a generalization of both the PH and PO models. The Yang-Prentice (YP) model was described in Yang and

Prentice (2005) and has hazard and survival functions

$$\lambda(t|\mathbf{z}) = \frac{\lambda_0(t) \exp [(\boldsymbol{\beta} + \boldsymbol{\gamma})'\mathbf{z}]}{\exp(\boldsymbol{\beta}'\mathbf{z}) - S_0(t) \exp(\boldsymbol{\beta}'\mathbf{z}) + S_0(t) \exp(\boldsymbol{\gamma}'\mathbf{z})} \quad (2.2.12)$$

and

$$S(t|\mathbf{z}) = \left(\frac{S_0(t) \exp(\boldsymbol{\gamma}'\mathbf{z})}{\exp(\boldsymbol{\beta}'\mathbf{z}) - S_0(t) \exp(\boldsymbol{\beta}'\mathbf{z}) + S_0(t) \exp(\boldsymbol{\gamma}'\mathbf{z})} \right)^{\exp(\boldsymbol{\gamma}'\mathbf{z})}, \quad (2.2.13)$$

respectively. When $\boldsymbol{\gamma} = \boldsymbol{\beta}$, it becomes the PH model, and when $\boldsymbol{\gamma} = 0$, it becomes the PO model. Thus, it is a versatile and useful model that encompasses both the PH and PO models and allows for varying types of hazards, including both PH and NPH. However, this model has its own limitation because, currently, it has only been implemented for the dichotomized covariate case.

The Accelerated Failure Time (AFT) Model The AFT model is another special case of the linear transformation model, where $h(T) = \log(T)$ in Equation 2.2.6. This model differs from the PH model in the sense that it measures the direct effect of the covariate on survival time instead of the hazard. It is a censored linear regression model that represents the relationship between covariates and log time and has the form

$$Y = \ln T = \boldsymbol{\mu} + \boldsymbol{\beta}'\mathbf{z} + \boldsymbol{\sigma}\epsilon, \quad (2.2.14)$$

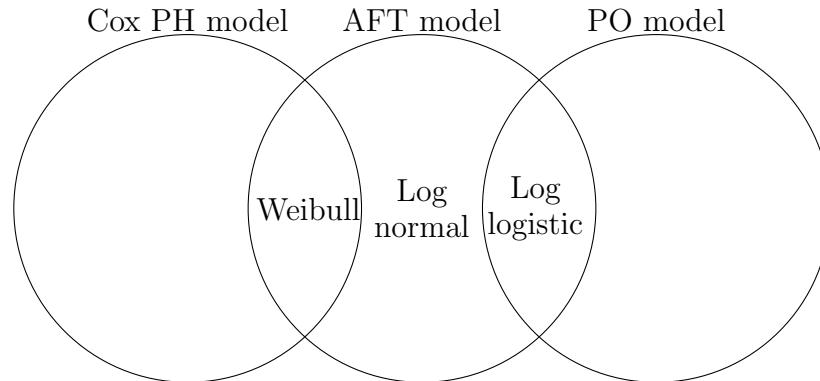
where $\boldsymbol{\mu}$ is the intercept, $\boldsymbol{\beta}'$ is a vector of regression coefficients, $\boldsymbol{\sigma}$ is a

scale parameter and ϵ is the error term that has a certain distribution. This model is equivalent to the linear transformation model in Equation 2.2.6, with $h(T) = \log(T)$. The linear form of this model makes it easy to interpret, as the regression coefficients are interpreted similar to those in multiple linear regression. Unlike the PH model, it also allows for NPH. A rank-based inference for the model is described in Jin *et al.* (2003), and a regularized estimation procedure is described in Cai *et al.* (2009).

This model also has the benefit of having many special cases. In the parametric AFT model, we will look at four specific distributions for ϵ , log normal, log logistic, Weibull and generalized F (GF). The GF model is a broad, flexible parametric model that includes the other distributions as special cases.

- **Log Logistic:** If $\epsilon \sim$ standard logistic, then $T \sim$ log logistic.
- **Log Normal:** If $\epsilon \sim$ standard normal, then $T \sim$ log normal.
- **Weibull:** If $\epsilon \sim$ EV (extreme value), then $T \sim$ Weibull.
- **Generalized F :** The distribution of ϵ is described in Chapter 5.

There is some overlap between the parametric models described above, which can be seen in the following diagram.



The semiparametric AFT model is similar to the parametric AFT model in that it takes on the same form and relates log time to the covariates. However, in the semiparametric case the error term is completely unspecified. An iterative solution was developed to estimate the regression parameters for this model (Jin *et al.* 2006). The procedure is based on the least-squares principle, while taking censoring into account, but a disadvantage of this procedure is that it is computationally slow. On the other hand, the GF model is computationally efficient and, thus, offers a flexible alternative. The GF model is described in more detail below.

Generalized F (GF) Model Another model of interest is the generalized F model discussed by Ciampi *et al.* (1986) and more recently by Cox (2008). Both papers discuss the use and benefits of using the generalized F model for parametric survival analysis. Its benefit lies in its umbrella structure, where the choice of various parameters can turn this model into Weibull, log-normal, log-logistic, and many other parametric AFT models. Its distribution has the form seen in Equation 5.1.2.

The PO, PLO, YP, and AFT models were chosen because they do not

assume PH and allow for various types of NPH. However, while the goal of this study is to find methods that can handle NPH, the ideal situation is to create a general framework under which the Cox PH model and other models can be chosen, after checking to see if the PH assumption holds. The linear transformation model proves to be a useful tool in this setting because of its flexibility.

Semi-parametric Generalization of the Cox PH Model Devarajan and Ebrahimi (2011) directly address the issue of NPH with their semi-parametric generalization of the Cox PH model. Unlike the PH model, their model allows for crossing hazards. Its form involves raising the cumulative hazard function to a power. The survival and cumulative hazard functions are, respectively,

$$\Lambda(t|\mathbf{z}) = e^{\beta'\mathbf{z}}\{\Lambda_0(t)\}^{\exp(\gamma'\mathbf{z})} \text{ and } S(t|\mathbf{z}) = \exp\{-e^{\beta'\mathbf{z}}[-\log S_0(t)]^{\exp(\gamma'\mathbf{z})}\}. \quad (2.2.15)$$

We can then apply a $\log(-\log)$ transformation to the survival function in 2.2.15. Doing this, we get $\log\{-\log\{S(t|\mathbf{z})\}\} = \beta'\mathbf{z} + \exp(\gamma'\mathbf{z}) \log\{-\log\{S_0(t)\}\} = \beta'\mathbf{z} + \exp(\gamma'\mathbf{z})h(t)$. This is similar to the PLO model, but here $g\{\cdot\} = \log(-\log)$ in equation 2.2.7. Thus, this model is also related to the linear transformation model. Because the partial likelihood approach cannot be applied, Devarajan and Ebrahimi (2011) use the full likelihood approach via a cubic B-spline approximation for the baseline hazard to estimate the model parameters. Penalized likelihood estimation is described in Devarajan and Ebrahimi (2013), and a goodness of fit test is described in Devarajan and Ebrahimi (2002). The crossing hazards (CH) model used in Rouam *et al.* (2011) is a special case of this model when $\gamma = \beta$.

Hypertabastic Survival Model Tabatabai *et al.* (2012) use a new proportional hazards model called the hypertabastic survival model. Although it is a parametric PH model, it resulted in better goodness-of-fit results compared to the standard PH and AFT models. Its hazard function has the form

$$h(t|\mathbf{z}, \theta) = h_0(t)g(\mathbf{z}|\theta), \quad (2.2.16)$$

where $h_0(t) = \alpha [t^{2\beta-1} \operatorname{csch}^2(t^\beta) - t^{\beta-1} \operatorname{coth}(t^\beta)] \tanh[W(t)]$ and $W(t) = \alpha [1 - t^\beta \operatorname{coth}(t^\beta)] / \beta$. One valuable feature is that its hazard function has the ability to assume many different shapes, unlike some standard AFT models.

2.2.4 Comparing the Univariate Models

In this section, we were able to introduce some viable alternatives to the Cox PH model, where most of the models we introduced allowed for varying types of NPH. The PO model is particularly useful because it allows for both PH and NPH, but it does assume the odds are proportional. The YP model is useful due to its generalized form that encompasses both the PO and PH model, but it has a clear disadvantage in its implementation as it can only be applied to dichotomized covariates. Alternatively, the AFT model has an advantage in its interpretation because of its simple linear form. The semiparametric version does not require a distribution for the error term. However, this semiparametric AFT model's current implementation is computationally slow. Thus, the GF model offers a flexible alternative with computational efficiency. While there is not one perfect model, each of the proposed univariate models provide certain benefits for dealing with the possibility of NPH in genomic data.

In large-scale genomic studies, the number of genes, p , far exceeds the

number of observations, n , (i.e. $p \gg n$). In §2.3 and Chapter 4, we will be dealing with univariate models involving individual gene j , for $j = 1, \dots, p$. Each of the models described above can be applied in this setting to an individual gene. The expression of gene j for individual i is denoted by z_{ij} . We will use these univariate models as foundations for the proposed measures in Chapters 4 and 5.

2.3 Existing Methods for Variable Selection and Ranking

To study the relationship between gene expression and survival in large-scale genomic data sets, the most common method has been to use the Cox PH model. However, as discussed in the previous section, if the PH assumption does not hold, then it is not appropriate to use the Cox PH model. An alternative method must be found, such as adding an interaction term between the covariates and time, using another model that allows for various types of hazards, using separate models for different time periods, and so on. This section will discuss some specific methods that have been explored so far in the literature.

2.3.1 Weighted Cox Regression

Some of the proposed methods use the Cox PH model as a foundation but with some necessary changes to address potential NPH. Schemper *et al.* (2009) uses a weighted Cox regression, a method that is independent of the type of non-proportionality. This method does not introduce other parameters, but instead introduces weights into the score function of Cox's log partial likelihood. The weights

used are defined as

$$w(t_i) = S(t_i)G(t_i)^{-1}, \quad (2.3.1)$$

where $S(t_i)$ is the Kaplan-Meier (KM) estimate of the survival function at time t_i and $G(t_i)$ is the probability that a subject is under follow-up at time t_i . These weights allow for the estimation of $\hat{\beta}_w$, interpreted as the log odds of concordance, with the critical point being that these approximations are independent of the type of non-proportionality. At the time the paper was published, they had not applied the method to any gene expression data, but Dunkler *et al.* (2010) applied the weighted Cox analysis to three large genomic data sets. They demonstrated that it performed well in estimating the prediction model without having to assume PH, but it had a disadvantage of seeming to favor converging hazards.

2.3.2 Concordance Regression (CON)

Dunkler *et al.* (2010) proposed a method called concordance regression (CON) to select genes that are related to survival irrespective of the type of hazard. They use a generalized concordance probability as a measure of the effect size. The basic concordance probability is

$$c = P(T_1 < T_0) \quad (2.3.2)$$

where T_1 is a randomly chosen survival time from group 1 and T_0 is a randomly chosen survival time from group 0. This probability is interpreted as the probability that a randomly chosen subject from group 1 dies or experiences the event of interest before a randomly chosen subject from group 0. This probability is independent of the PH assumption.

First, they define c' , a generalization of c to continuous data, which has the form

$$c' = \frac{\exp(\gamma)}{1 + \exp(\gamma)}, \quad (2.3.3)$$

where γ are the log odds that the survival time decreases if the gene expression is doubled. Then, they model c' through its log odds by

$$P(T_i < T_j | x_i > x_j) = \frac{\exp(x_i \gamma)}{\exp(x_i \gamma) + \exp(x_j \gamma)}, \quad (2.3.4)$$

where x_i is equivalent to gene expression z_i in our application. They compute its log-likelihood and derivative as

$$\ell(\gamma) = \sum_{(i,j)} \{x_i \gamma - \log [\exp(x_i \gamma) + \exp(x_j \gamma)]\} \quad (2.3.5)$$

and

$$\frac{\partial \ell(\gamma)}{\partial \gamma} = \sum_{(i,j)} \left[x_i - \frac{x_i \exp(x_i \gamma) + x_j \exp(x_j \gamma)}{\exp(x_i \gamma) + \exp(x_j \gamma)} \right], \quad (2.3.6)$$

respectively, where the summation is over all risk pairs (i, j) , such that $t_i < t_j$. This model can be seen as a conditional logistic regression where the dependent variable is the concordance of the risk pair (i, j) . To estimate the log odds γ , the derivative of the log likelihood can be set to zero and solved. Once $\hat{\gamma}$ is computed, c' can be estimated by

$$\hat{c}' = \frac{\exp(\hat{\gamma})}{1 + \exp(\hat{\gamma})}. \quad (2.3.7)$$

When t_i is censored, it is not clear if $t_i < t_j$. Thus, those risk pairs were omitted. Because of this, the risk pairs were weighted to make up for this possible over-representation of some subjects. These weights are described in Dunkler *et al.* (2010). They then define the absolute effect size as $\hat{c}'_+ = .5 + |\hat{c}' - .5|$ and the genes

can be ranked based on that value.

They showed that when some of the genes showed a time-dependent effect on survival, CON produced the least biased and most stable estimates compared to the Cox PH model. In fact, despite the type of hazards present (proportional, converging or diverging), CON selected the most correct genes (genes having an effect on survival) compared to both weighted Cox regression (Schemper 2009) and the Cox PH model. However, one drawback of this method is that the quantity used for ranking lacks an intuitive interpretation.

2.3.3 Test for Covariate Effect

Devarajan and Ebrahimi (2009) developed a method to test the covariate effect in the Cox PH model using Kullback-Leibler (KL) divergence and Renyi's information measure. Unlike the other methods discussed, this does not address the issue of NPH. However, their tests are easy to compute and perform better than other statistical measures. Another advantage is that these methods are invariant under any non-singular transformation of the failure time data. In Chapter 4, we will create a new method based on this idea using both the YP and PO models, which do not rely on the PH assumption.

In general, KL divergence is defined as $I(F : F_0) = \int_0^{\infty} f(x) \log \left\{ \frac{f(x)}{f_0(x)} \right\} dx$, where F and F_0 are the distribution functions. This measure is also called the directed divergence as it measures the discrepancy between F and F_0 in the direction of F . Alternatively, one can define $I(F_0 : F) = \int_0^{\infty} f_0(x) \log \left\{ \frac{f_0(x)}{f(x)} \right\} dx$, which measures the discrepancy between F and F_0 in the direction of F_0 . Lastly, these two measures can be combined to quantify the difficulty of discriminating between F and F_0 , by defining $J(F : F_0) = I(F : F_0) + I(F_0 : F)$.

Devarajan and Ebrahimi (2009) test the hypotheses $H_0 : \boldsymbol{\beta} = 0$ against $H_1 : \boldsymbol{\beta} \neq 0$ by defining f as $f(t|\mathbf{z})$ under H_1 and f_0 as $f_0(t)$ under H_0 . Then,

$$I_1(F : F_0|\mathbf{z}) = \int_0^{\infty} f(t|\mathbf{z}) \log \left\{ \frac{f(t|\mathbf{z})}{f_0(t)} \right\} dt, \quad (2.3.8)$$

$$I_2(F_0 : F|\mathbf{z}) = \int_0^{\infty} f_0(t) \log \left\{ \frac{f_0(t)}{f(t|\mathbf{z})} \right\} dt, \quad (2.3.9)$$

and

$$J(F : F_0|\mathbf{z}) = I_1(F : F_0|\mathbf{z}) + I_2(F_0 : F|\mathbf{z}). \quad (2.3.10)$$

These measures are derived using $f(t|\mathbf{z})$ from the PH model, and then \hat{I}_1, \hat{I}_2 , and \hat{J} are obtained by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$, the maximum partial likelihood estimates. From these, the following statistics are used to test for covariate effect:

- Using \hat{I}_i , reject H_0 if $\frac{\hat{I}_i^2}{\widehat{Var}(\hat{I}_i)} > \chi_p^2, i = 1, 2$,
- Using \hat{J} , reject H_0 if $\frac{\hat{J}^2}{\widehat{Var}(\hat{J})} > \chi_p^2$.

These measures outperformed existing methods in terms of power specifically with smaller sample sizes. For larger sample sizes, the tests performed about the same. We also note that this measure is easy and relatively quick to compute, and it does not require an estimate of the baseline hazard.

2.3.4 R^2 -type Measures

Pseudo R^2 Measures

Rouam *et al.* (2010, 2011) developed pseudo- R^2 measures for gene selection that rely on the partial likelihood function of a particular model and is based on

the score statistic. In general, the partial likelihood is the joint probability of all realized events conditioned upon the existence of events at those times. In their 2010 paper, they create the R^2 measure based on the Cox PH model, and in the 2011 paper, they extend the same idea to a semiparametric crossing hazards model. The calculation of their measures is relatively easy because it does not require the estimate of the parameter β in the models. Also, the R^2 measure itself has a straightforward interpretation and represents the percentage of separability between the event and non-event groups over time. We will now discuss the two measures.

R^2 for the Cox PH Model Rouam *et al.* (2010) consider the derivative of the log partial likelihood as the score function, denoted by U_i , under the null hypothesis ($\beta = 0$). Using the Cox partial likelihood described in equation 2.2.4, they show that an estimate for the score function can be written as

$$\hat{U}_i = \delta_i \hat{w}(t_i) \left(z_i - \frac{\sum_{j=1}^n Y_j z_j}{\sum_{j=1}^n Y_j} \right), \quad (2.3.11)$$

where $\hat{w}(t_i) = 1$. They then define a closely related quantity W_i , which are robust scores that are similar to U_i but independent. The pseudo- R^2 measure is then calculated as

$$R_{PH}^2 = \frac{1}{k} \frac{\left(\sum_{i=1}^n \hat{W}_i \right)^2}{\sum_{i=1}^n \hat{W}_i^2}, \quad (2.3.12)$$

where k is the number of distinct uncensored failure times, a quantity that falls between 0 and 1. Because this measure is based on the PH model, it has the disadvantage of not allowing for NPH.

R^2 for the Crossing Hazards (CH) Model Rouam *et al.* (2011) follow a similar process as the 2010 paper, but instead use a CH model with the goal of detecting genes that specifically exhibit CH. This model is a special case of the generalized Cox model described in Devarajan and Ebrahimi (2011) in Equation 2.2.15, but here they set $\gamma = \beta$. Their semi-parametric crossing hazards model has survival function

$$S_i(t|z_i) = \exp \left\{ - \left(\int_0^t \lambda_0(s) ds \right)^{e^{\beta z_i}} \right\} \quad (2.3.13)$$

for subject i with covariate z_i , where $\lambda_0(t)$ is the baseline hazard function and β the regression parameter. For $i = 1 \dots n$, the hazard function for subject i is defined as

$$\lambda_i(t|z_i) = \lambda_0(t) e^{\beta z_i} \Lambda_0(t)^{(e^{\beta z_i} - 1)}, \quad (2.3.14)$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u) du$. Using this model, they calculate the score function which has the same form as seen in equation 2.3.11, but for this model $\hat{w}(t_i) = (1 + \log\{\hat{\Lambda}_0(t_i)\})$. $\hat{\Lambda}_0(t_i)$ is estimated by the left-continuous version of Nelson's estimator. The R_{CH}^2 measure has the same form depicted in equation 2.3.12. When they applied the method to lung cancer data, it was able to detect genes that were related to tumor evolution but were not selected under the PH assumption. Unlike their previous R^2 measure, this method does not require proportional hazards, but it still has a disadvantage. The model itself forces crossing hazards and does not allow for other types of NPH. We would not expect every gene to exhibit crossing hazards, so this presents similar issues to what we face with the Cox PH-based methods, where the method is too strict on the types of hazards present. This measure also has a computational issue that we address in Chapter 4.

Other R^2 -type Measures

Several other indices have been proposed in the literature, most of which are based on the Cox PH model. Allison's index (1995), ρ_N^2 , is based on a transformation of the log partial likelihood ratio test. O'Quigley *et al* (2005) proposed a modified version of Allison's index, ρ_k^2 . Nagelkerke (1991) also proposed a modification of Allison's index, R_N^2 , dividing it by its maximum value. Xu and O'Quigley (1999) developed a measure, ρ_{XOQ}^2 , based on a transformation of the Kullback-Leibler distance between the null and the alternative models. These four measures are, respectively,

$$\rho_N^2 = 1 - \exp\left(-\frac{2}{N} \times [\log L(\hat{\beta}) - \log L(0)]\right),$$

$$\rho_k^2 = 1 - \exp\left(-\frac{2}{k} \times [\log L(\hat{\beta}) - \log L(0)]\right), \text{ where } k \text{ is the number of failures,}$$

$$R_N^2 = \frac{\rho_N^2}{R_{max}^2}, \text{ where } R_{max}^2 = 1 - \exp\left(\frac{2}{N} \times \log L(0)\right), \text{ and}$$

$\rho_{XOQ}^2 = 1 - \exp\left\{-\Gamma(\hat{\beta}) / \sum_{j=1}^N V(t_j)\right\}$, where $V(t_j) = S_j^+(t_j) - S_j(t_j)$ and \hat{S} is the Kaplan-Meier estimator. Note, $\hat{\Gamma}(\hat{\beta})$ is derived from twice the KL distance between the null model ($\beta = 0$) and the model taking covariates into account ($\beta \neq 0$). See Xu and O'Quigley (1999) for details.

These measures have a disadvantage because they do not address NPH. In Chapter 4, we propose similar R^2 measures based on the PO model.

2.3.5 Comparing the Methods for Variable Selection

The methods discussed in this section each have their own advantages and disadvantages. Some address the issue of NPH, while others are based on the Cox PH

model but have potential for extensions to other models that allow for varying types of hazards. For example, the test for covariate effect proposed by Devarajan and Ebrahimi (2009) has a computational advantage of not requiring an estimate of the baseline hazard, but it is based on the Cox PH model and has not been applied in a genomic setting. Alternatively, Rouam’s R^2 measures require the estimate of the baseline hazard, but they do not require an estimate of the model parameter β . Rouam *et al.* (2011) also attempted to address the issue of NPH by introducing an R^2 measure based on the crossing hazards model, but their method forces crossing hazards so it is not flexible in the types of hazards present. Lastly, while CON addresses the issue of NPH, it lacks the intuitive interpretation of both the R^2 and test for covariate effect measures. Thus, in Chapter 4, we will propose two general frameworks for the R^2 and test for covariate effect measures. These proposed methods extend the existing measures to the PO, PLO and YP models that allow for varying types of NPH and also include the existing measures as special cases.

2.4 Existing Methods for Supervised Dimension Reduction

As stated in the introduction, the goal of this proposal is to address the issues of NPH in high-dimensional genomic data in terms of both variable selection and dimension reduction. The methods in the previous section focused on variable selection or marginal screening, meaning each method could be applied to each gene separately, resulting in a single quantity for each of the p genes. They did not address the issue of high-dimensionality. In this section, we review existing methods that specifically address the $p \gg n$ issue through dimension reduction. Each method has one of two

goals: gene selection or survival prediction (from building a prognostic model).

2.4.1 Classic Approach

Univariate Model and Multiple Testing The most basic way to identify genes associated with survival is to use the classical approach of fitting the univariate Cox PH model to each gene. Essentially, the null hypothesis $H_0 : \beta = 0$ hypothesis is tested for each gene, resulting in p p -values, which can then be adjusted using some multiple testing procedure, such as Benjamini-Hochberg. These multiple testing procedures are meant to reduce the risk of a Type I error, but in the genomic setting it often gives poor results, as explained in Dudoit *et al.* (2003). Instead of using p -values, the classical approach can also be used to calculate the score statistic, or Cox Score, for each gene. The score statistic is

$$S_j = \left(\frac{dl(0)}{d\beta_j} \right) / \sqrt{\frac{d^2l(0)}{d\beta_j^2}}, \quad (2.4.1)$$

where $l(0)$ is the log partial likelihood in Equation 2.2.4 when $\beta = 0$. These S_j values can be ranked, where the larger score values represent a relationship with survival (i.e., reject H_0). Beer *et al* (2002) used this method to identify significant genes in an early-stage lung adenocarcinoma study.

Another logical extension to the univariate selection method is to use the classic forward stepwise selection procedure. This takes into account correlation between genes, which the univariate procedure discussed above neglects. Here, the most significant gene is selected first using the univariate score test described above, and then models including this gene and the remaining $p - 1$ genes are tested and compared to the model with the first selected gene. The gene from the most significant model here is ranked as number two, and now the first and second choice genes

are included in a multivariable PH model. This process continues until the desired number of genes are selected.

Significance Analysis of Microarrays (SAM) To improve the use of the Cox score, Tusher *et al.* (2001) proposed a procedure called Significance Analysis of Microarrays (SAM) that uses a modified score based on the change in gene expression relative to the standard deviation of observations. Essentially, SAM assigns a score to each gene, and genes with a score greater than a certain threshold are deemed significant. Their score can be adapted to censored survival data, where they define it in terms of Cox's proportional hazards function, in which some of the patients remain alive or are lost to follow-up at the time of the study. Once their measure is calculated, genes are ranked based on this value. They found that this modified Cox PH score performed better than the basic Cox PH score. One obvious disadvantage to this method is that it is based on the Cox PH model and, therefore, does not address the potential for NPH.

2.4.2 CPR: PCR and PLS-based Methods

We now draw our attention away from the classic univariate approach and focus on the most common types of methods in the literature. To deal with high-dimensionality, many existing methods use the PH model in combinations with a dimension reduction method. We will first describe these dimension reduction techniques separately and then discuss methods that combine this with survival analysis.

Continuum Regression and CPR

Some of the more familiar dimension reduction methods are principal components regression (PCR) and partial least squares regression (PLS). Sundberg (2002) dis-

cusses continuum regression (CR), a regularization method that includes ordinary least squares (OLS), PLS and PCR as special cases with the simple choice of one parameter. In general, CR deals with the collinearity issue, which means that there may be approximate linear relationships between various predictors, or genes in our application. The ultimate goal is to reduce the dimension of the data by creating linear combinations of variables called regressors or factors, say $\mathbf{v} = \mathbf{X}\mathbf{w}$, where \mathbf{X} is the covariate matrix, that satisfy certain conditions. In CR, the rule is to maximize the function

$$g_\gamma(R^2(\mathbf{v}, \mathbf{y}), \text{Var}(\mathbf{t})) = R^2(\mathbf{v}, \mathbf{y})\text{Var}(\mathbf{v})^\gamma \quad (2.4.2)$$

over directions \mathbf{w} , where $|\mathbf{w}| = 1$, $R(\mathbf{t}, \mathbf{y})$ is the correlation between survival time \mathbf{y} and factor \mathbf{v} , and $\gamma \geq 0$. Choice of the parameter γ leads to the following special cases: OLS ($\gamma = 0$), PLS ($\gamma = 1$), and PCR ($\gamma \rightarrow \infty$). An alternative parameter, α , can also be defined as $\gamma \equiv \alpha/(1 - \alpha)$, where OLS, PLS and PCR have α values of 0, 1/2 and 1, respectively. In other words, the main differences between OLS, PCR and PLS are what they are maximizing. OLS can be seen as maximizing correlation, while PLS and PCR work to maximize the covariance and variance, respectively. One can either choose γ arbitrarily or use CV to choose an optimal γ , as discussed in Spitzner (2004).

An alternative to CR was described by de Jong *et al.* (2001), which they refer to as continuum power regression (CPR). In CPR, they set $\mathbf{v} = \mathbf{X}^{(\gamma)}\mathbf{w}$ where $\mathbf{X}^{(\gamma)}$ is found via the spectral decomposition of \mathbf{X} . After this step, standard PLS can be applied to $\mathbf{X}^{(\gamma)}$. Just like CR, any of the special cases described above (OLS, PLS, and PCR) can be performed using this new transformed \mathbf{X} . In their paper, they discuss an algorithm for canonical PLS, which they extend to the general CPR framework. They also provide an easily implementable MATLAB code for this CPR,

and in Chapter 5, we develop a method using this CPR algorithm.

PCR-based Methods

Principal components regression (PCR) is one of the more commonly used methods in the literature that aims to reduce the number of covariates by creating linear combinations of highly correlated covariates. This can be applied to genomic data, where the highly correlated genes would be decomposed in this manner, and the first k linear combinations that account for the majority of variation in gene expression would be selected. The issue with PCR is that it does not utilize survival information, and therefore, it is not guaranteed that the components selected are related to survival. Bair *et al.* (2006) addressed this issue by creating a supervised PCR approach. Here, they first select the subset of genes that is related to survival using the univariate model fitting approach. PCR is then applied to this subset. However, they use Cox regression to select the subset of genes, which does not allow for NPH.

Li, L. and Li, H. (2004) also developed a method to handle high-dimension genomic data using PCR. Their method combines PCR with sliced inverse regression (SIR) to identify linear combinations of genes, which is an advantage because SIR takes survival information into account. Once the components are found, a Cox PH model is fitted to the components to predict survival. Their method was shown to have good predictive performance when applied to a large-B-cell lymphoma dataset, but it is based on the Cox PH model. Thus, the potential for NPH is not taken into account, so the results could be inaccurate.

PLS-based Methods

Partial least squares (PLS) is also a commonly used technique for dimension reduction. It has been shown to be a versatile tool in the analysis of high-dimensional genomic data (Boulesteix and Strimmer 2007). Unlike PCR, this takes survival information into account as it works to find components that maximize the covariance between gene expression and the survival outcome. Specific details behind PLS can be found in §5.1. Below, we discuss various PLS applications using both the Cox and AFT models.

PLS + Cox PH In the literature, most PLS application to genomic data is used in conjunction with the Cox PH model, which does not allow for NPH. Also, because PLS assumes a linear relationship between the outcome and gene expression, it cannot be directly applied to the PH model which is non-linear. To address this, Park *et al.* (2002) reformulated the likelihood for the PH model as the likelihood of a Poisson model, a generalized linear model. Nygard *et al.* (2006) then modified this approach by estimating the baseline hazard and gene effects in separate steps. Their estimation procedure is similar to Park *et al.* (2002) but computationally faster. They also extend their own approach by applying their method only to a subset of genes found to be significant using the univariate approach described earlier. This approach outperforms Park *et al.* (2002) but performs similarly as supervised PCR.

Li and Gui (2004) proposed a partial Cox PH regression method that extends the PLS idea to the PH model to create uncorrelated components from genomic data, with the ultimate goal being survival prediction. Their method works through repeated least squares fitting of residuals and Cox PH regression fitting, and once the components are found the PH model is fitted to predict survival. The

algorithm works as follows:

1. Construct the first component, T_1 .
 - a. Define $V_{1j} = X_j - \bar{x}_{.j}$, where $\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^n x_{ij}$
 - b. For each gene j , fit the Cox model $\lambda(t) = \lambda_0(t) \exp(\beta_{1j} V_{1j})$, and obtain $\hat{\beta}_{1j}$
 - c. $T_1 = \sum_{j=1}^p w_{1j} \hat{\beta}_{1j} V_{1j}$, where w_{1j} is a weight with $\sum w_{1j} = 1$
2. Construct the second component, T_2 .
 - a. Define $V_{2j} = V_{1j} - \frac{V'_{1j} T_1}{T'_1 T_1} T_1$
 - b. Fit the Cox model $\lambda(t) = \lambda_0(t) \exp(\beta_1 T_1 + \beta_{2j} V_{2j})$, and obtain $\hat{\beta}_{2j}$
 - c. $T_2 = \sum_{j=1}^p w_{2j} \hat{\beta}_{2j} V_{2j}$, where w_{2j} are the weights
3. Repeat this process to obtain T_3, \dots, T_K , where the general process is:
 - a. Define $V_{(i+1)j} = V_{ij} - \frac{V'_{ij} T_i}{T'_i T_i} T_i$
 - b. Fit the Cox model $\lambda(t) = \lambda_0(t) \exp(\beta_1 T_1 + \beta_2 T_2 + \dots + \beta_i T_i + \beta_{(i+1)j} V_{(i+1)j})$, and obtain $\hat{\beta}_{(i+1)j}$
 - c. $T_{i+1} = \sum_{j=1}^p w_{(i+1)j} \hat{\beta}_{(i+1)j} V_{(i+1)j}$, where $w_{(i+1)j}$ are the weights

After the components T_1, \dots, T_K are determined, a Cox PH model is fitted to predict survival. This method relies on the proportional hazards assumption. Also, unlike Park *et al.* (2002) and Nygard *et al.* (2006), they do not re-formulate the likelihood of the Cox model to have a generalized linear form. Nguyen and Rocke (2002, 2005) also used PLS PH regression on microarray survival data, and Bastien *et al.* (2015) developed a residuals-based sparse PLS using the Cox PH model.

PLS + AFT Devarajan *et al.* (2010) developed a method that combined the PLS method and the AFT model, a model that allows for NPH. This method has two advantages over the PLS + Cox PH method from Li and Gui (2004) described earlier. First, it adjusts the censored observations based on AFT model fitting, providing more accurate results when censoring is present. Second, it uses a model that has a linear form and allows for various types of proportional and non-proportional hazards. This PLS + AFT method is fully described and extended in Chapter 5.

2.4.3 Regularization and Parameter Shrinkage Procedures

Many researchers have tried using various parameter shrinkage techniques, specifically involving ridge regression and LASSO. These methods shrink some of the coefficients towards zero, and use a tuning parameter, say λ , that controls the amount of shrinkage. This tuning parameter is usually chosen through cross validation (CV).

Ridge Regression

First, we discuss ridge regression, which imposes a penalty on the squared values. In other words, the regression coefficients for the PH model would be estimated by maximizing

$$l(\beta) - \lambda \sum_{j=1}^p \beta_j^2, \quad (2.4.3)$$

where $l(\beta)$ is the log partial likelihood of the PH model seen in Equation 2.2.4 and $\lambda \sum_{j=1}^p \beta_j^2$ is the penalty term. Ridge regression has an added benefit of being computationally faster than LASSO, but in application, it has only been applied using the Cox PH model, which does not allow for NPH. Verweij and van Houwelingen

(1994) discuss this penalized likelihood.

LASSO

The second, and more commonly used, parameter shrinkage approach is LASSO. It is similar to ridge regression, but it penalizes based on the absolute value. This allows the method to shrink some of the coefficients to exactly 0, making it a useful gene selection technique as well. The penalized log partial likelihood can be written as

$$l(\beta) - \lambda \sum_{j=1}^p |\beta_j|, \quad (2.4.4)$$

where $l(\beta)$ is the log partial likelihood of the PH model seen in Equation 2.2.4 and $\lambda \sum_{j=1}^p |\beta_j|$ is the penalty term. One disadvantage is that the number of non-zero coefficients is at most N , the number of subjects, which means it only allows N genes in the model. In our $p \gg N$ scenario, there may be a need for more than N genes.

LASSO + Cox PH This LASSO method has been most commonly applied using the PH model, which does not allow for NPH. Tibshirani (1997) was one of the first to use LASSO based on the PH model for variable selection. Zhang and Lu (2007) also adapted Lasso to the PH model. Kaneko *et al.* (2012) then worked to improve on the LASSO Cox PH method by recognizing that the choice of the tuning parameter λ directly determines the number of genes selected, where typically a higher λ implies fewer genes. However, this set of chosen genes often includes a large number of false positives (FP) which leads to inaccurate predictions. To address this, they develop a method for estimating the false positive rate (FPR) for LASSO estimates in the high-dimensional Cox PH model. They use this method to

calculate the FPR using a mixture distribution based on the coefficients estimated by LASSO. This then helps them identify the FP genes, which they can then remove to get more accurate results.

Gui and Li (2005) created a procedure that they call LARS-Cox, which uses an L_1 penalty in the PH model to select genes with a relationship to survival. However, their approach differs from others in their use of least-angle regression (LARS) to obtain solutions for the Cox model. They use the standard LASSO penalty described in Equation 2.4.4, but the difference lies in their estimation technique. First, they reformulate the LASSO technique using an iterative procedure similar to Tibshirani (1997). Let $\eta = \beta'x$, $\mu = \partial l / \partial \eta$, $A = -\partial^2 l / \partial \eta \eta^T$, and $z = \eta + A^{-1}\mu$. With these newly defined parameters, a one-term Taylor series expansion for $l(\beta)$ has the form $(z - \eta)^T A(z - \eta)$. Then, the iterative procedure for LASSO from Tibshirani (1997) is defined as follows:

1. Fix s and initialize $\hat{\beta} = 0$
2. Compute η , μ , A and z based on the value of $\hat{\beta}$
3. Minimize $(z - \beta'x)^T A(z - \beta'x)$ subject to $\sum |\beta_j| \leq s$
4. Repeat step 2 and 3 until $\hat{\beta}$ does not change.

Tibshirani (1997) used quadratic programming to solve Step 3, but in Gui and Li (2005) they propose using a modification of the LARS algorithm, which is more computationally efficient. To do so, they first apply the Cholesky decomposition to get $T = A^{1/2}$ such that $T'T = A$. They then rewrite Step 3 as

$$\text{Step 3: Minimize } (y - \beta'\hat{x})^T (y - \beta'\hat{x}) \text{ subject to } \sum |\beta_j| \leq s$$

where $y = Tz$ and $\hat{x} = Tx$. This is then the LASSO of y on \hat{x} and can be solved by the LARS algorithm given s . Zou (2008) also developed a shrinkage method for variable selection in the PH model using LAR, but their method was not applied in the genomic setting.

LASSO + AFT Huang *et al.* (2005) also propose a method for variable selection involving LASSO, but unlike most methods theirs is not based on the Cox PH model. Instead, they base their method on the AFT model, which allows for NPH, and use Stute's weighted least squares method (Stute 1993). Stute's estimator for AFT is computationally better for more covariates, which is a benefit in our high-dimensional genomic setting. Recall from Equation 2.2.14 that the AFT model has a linear form. Here, let $T_i = \beta_0 + X_i'\beta + \epsilon$. In general, the Stute estimator uses KM weights to account for censoring in the least squares criterion. To display Stute's estimator, first let \hat{F}_n be the KM estimator of the distribution function $F(T)$, and write $\hat{F}_n(y) = \sum_{i=1}^n w_{ni} I\{Y_{(i)} \leq y\}$, where w_{ni} are the jumps in the KM estimator expressed as

$$w_{n1} = \frac{\delta_{(1)}}{n}, \quad w_{ni} = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, \quad (2.4.5)$$

where $\delta_{(1)}, \dots, \delta_{(n)}$ are the censoring indicators corresponding to the ordered Y_i s. Then, Stute's weighted least squares (LS) estimator $\hat{\theta} \equiv (\hat{\beta}_0, \hat{\beta})$ are the estimators that minimize

$$L(\theta) = \frac{1}{2} \sum_{i=1}^n w_{ni} (Y_{(i)} - \beta_0 - X'_{(i)}\beta)^2, \quad (2.4.6)$$

and thus, the LASSO estimator for $\hat{\theta}$ is described as minimizing

$$L_{\lambda}(\theta) = \frac{1}{2} \sum_{i=1}^n w_{ni} (Y_{(i)} - \beta_0 - X'_{(i)}\beta)^2 + \frac{\lambda}{n} \sum_{j=1}^d |\beta_j|. \quad (2.4.7)$$

One disadvantage to this method is that its threshold-gradient-directed regularization may overestimate the number of non-zero coefficients. Wang *et al.* (2008) uses a doubly penalized Buckley-James (BJ) method for survival data with high-dimensional covariates, and Wang and Wang (2010) also developed a BJ boosting procedure using the AFT model in high-dimensional biomarker data. Johnson (2008, 2009) also studied the use of penalized estimation functions for variable selection in the AFT model.

Lassoed Principal Components (LPC) We have discussed both LASSO and principal component (PC) methods, and now we will discuss a method that combines both ideas. Lassoed principal components (LPC) was proposed by Witten and Tibshirani (2008) and strives to use all features of a genomic data set. The motivation for this method comes from the biology behind genes, where we expect that genes work together, rather than individually, to affect a particular phenotype. LPC addresses this idea by favoring groups of correlated genes having an effect on survival over those with an individual effect and no correlation with other genes. The method itself has a general two-step process. First, the Cox scores are computed for each gene, call these T . Then, these scores are regressed onto the eigenvectors of the gene expression data matrix, X , subject to the L_1 penalty seen in LASSO. The LPC scores are the resulting fitted values. In other words, if we let $\mathbf{v}_1, \dots, \mathbf{v}_m \in \mathbb{R}^p$

represent the right singular vectors of X , then the LPC scores \hat{T} are given by

$$\hat{T} = \hat{\beta}_0 + \sum_{i=1}^n \mathbf{v}_i \hat{\beta}_i, \quad (2.4.8)$$

where $\hat{\beta} = \operatorname{argmin}_{\beta} \{ \|T - \beta_0 - \sum_{i=1}^n \mathbf{v}_i \beta_i\|^2 + \lambda \sum_{i=1}^n |\beta_i| \}$ and λ is chosen using CV. They demonstrate in their paper that this method leads to more accurate gene rankings compared to other methods. The obvious disadvantage of this method is that it uses Cox scores, rather than a model that allows for NPH.

2.4.4 Machine Learning Methods

In this section, we discuss methods with foundations in machine learning techniques, such as boosting, random forests, and networks.

Zhang *et al.* (2013) developed a method using networks but based on Cox regression, called Net-Cox. It works to integrate network information from gene co-expression into the Cox PH model. They first construct \mathbf{L} , a positive semidefinite matrix derived from gene expression network information, and create the following network constraint to the Cox model

$$l(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2} \lambda \boldsymbol{\beta}' [(1 - \alpha) \mathbf{L} + \alpha \mathbf{I}] \boldsymbol{\beta}, \quad (2.4.9)$$

where $l(\boldsymbol{\beta})$ is the Cox PH log partial likelihood described in Equation 2.2.4, \mathbf{I} is the identity matrix, λ is the parameter controlling the amount of shrinkage, and α is the parameter weighting \mathbf{L} and \mathbf{I} in the network constraint. They found that this network approach performed better than the more standard LASSO Cox methods. However, their method does not allow for NPH since it is based on the Cox PH model.

Li and Luan (2005) discuss a method for inference in the Cox PH model with a boosting procedure using smoothing splines. The procedure is related to the previously discussed method of Gui and Li (2005), without the linear assumption for the predictive variables, the genes. In the Gui and Li (2005) paper, they assume a simple linear risk function, $f(x) = \beta'x$, where the hazard function of the PH model can be written as $\lambda(t|X) = \lambda_0(t) \exp\{f(x)\}$. Li and Luan (2005) do not assume this form and instead estimate this risk function $f(x)$ non-parametrically. Because of this, this method can be used to identify non-linear effects of genes with a risk of developing an event. Leng and Ma (2007) also used the idea of smoothing but under the AFT model, which allows for NPH, and estimates were found using Stute's estimator which is computationally efficient. However, their method has only been applied to small data sets and not the large, genomic data we are interested in. Lu and Li (2008) also developed a boosting procedure, but their method is based on a nonlinear transformation model. A nonparametric pathway-based regression model was developed by Wei, Z. and Li, H. (2007) and uses a boosting procedure with an application in genomic data.

Geng *et al.* (2012) developed a more recent model-free method for both risk classification and survival prediction. The benefit to this method is that it requires no model assumption for the data. The main idea is to create a non-parametric estimator for t-year survival probability using a series of weighted Support Vector Machine (SVM) decision rules. These are used for both classification - high vs. low-risk - and survival prediction, but they do not address variable selection. Belle *et al.* (2010) also discuss a method that uses SVM within the survival framework and compares it to other methods involving the standard Cox model and neural networks. They find that the SVM method outperforms the others, but its variable selection performance was not evaluated.

Pang *et al.* (2012) discuss the use of random forests to select a group of prognostic genes. Essentially, a binary survival tree is grown by splitting the continuous values of the genes using a deterministic algorithm and bootstrap sampling. They offer various suggestions for splitting criteria and a method for aggregating to get the cumulative hazard estimates. They also show that their method performs well in the high-dimensional setting, as they incorporate multivariate correlations in genomic data. However, this method is computationally intensive and would require a parallel version for it to be useful in the genomic setting. Hastie *et al.* (2001) also developed an approach called ‘tree harvesting’ that starts with a hierarchical clustering of genes and then models the outcome variable as a sum of the average expression profiles of chosen clusters.

2.4.5 Other Methods

While the methods described above are the most common in the literature, there have been other types of approaches that address the high-dimensional issue. For example, Fan *et al.* (2010) proposed a sure screening procedure based on the Cox PH model. They discuss the issues with penalty based methods, citing their inability to perform well when the number of covariates is ultra-high, $p \gg n$, a common trend in genomic data. The sure screening procedure performs better in this case. In general, a method has the sure screening property if all the important variables survive after applying a variable screening procedure. Thus, a dimension reduction method is desirable if it has the sure screening property. The original sure independence screening was based on the basic idea of marginal correlation ranking. They extend this work to the PH model and iteratively apply large-scale screening that eliminates unimportant variables using conditional marginal utility. They show that their

method is computationally efficient compared to LASSO and more conservative in selecting variables. However, their approach is based on the PH model, which does not allow for NPH.

Datta *et al.* (2007) developed three approaches based on the AFT model. They also directly address the issue of censoring by proposing three methods for handling the censored observations. They found mean imputation to be the best approach. Their paper did not develop any new methodology in dimension reduction. Instead, their method focused mostly on the imputation procedure and then applied standard PLS and LASSO.

Engler *et al.* (2009) present a method using an elastic net approach under both the Cox PH and AFT model. This is one of few methods that describes a procedure based on a model other than the Cox PH model. The AFT model is beneficial in general because it allows for various forms of NPH. Unlike LASSO, this method allows one to select a number of variables greater than the number of subjects. Their method outperformed other Cox-based and AFT-based approaches in terms of computational time, and they also either matched or exceeded other approaches in terms of predictive performance. Also, the AFT-based method outperformed the Cox-based method in gene selection.

Laimighofer *et al.* (2016) published one of the most recent articles on the subject, attempting to address both issues of over-fitting a model and detecting genes with high predictive power. They use a nested CV approach and compare their results to the standard LASSO Cox, similar to what was described earlier. Their method contains both an inner and outer CV, where the inner CV is used to obtain the optimal number of parameters and the outer CV estimates unbiased prediction accuracy. The outer CV utilizes the C index (Harrell *et al.* 1996) from the Cox model. In general, within this nested CV framework they fit a survival model,

evaluate it in an unbiased fashion by splitting the data into separate training and test sets, and then select genes with the highest predictive power using results across different CV folds. To select these genes, they aggregate the CV information and weight the features based on their performance across all CV runs. This is the fundamental difference between their method and LASSO Cox. LASSO Cox applies CV to the whole data set, but this method aggregates the results of different CV runs and creates a weighting scheme to ensure the selection of highly predictive genes. This method has two disadvantages. First, it is based on the Cox PH model, and second, the nested CV approach makes it computationally slow. Li and Luan (2003) developed a kernel Cox PH regression model for linking gene expression profiles to censored survival data and applied their approach to genomic data from patients with diffuse large B-cell lymphoma.

Eng and Hanlon (2012) developed an approach called Cox Assisted Clustering. Here, they define a discrete mixture model that can identify and model genetic subgroups for time-to-event data. To handle heterogeneity, they define K classes, where π_k is the probability that a patient comes from class k , $k = 1, \dots, K$. They assume that each class k follows a Cox PH model with distribution function $f_k(y, \delta|x)$. The mixture density is then written as

$$f(Y, \Delta|x) = \prod_{i=1}^n \sum_{k=1}^K \pi_k f_k(Y_i, \delta_i|x_i). \quad (2.4.10)$$

Regression coefficients for this model are estimated using the expectation-maximization (EM) procedure (Dempster *et al.* 1977). One issue with this approach is that the covariates may have different distributions across groups, and the baseline hazards may also be different.

Wang *et al.* (2006) create what they call a double penalized Buckley-James

method for the semiparametric AFT model to correlate genomic data to censored survival outcomes. Their method uses a mixture of L_1 (absolute value) and L_2 (quadratic)-norm penalties and performs automatic gene selection and parameter estimation where highly correlated genes are either selected or removed together. While their method is based on a model that allows for NPH, it still has some restrictions. First, similar to LASSO, it does not allow for the number of genes selected to be greater than N , and second, it is computationally slow.

Other Applications Here we discuss papers that are application based. First, Broet *et al.* (2009) sought to identify subgroups with the greatest risk of relapse in large cell lung carcinoma. They used a Bayesian approach to restrict the model to genes exhibiting only copy number-driven expression. Then, they performed their gene selection strategy, which had two parts. First, the feature selection component was performed, where genes were individually ranked based on their genomic survival posterior probabilities. Then, linear combinations of genes weighted by their respective Cox PH regression coefficients were created and used to compute a gene expression signature that was then used to classify genes into one of two groups, high and low-risk.

Peri *et al.* (2013) studies genes in clear cell renal cell carcinoma (ccRCC) and their association with survival outcomes. They applied a non-parametric approach using a function called RankProd (Hong *et al.* 2006), a method that identifies genes under one condition compared to another. In their paper, they used ccRCC gene expression data sets compared to normal, and genes with false discovery rate (FDR) $< .05$ based on 10,000 permutations were considered significant. They then identified the signatures of interest and use the relative risk from the Cox PH model and the coefficients from the AFT model, as well as the model p-values, to deter-

mine the magnitude of association between gene expression and survival. Using this approach, they identified a subset of potentially-druggable targets whose elevated expression correlates with poor prognosis and survival. While this method has the advantage of using the semi-parametric AFT model, which allows for NPH, it is univariate in nature and does not account for the correlation between genes.

Huang and Harrington (2004) examine methods for dimension reduction where the goal is to find predictors of potentially censored outcomes. In their study, they are specifically interested in predicting change in HIV viral load, and the predictors are both prognostic measures and gene expressions. More specifically, they use the data to cluster subjects into groups with either a “good” or “poor” response. To accomplish this, they develop a method for ranking subjects according to predicted outcome using stepwise regression, PCR and PLS, all adapted to the Buckley-James algorithm, an iterative procedure for determining parameter estimates in the AFT model. They find that the Buckley-James PLS and PCR approaches perform similarly in terms of predictive power, but the PLS approach has several advantages in terms of interpretation and computation. One downfall to their approach is that it sometimes has convergence issues.

2.4.6 Comparing Methods

The methods described in this section all work to address the high-dimensional structure of genomic data where $p \gg n$. We have seen classical approaches based on the Cox model, dimension reduction techniques, and parameter shrinkage estimation. Bovelstad *et al.* (2007) compared the performance of seven of the methods described above: univariate selection, forward stepwise selection, PCR, supervised PCR, PLS regression, and LASSO based on the Cox PH model. They found that

methods involving either a shrinkage estimator or linear combination of the covariates performed better than other methods. Witten and Tibshirani (2010) also compared some of the methods described above, including their own LPC method, and found that their LPC method provided a slight improvement over the other methods. Wieringen *et al.* (2008) also compared the various methods discussed in this section, such as univariate selection, supervised PC, partial Cox PH regression, LASSO Cox and tree based methods. While the results varied based on each data set, they found similar results as Bovelstad *et al.* (2007), where penalized Cox PH regression appeared to perform well in relation to other methods.

The problem with LASSO, however, is that it does not allow for the selection of more than N genes, which may be necessary in the genomic setting. Thus, in general, PLS based methods have a benefit. First, it has an advantage over PCR because it takes survival information into account, maximizing the covariance between gene expression and the survival outcome. Second, PLS can be applied in combination with AFT, a versatile linear model with many special cases that allows for NPH. Furthermore, the development of a method based on CPR would be beneficial, as it encompasses the same advantages of PLS with the added benefit of the choice of CPR parameter.

In general, while all of the methods in this section differed in their approach, there was one common thread: most methods were based on or used in conjunction with the Cox PH model. Thus, they do not address or recognize the potential for gene expressions exhibiting some form of NPH. Also, none of the methods, except for Devarajan *et al.* (2010) and Datta *et al.* (2007), address the issue of censoring. This demonstrates that there is still a need for dimension reduction techniques that address both the censoring and NPH issues. We will propose methods to handle these challenges in Chapter 5.

CHAPTER 3

EXAMPLES: MODEL FITTING AND VARIABLE SELECTION

In this chapter, we explore large-scale genomic datasets with censored survival outcomes in oral and ovarian cancer. We fit various models described in Chapter 2 at the univariate level to test for model significance and goodness-of-fit (GOF). The goal is to identify genes that exhibit some form of NPH, thus demonstrating the need for alternatives to the PH model. First, we will give a description of each data set, emphasizing the large number of covariates and high censoring proportions. Then, we will clearly demonstrate evidence of NPH and reinforce the need for alternative models and methods. The two datasets of interest are described below.

- ORAL: Saintigny *et al.* (2011) studied 86 subjects enrolled in a clinical chemoprevention trial that used the development of oral cancer as the endpoint. This

data had been pre-filtered and contained 12,776 genes.

- OVARIAN: Tothill *et al.* (2008) studied the relationship between survival and gene expression data for subjects with ovarian cancer. Only genes with non-missing values were used in the analysis. After filtering the data with a CV threshold of 5%, there were 276 subjects with 32,575 genes to analyze.
- ORAL (RNA-Seq): This data was generated by the TCGA Research Network: <http://cancergenome.nih.gov/>. It looks at the survival and gene expression information for 221 subjects with head and neck squamous cell carcinoma, with a focus on the sub-type oral cavity. Gene expression was measured using RNA-seq technology on 19,341 genes.

3.1 Model Fitting

In the oral and ovarian set, gene expression is measured on the \log_2 scale. The initial step was to employ univariate Cox regression and determine the goodness of fit (Grambsch and Therneau 1994) of the model. In other words, genes that did not follow the PH assumption were identified. The proposed methods in Chapters 4 and 5 rely heavily on the PO, semiparametric AFT and YP models, so we also perform a goodness of fit test for those (Yang and Prentice 2005; Martinussen and Scheike 2006; Novak 2010) to get a clearer picture of the proportion of genes that typically fit these models and determine whether they pose a good alternative to the PH model. For both datasets, clinical covariates such as age at diagnosis and stage of cancer were available. Therefore, the PH, PO and semiparametric AFT models were fit adjusting for these clinical variables. Currently, the YP model implementation is only capable of handling a single covariate (Yang and Prentice 2005), and, hence,

we did not adjust for age and stage for analyses involving this model. We also note that the PH, PO and semiparametric AFT models are fit to the continuous genomic data, but there is no estimation in the YP model for continuous covariates. Thus, for this application, the YP model is fit to the dichotomized case, where gene expression is dichotomized based on the median as either high (1) or low (0).

Table 3.1: Univariate Analysis Results

Data Set:	Oral	Ovarian
Total Genes:	12,776	32,575
% Censored	59%	59%
CoxNF	1,810 (14%)	4,617 (14%)
POF	11,544 (90%)	29,571 (91%)
CoxNF \cap POF	947	2,725
CoxNF \cap PONF	863	1,892
SAF	9,753 (76%)	29,902 (92%)
CoxNF \cap SAF	1,379	4,224
CoxNF \cap PONF \cap SAF	694	1,685
YPF	11,609 (91%)	29,960 (92%)
CoxNF \cap YPF	1,544	4,070
CoxNF \cap PONF \cap YPF	717	1,645

Table 3.1 shows the size of the datasets, the proportion censored, and the results of GOF tests at the 5% significance level. In this analysis, Storey's q -value multiple testing procedure (Storey 2002) was employed to account for multiple hypotheses. For the oral data, the q thresholds corresponding to $p < .05$ were .24, .12, .10 and .55 for the PH, PO, semiparametric AFT and YP models, respectively. The p -value histogram for the YP model shows a U-type shape, which may be causing

the large q value. This may be due to the dichotomization of the covariate. For the ovarian data, the q thresholds corresponding to $p < .05$ were .25, .25, .53 and .33 for the PH, PO, semiparametric AFT and YP models, respectively. The p-value histogram for the AFT model did show a sharp peak towards zero but an insignificant dropped plateau for larger p-values, which may be causing the large q value.

In Table 3.1, we employ the following abbreviations: POF, SAF and YPF represent sets of genes for which the PO, semi-parametric AFT and YP models fit, respectively; and CoxNF and PONF refers to sets of genes for which the PH and PO models do not fit, respectively. It is clear that each set has a significant number of genes (14%) for which the PH model does not fit. In each case, there are also a large number of genes for which the PO model fits. However, we are most interested in the subsets created from these sets as outlined below. First we examine the genes for which PH does not fit but PO fits. This subset has a large number of genes, 947 and 2,725 for the oral and ovarian datasets, respectively. From this, we observe that in each case, the PO model fits a large fraction (over 52%) of genes for which PH does not fit. Thus, it would be beneficial to develop methods based on the PO model because it has been shown to handle NPH (Martinussen and Scheike 2006). However, we also note that the PO model does not fit 41-48% of genes for which the PH model does not provide a good fit in both datasets. Thus, it would also be useful to develop methods based on an alternative model that can handle NPH, such as AFT or YP. Looking at the semiparametric AFT GOF, we observe that this model fits 76% and 91% of the 1,810 and 4,617 genes for which the PH model does not fit, respectively, for the oral and ovarian datasets. In fact, the semiparametric AFT model fits a large fraction (80% and 89%, respectively) of the genes that do not fit both the PH and PO models, and thus, it appears to offer a more flexible alternative to the other models. The YP model shows a similar advantage. This

model fits 85% and 88% of genes for which the PH model does not fit, for the oral and ovarian datasets, respectively, and it also fits a large fraction (83% and 87%, respectively) of genes that do not fit both the PH and PO models. Thus, similar to the semiparametric AFT model, the YP model shows versatility and the ability to fit a large number of genes when the PH and PO models do not. In summary, these results serve as motivation for developing methods based on the PO, semiparametric AFT, and YP models because of their ability to fit not only a large number of genes in general, but specifically genes for which the PH assumption is violated.

Next, we look at two subsets. The first subset, Subset A, is described below and represents genes discussed in Table 3.1 for which the PH model does not fit but alternative models (PO, semiparametric AFT, and YP) do fit. The oral dataset has over 1,800 genes in Subset A and ovarian has over 4,600.

Subset A: PH does not fit

Subset A-PO: PH does not fit \cap PO fits

Subset A-SA: PH does not fit \cap Semipar AFT fits

Subset A-YP: PH does not fit \cap YP fits

Next, we applied the PH, semi-parametric AFT, PO, and YP models to identify genes that have a significant relationship with survival time and looked at overlaps between the significant genes found in each model and those for which that particular model fits. This is referred to as Subset B. Then, for the semi-parametric AFT, PO, and YP models, we looked at the overlap between Subset B and the genes for which the PH model did not fit. This subset is referred to as Subset B-CNF. In general, the goal was to identify various subsets of genes for which the PH

assumption fails to hold but another model fit and showed significance and provided a good fit.

Subset B: Model Fits \cap Significant

Subset B-CNF: Model Fits \cap Sig \cap PH does not fit

The venn diagrams in Figure 3.1 show the number of genes in Subset B and B-CNF. These subsets were chosen because they specifically narrow down genes that exhibit some form of NPH but fit and are significant for another model. We observe in Figure 3.1 that Subset B has a large number of genes in both data sets for the PO, AFT and YP models. Furthermore, the oral data has 263, 301, and 437 genes in Subset B-CNF for PO, AFT, and YP, respectively. The ovarian data also has a large number of genes in Subset B-CNF, with 523, 635, and 816 for PO, AFT, and YP, respectively.

3.2 Subset Analysis Using a Weighted Average of Gene Expressions

Once the subsets in §3.1 were established, an analysis was run on a weighted average of the gene expressions. Let m be the number of genes in the subset of interest. If $\beta = \{\beta_1, \dots, \beta_m\}'$ are the regression coefficients for the respective model and \mathbf{Z} is the $n \times m$ gene expression matrix including only those genes in the subset of interest, then the weighted average, η , can be calculated as

$$\eta = \mathbf{Z} \times \beta \tag{3.2.1}$$

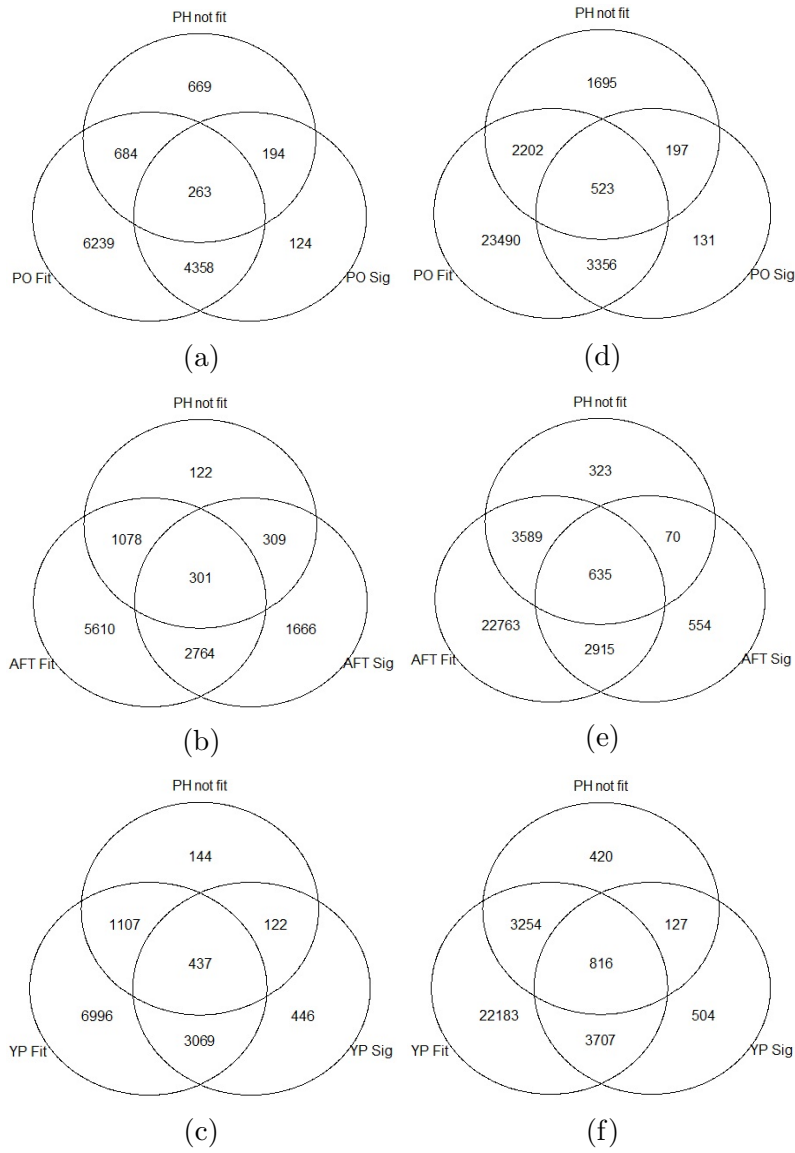


Figure 3.1: Venn Diagrams for Subset B and B-CNF.
(a)-(c) Oral, (d)-(e) Ovarian

This results in an n -dimensional vector, where each subject has a single value representing their weighted average across all genes, and can be interpreted as a linear predictor. Also, we note that for the YP model with two parameters, the weighted average is calculated as $\boldsymbol{\eta} = \mathbf{Z} \times (\boldsymbol{\beta} + \boldsymbol{\gamma})$. Exploratory analysis was performed on

these weighted averages for each subset. The objective was to determine if and how average gene expression was related to patient survival and specifically whether NPH were present.

We first look at the results for Subset A. Recall, Subset A contains only genes for which the PH does not fit, and its subcategories refer to the overlaps between these genes and genes that fit the PO, semiparametric AFT, and YP models. For Subsets A-PO, A-SA, and A-YP, we used the regression coefficients from their respective models, but because Subset A does not have a specific model, we used the coefficient from CON, the concordance regression described in §2.5.3. A GOF test was run for the PH, PO and semiparametric AFT models on each of the continuous weighted averages, and the YP GOF test was run on the dichotomized weighted average. The resulting p -values are shown in Table 3.2, where a large p -value implies the model fits.

Based on the p -values, we observe some clear evidence of NPH. In each case, the PH GOF p -value is very low, in fact in most cases it is 0, verifying that the PH assumption is violated. We also observe low GOF p -values for the PO model, except for the oral A-PO and A-SA cases, indicating that while the PO model does show some versatility, it does not fit the majority of the subsets. This suggests time-varying gene effects and the need for alternative methods to handle such cases. In each case, we observe large p -values for both the semiparametric AFT and YP GOF, indicating that these models fit well. This further emphasizes their versatility and ability to handle NPH.

Next, we perform a similar analysis for Subset B. Recall, this subset examines the genes that are both significant and fit for a particular model. For each of these subsets, we use the regression coefficients associated with the specific model (PH, PO, Semiparametric AFT, YP) to calculate the weighted gene expression, and

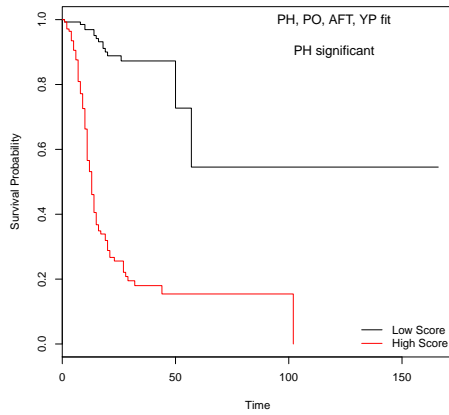
Table 3.2: GOF p -values for Weighted Averages, Subsets A

Subset	Oral				Ovarian			
	A	A-PO	A-SA	A-YP	A	A-PO	A-SA	A-YP
Cox PH	.03	.04	.03	0	0	0	0	0
PO	.04	.20	.57	0	0	0	0	0
Semipar AFT	.87	.06	.28	.85	.14	.74	.08	.07
YP	.98	.28	.22	.26	.05	.27	.39	.29

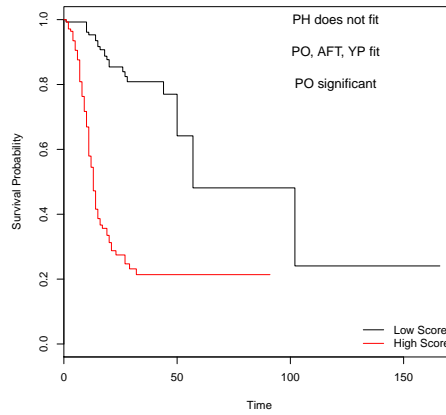
then GOF tests are applied. Also, to get a visual representation, η was dichotomized by the median into high (H) and low (L) values and Kaplan-Meier (KM) survival curves were plotted comparing the two groups. The survival curves and GOF results for the ovarian data can be seen in Figure 3.2.

As expected, we observe physical evidence of PH for the Cox subset, and the GOF reflect this as well. The Cox GOF p -value is 0.87, indicating that the PH model does fit, but we also note that each of the other models fit as well. This is because the PO, SA, and YP models also have the ability to handle PH and our results indicate that they do provide a good fit. For the PO subset, we see physical evidence of converging hazards and the Cox GOF p -value drops to 0.04, but the other three models still fit (GOF $p > .05$). For the SA subset, the PO, SA, and YP models fit, but the Cox model does not. Last, for the YP subset, we observe clear evidence of crossing hazards and the YP model is the only model that fits (GOF $p > .05$), while the Cox, PO and SA models have GOF p -values close to zero. Thus, we observe that the PO, SA, and YP models are flexible enough that they are able to fit well in both PH and NPH scenarios. On top of GOF, model significance was also checked, and in each case, the model of interest is significant ($p < .05$) for its respective subset.

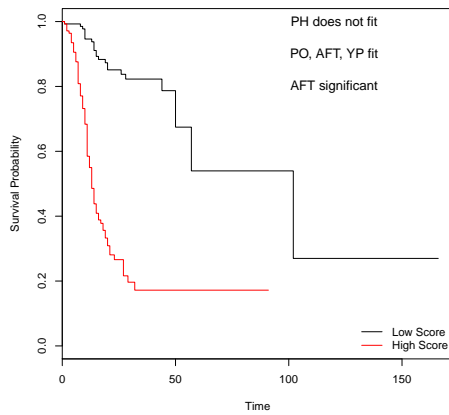
Table 3.3 shows the resulting p -values for the ovarian Subset B-CNF. Recall, this subset eliminates genes for which the PH model fits from the PO, SA, and YP



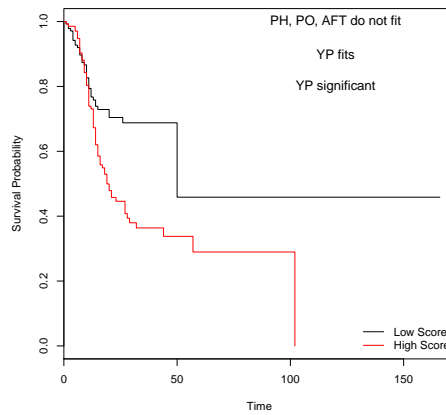
(a) Cox Fit and Sig



(b) PO Fit and Sig



(c) SA Fit and Sig



(d) YP Fit and Sig

Figure 3.2: Survival Curves for Subset B, Ovarian

cases of Subset B. This will produce a clean subset of genes that have no connection to the PH model. Looking at the GOF p -values, we see that the results mimic what we found in Subset B. Once the genes for which the PH model does not fit are removed, the PH model does not fit any weighted average. For the PO GOF p -values, we observe varying results, but we take note that the PO model fits in both PO subset cases. For cases where neither PH nor PO fit, it would be valuable

to have an alternative approach, such as a method based on the SA or YP model. In most of those cases, the semiparametric AFT and YP GOF p -values are large, indicating that both models fit well.

Table 3.3: Ovarian GOF p -values for Weighted Averages
Subset B-CNF

Subset	GOF	PO	Semipar AFT	YP
B-CNF	Cox PH	0	0	0
	PO	.20	.03	0
	Semipar AFT	.84	.86	.04
	YP	.11	.05	.70

We now examine the results for Subset B and B-CNF in the oral data. Table 3.4 shows the resulting GOF p -values. In most cases, particularly for B-CNF, the results parallel what was seen in the ovarian data. However, for Subset B, we observe that all of the resulting p -values are high, indicating that all four models fit in each case. Conversely, as the genes for which the PH model fits are eliminated, we see that the Cox GOF p -values drop in most cases. Thus, the PH model shows less versatility than the other models. Alternatively, the PO model fits in many cases, including cases where the PH model does not. Most interestingly, the SA and YP models fit in every case, and thus, show the most versatility. Thus, while the PH model does appear to fit in some cases, the PO, SA and YP models demonstrate an advantage, as they are able to fit in cases where the PH model fits and does not fit. Therefore, they are able to handle both PH and NPH scenarios.

Table 3.4: Oral GOF p -values for Weighted Averages
Subset B and B-CNF

Subset	GOF	PH	PO	Semipar AFT	YP
B	Cox PH	.28	.45	.73	.22
	PO	.17	.26	.39	.04
	Semipar AFT	.76	.87	.85	.70
	YP	.12	.11	.10	.55
B-CNF	Cox PH	-	0	.04	.08
	PO	-	.04	.08	.06
	Semipar AFT	-	.94	.49	.84
	YP	-	.42	.82	.35

3.3 Graphical Illustration of Time-Varying Effects in Individual Gene Expression

We selected genes in both the oral and ovarian datasets that exhibit some form of NPH and graphically displayed their relationship to survival. This differs from the analysis done previously where we looked at specific subsets of genes. Here, we used those subsets to narrow down the gene pool, and then select one gene at a time to examine at random. This is a solely univariate process, where we dichotomize each gene expression by the median, and plot KM survival curves based on high and low gene expression dichotomized by the median. The point of this exercise is to provide further evidence of the presence of NPH, and thus, the need for models and methods that can handle these hazard forms.

Oral Data Figure 3.3 shows the survival curves of four genes from the oral dataset found after examining only those genes for which the PH model did not fit. This is not an exhaustive list, but rather a few examples showing that NPH is present at the univariate level. As seen below, all four genes display some form of NPH. The

first three show crossing hazards, while the third shows diverging hazards. In each case, the Cox model does not fit ($GOF\ p < .05$), indicating the PH assumption is violated. On the other hand, the PO model fits two genes while, the semiparametric AFT and YP models fit all four genes.

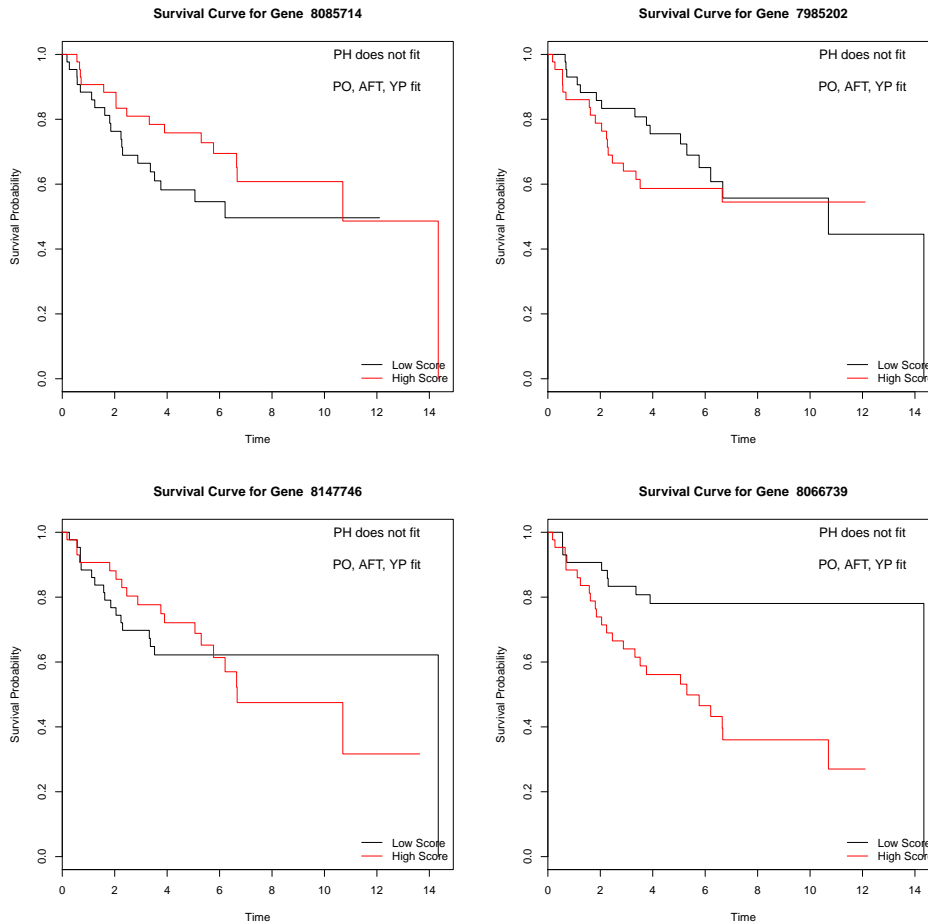


Figure 3.3: Survival Curves for Individual Genes: Oral

Ovarian Data Similarly, we display specific examples of genes exhibiting NPH in Figure 3.4. Here, all four examples show signs of crossing, and the last gene also shows diverging survival curves. Based on the GOF results, we observe that while

the PH model does not fit in each case, the PO, YP and AFT models do. Again, this is not an exhaustive list, but merely a few examples of NPH found in the ovarian dataset. The semiparametric AFT model fits three out of the four genes, and the YP model fits in each case.

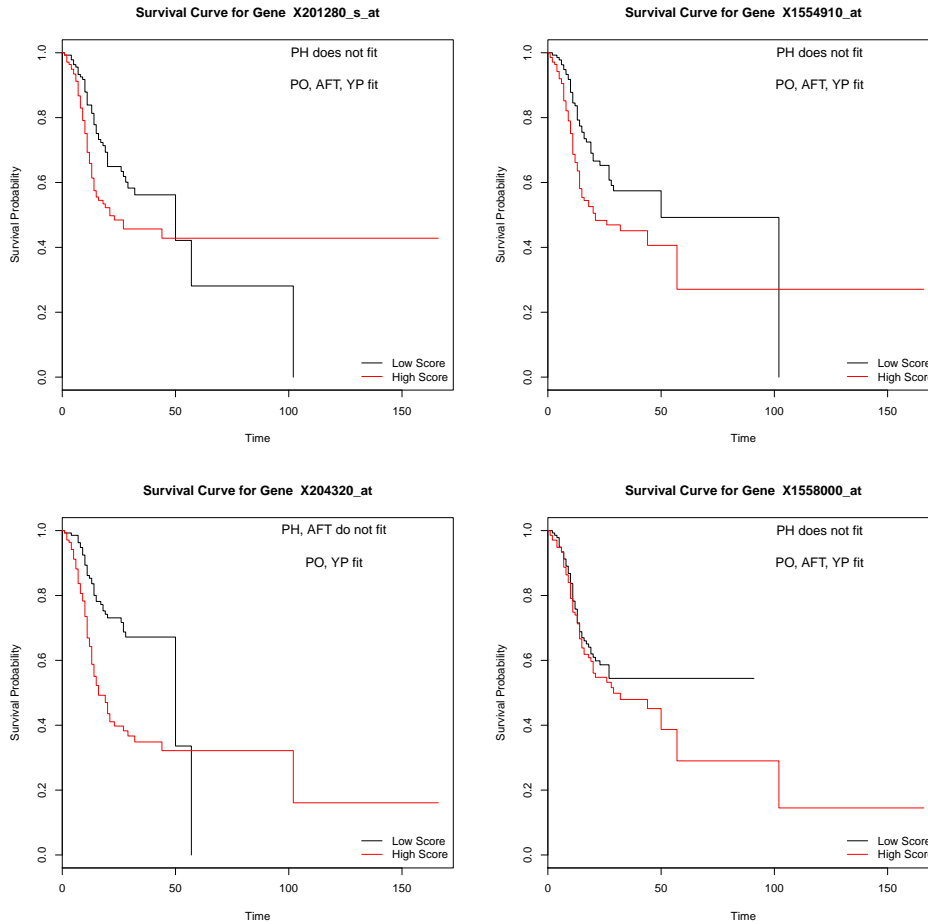


Figure 3.4: Survival Curves for Individual Genes: Ovarian

Gene Interactions Next, we look at interactions between the few genes selected above to check if some type of modulating effect is present. This idea is consistent with the known biological activity of some genes that tend to be linked, interact

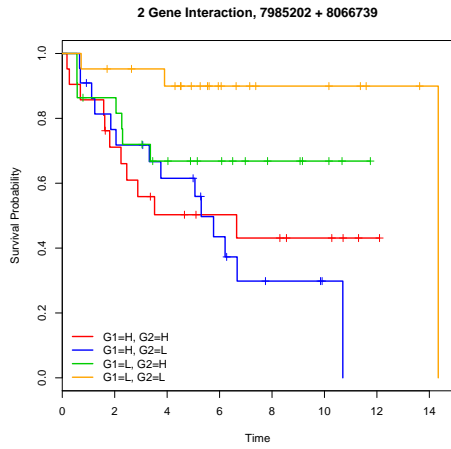
with and affect each other. We look at the dichotomized values for two genes and split the subjects into four sub-groups: Genes 1 & 2, High; Gene 1, High & Gene 2, low; Gene 1, Low & Gene 2, High; and Genes 1 & 2, Low. The KM survival curves for the four sub-groups are plotted on one axes to compare in Figure 3.5.

For the oral data, we observe varying results depending on the genes involved. In Figure 3.5a, it is clear that patients with both lowly expressed genes have the highest survival probability, while in Figure 3.5b, the lowly expressed genes have the worst prognosis. We also observe crossing in both plots. For example, in the first plot, the subjects with both highly expressed genes have the lowest survival until the 5 year mark, where the subjects with a low expression for gene 2 then have the worst prognosis. In the second plot, subjects with a lowly expressed gene 1 and highly expressed gene 2 cross the subjects with two highly expressed genes at the 5 year mark and then have the highest survival. Based on these results, we could hypothesize that the crossing effects seen in the individual cases in Figure 3.3 could be related to some modulating effect between the genes.

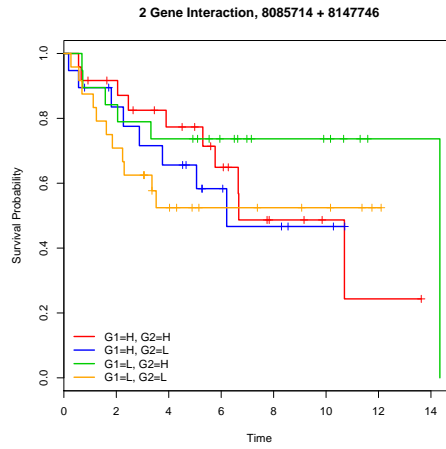
For the ovarian data, we discover similar results for both interactions. In both Figures 3.5c and 3.5d, survival is improved for subjects with a low expression for both genes, while subjects with a high expression for both genes have the lowest prognosis. In the first plot, subjects with a lowly expressed gene 1 and highly expressed gene 2 do not show much improvement compared to the group for which both are highly expressed. We also observe evidence of crossing, specifically for the lowly expressed gene 1 and highly expressed gene 2 group, which has a significant drop in survival at 100 months. Again, we could hypothesize that the crossing effects seen in the individual cases in Figure 3.4 could be related to some modulating effect between the genes.

In this section, we were able to demonstrate the challenges commonly en-

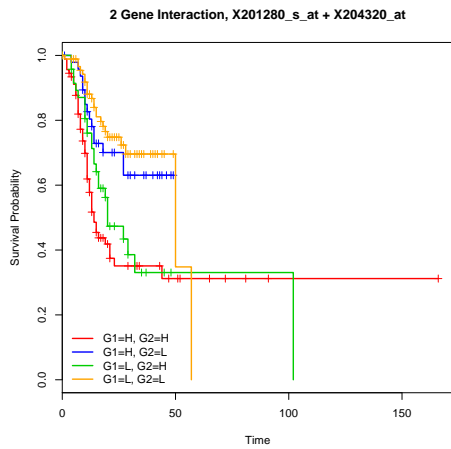
countered in the analyses of large-scale genomic data with censored survival outcomes. In addition to high-dimensional and low sample size issues, the observed censoring proportions are often high (Beer *et al.* 2002; Tothill *et al.* 2008; Van der Net *et al.* 2008; Saintigny *et al.* 2011), and as seen in the tables and figures throughout this chapter, there is often evidence of NPH. In the following chapters, we will propose methods to handle these challenges.



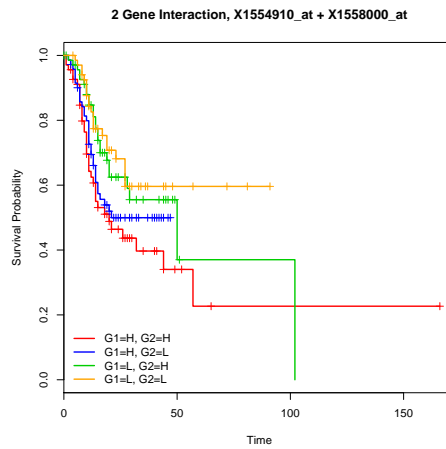
(a) Oral



(b) Oral



(c) Ovarian



(d) Ovarian

Figure 3.5: Survival Curves for 2 Gene Interactions

CHAPTER 4

PROPOSED METHODS FOR VARIABLE SELECTION AND RANKING

It has been shown in Chapter 3 that there is a need for alternative methods to the PH model that can handle various types of NPH. It was also demonstrated that when various goodness-of-fit tests are applied to real-life high-dimensional genomic data sets, there are many instances where the PH model does not fit but the PO or YP model does. Thus, it would be useful to develop other methods that deal specifically with these models.

In this section, we construct several marginal screening approaches based on the Proportional Odds (PO), Proportional Log Odds (PLO), and the Yang-Prentice (YP) models. In §4.1, we generalize the information-theoretic approach of Devarajan and Ebrahimi (2009) and develop a test for gene effect using the YP model, which includes tests for the PH and PO models as special cases. Then, we propose R^2

measures for the PH and PO models using these tests for gene effects. In §4.2, we propose a variety of R^2 measures covering a wide range of survival models. First, we develop a unified R^2 measure by generalizing the work of Rouam et al. (2010, 2011) to include the PO, PLO, CH and PH models. We then propose R^2 measures for the PO model based on the likelihood ratio test. These new variable selection methods are applied to both simulated and real-life genomic data and compared to previous methods to demonstrate their usefulness and advantages.

4.1 Test for Gene Effect

As outlined in §2.3.3, Devarajan and Ebrahimi (2009) developed a test for covariate effect in the PH model based on Kullback-Leibler (KL) divergence. Their methods were easy to implement and performed better than other known tests in terms of power, specifically when handling small to medium sample sizes. While their approach had certain benefits in terms of power and computation time, it was based on the PH model. Thus, it could potentially over- or under-estimate gene effects if the PH assumption is violated. In this section, we propose a more generalized test for gene effect based on KL divergence and the YP model, which does not rely on the PH assumption. We will then show that this contains tests for both the PH and PO models as special cases. The PO model allows for various types of hazards, including proportional and crossing, and the YP model is a generalized version of both the PH and PO models. Thus, with these new measures, we can utilize the computational benefits provided by the approach of Devarajan and Ebrahimi (2009), while also addressing the issue of NPH. Lastly, R^2 measures for the PH and PO models using these tests for gene effects are developed in §4.1.4.

4.1.1 General Framework: YP Model

The survival and hazard functions of the YP model in Equations 2.2.13 and 2.2.12, respectively, can be used to create the YP density function

$$f(t|z) = \lambda(t|z)S(t|z) = \frac{\lambda_0(t) \exp[(\beta + \gamma)z] S_0(t)^{\exp(\gamma z)} \exp(\gamma z)^{\exp(\gamma z)}}{[\exp(\beta z) - S_0(t) \exp(\beta z) + S_0(t) \exp(\gamma z)]^{\exp(\gamma z)+1}}. \quad (4.1.1)$$

Here, z refers to the gene expression for a given gene j , where $j = 1, \dots, p$. Note, when $\gamma = \beta$, it becomes the PH model, and when $\gamma = 0$, it becomes the PO model. Thus, it is a versatile, useful model that encompasses both the PH and PO models, as well as a host of other models.

Hypotheses We wish to test the null hypothesis

$$H_0 : \Lambda(t|z) = \Lambda_0(t) \quad (4.1.2)$$

against the alternative

$$H_1 : \Lambda(t|z) = \int_0^t \lambda(t|z) dt, \quad (4.1.3)$$

where $\lambda(t|z) = \frac{\lambda_0(t) \exp[(\beta+\gamma)z]}{\exp(\beta z) - S_0(t) \exp(\beta z) + S_0(t) \exp(\gamma z)}$ from the YP model. These hypotheses can be re-written as

$$H_0 : \beta = 0, \gamma = 0 \text{ against } H_1 : \beta \neq 0, \gamma \neq 0. \quad (4.1.4)$$

To summarize, we are testing whether the gene has an effect on survival time, $\beta \neq 0, \gamma \neq 0$, versus no effect $\beta = 0, \gamma = 0$.

Test Statistics We will now use KL divergence described in Equations 2.3.8 and 2.3.9 to build the test statistics. In §2.3.3, KL divergence is defined as $I_1(F : F_0) = \int_0^\infty f(x) \log \left\{ \frac{f(x)}{f_0(x)} \right\} dx$, where f and f_0 are the density functions. This measure is also called the directed divergence as it measures the discrepancy between F and F_0 in the direction of F . Alternatively, one can define $I_2(F_0 : F) = \int_0^\infty f_0(x) \log \left\{ \frac{f_0(x)}{f(x)} \right\} dx$, which measures the discrepancy between F and F_0 in the direction of F_0 . Thus, KL divergence measures the difference between two probability distributions. We denote $f(t|z)$ as f under H_1 and $f_0(t)$ as f_0 under H_0 , so both measures represent the divergence between the null and alternative hypotheses. We can rewrite I_1 and I_2 above as

$$I_1(F : F_0|z) = \int_0^\infty f(t|z) \log \left\{ \frac{f(t|z)}{f_0(t)} \right\} dt \quad (4.1.5)$$

and

$$I_2(F_0 : F|z) = \int_0^\infty f_0(t) \log \left\{ \frac{f_0(t)}{f(t|z)} \right\} dt. \quad (4.1.6)$$

These measures can be viewed as weighted log-likelihood ratios. For example, $I_2 \approx E_f \left(\log \left\{ \frac{f_0(t)}{f(t|z)} \right\} \right)$. For computational reasons, we focus our attention on I_2 .

We set $f(t|z) = \lambda(t|z)S(t|z)$ and $f_0(t) = \lambda_0(t)S_0(t)$ under the YP model using Equations 2.2.12 and 2.2.13, and we create the $I_{2,YP}(\beta, \gamma|z)$ measure as

$$\begin{aligned}
I_{2,YP}(\beta, \gamma|z) &= \int_0^{\infty} f_0(t) \log \left\{ \frac{\lambda_0(t)S_0(t)}{\lambda(t|z)S(t|z)} \right\} dt \\
&= -(\beta + \gamma)z - \gamma z \exp(\gamma z) - [\exp(\gamma z) - 1] \int_0^{\infty} f_0(t) \log [S_0(t)] dt \\
&\quad + [\exp(\gamma z) + 1] \int_0^{\infty} f_0(t) \log [\exp(\beta z) - S_0(t) \exp(\beta z) + S_0(t) \exp(\gamma z)] dt \\
&= \frac{\exp(\beta z) [\beta z \exp(\gamma z) - \gamma z - 2] - \exp(\gamma z) [\gamma z \exp(\beta z) - \beta z - 2]}{\exp(\beta z) - \exp(\gamma z)}.
\end{aligned} \tag{4.1.7}$$

Let z_{ij} represent the gene expression for individual i and gene j , where $i = 1, \dots, n$ and $j = 1, \dots, p$. To get an estimate for this measure, we replace β and γ with $\hat{\beta}$ and $\hat{\gamma}$, the maximum partial likelihood estimates using the approach in Yang and Prentice (2005). Using Equation 4.1.7, we sum over the individuals to get an estimate for our information statistic based on KL divergence. The measure for individual gene j is calculated as

$$\hat{I}_{2,YP}^j = \sum_{i=1}^n \frac{\exp(\hat{\beta}z_{ij}) [\hat{\beta}z_{ij} \exp(\hat{\gamma}z_{ij}) - \hat{\gamma}z_{ij} - 2] - \exp(\hat{\gamma}z_{ij}) [\hat{\gamma}z_{ij} \exp(\hat{\beta}z_{ij}) - \hat{\beta}z_{ij} - 2]}{\exp(\hat{\beta}z_{ij}) - \exp(\hat{\gamma}z_{ij})}. \tag{4.1.8}$$

In this thesis, we only consider the I_2 case, and thus, going forward, we will refer to this $\hat{I}_{2,YP}^j$ measure as \hat{I}_{YP} , where \hat{I}_{YP} can be calculated for each individual gene j .

Theorem 1. \hat{I}_{YP} is a maximum likelihood estimator and is asymptotically normal with mean I_{YP} .

Proof. Because $\hat{\beta}$ and $\hat{\gamma}$ are asymptotically normal with mean β and γ (Yang

and Prentice 2005) and because of the invariance property of maximum likelihood (ML) estimators, it can be concluded that our statistic is also asymptotically normal with the means above. See Devarajan and Ebrahimi (2009) for more details on this proof. \square

Using the ideas in Theorem 1, the variance of \hat{I}_{YP} can be computed as

$$\text{Var}(\hat{I}_{YP}) = \text{Var} \left(\sum_{i=1}^n \frac{\exp(\hat{\beta}z_{ij}) [\hat{\beta}z_{ij} \exp(\hat{\gamma}z_{ij}) - \hat{\gamma}z_{ij} - 2] - \exp(\hat{\gamma}z_{ij}) [\hat{\gamma}z_{ij} \exp(\hat{\beta}z_{ij}) - \hat{\beta}z_{ij} - 2]}{\exp(\hat{\beta}z_{ij}) - \exp(\hat{\gamma}z_{ij})} \right). \quad (4.1.9)$$

Because of the terms in the denominator, computation of the exact variance is complicated. Instead, we use the delta method and the Taylor series expansion to get an approximation for $\text{Var}(\hat{I}_{YP})$. In showing this calculation, we denote $I_{YP} = I$. The variance can be approximated using the quadratic Taylor series expansion,

$$I \approx I(0, 0) + I_{\beta}(0, 0)\beta + I_{\gamma}(0, 0)\gamma + \frac{1}{2}I_{\beta\beta}(0, 0)\beta^2 + I_{\beta\gamma}(0, 0)\beta\gamma + \frac{1}{2}I_{\gamma\gamma}(0, 0)\gamma^2, \quad (4.1.10)$$

$$\text{where } I_{\beta} = \frac{\partial I}{\partial \beta}, I_{\gamma} = \frac{\partial I}{\partial \gamma}, I_{\beta\beta} = \frac{\partial^2 I}{\partial \beta^2}, I_{\beta\gamma} = \frac{\partial^2 I}{\partial \beta \partial \gamma}, I_{\gamma\gamma} = \frac{\partial^2 I}{\partial \gamma^2}.$$

The derivatives noted above were calculated and the limits as $(\beta, \gamma) \rightarrow (0, 0)$ were found for each. The resulting limits are seen below.

$$\begin{aligned} \lim_{(\beta, \gamma) \rightarrow (0, 0)} I &= \lim_{(\beta, \gamma) \rightarrow (0, 0)} I_{\beta} = \lim_{(\beta, \gamma) \rightarrow (0, 0)} I_{\gamma} = 0 \\ \lim_{(\beta, \gamma) \rightarrow (0, 0)} I_{\beta\beta} &= \lim_{(\beta, \gamma) \rightarrow (0, 0)} I_{\gamma\gamma} = \frac{1}{3}z^2 \\ \lim_{(\beta, \gamma) \rightarrow (0, 0)} I_{\beta\gamma} &= \frac{1}{6}z^2 \end{aligned}$$

Thus, $I \approx 0 + 0 + 0 + \frac{1}{2}(\frac{1}{3}z^2)\beta^2 + \frac{1}{6}z^2\beta\gamma + \frac{1}{2}(\frac{1}{3}z^2)\gamma^2 = \frac{1}{6}z^2(\beta^2 + \beta\gamma + \gamma^2)$

We perform the following calculation under the assumption that β and γ are independent and both come from the distribution under H_0 , where the expected value of the parameter is zero. Therefore, the approximate $\text{Var}(\hat{I}_{YP})$ can be calculated as

$$\begin{aligned}
\text{Var}(\hat{I}_{YP}) &\approx \text{Cov} \left(\left[\frac{1}{6}z_i^2(\hat{\beta}^2 + \hat{\beta}\hat{\gamma} + \hat{\gamma}^2) \right], \left[\frac{1}{6}z_j^2(\hat{\beta}^2 + \hat{\beta}\hat{\gamma} + \hat{\gamma}^2) \right] \right) \\
&= \sum_{i=1}^n \sum_{j=1}^n \text{E} \left[\left(\frac{1}{6}z_i^2(\hat{\beta}^2 + \hat{\beta}\hat{\gamma} + \hat{\gamma}^2) \right) \left(\frac{1}{6}z_j^2(\hat{\beta}^2 + \hat{\beta}\hat{\gamma} + \hat{\gamma}^2) \right) \right] \\
&\quad - \text{E} \left[\frac{1}{6}z_i^2(\hat{\beta}^2 + \hat{\beta}\hat{\gamma} + \hat{\gamma}^2) \right] \text{E} \left[\frac{1}{6}z_j^2(\hat{\beta}^2 + \hat{\beta}\hat{\gamma} + \hat{\gamma}^2) \right] \\
&= \frac{1}{36} \sum_{i=1}^n \sum_{j=1}^n z_i^2 z_j^2 \left\{ \text{E} [\hat{\beta}^4] + 2\text{E} [\hat{\beta}^3\hat{\gamma}] + 3\text{E} [\hat{\beta}^2\hat{\gamma}^2] + 2\text{E} [\hat{\beta}\hat{\gamma}^3] + \text{E} [\hat{\gamma}^4] \right\} \\
&\quad - z_i^2 z_j^2 \left\{ \text{E} [\hat{\beta}^2] + \text{E} [\hat{\beta}\hat{\gamma}] + \text{E} [\hat{\gamma}^2] \right\}^2 \\
&= \frac{1}{36} \sum_{i=1}^n \sum_{j=1}^n z_i^2 z_j^2 \left\{ 2\sigma_{\hat{\beta}}^4 + \sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2 + 2\sigma_{\hat{\gamma}}^4 \right\},
\end{aligned} \tag{4.1.11}$$

where $\sigma_{\hat{\beta}}^2$ and $\sigma_{\hat{\gamma}}^2$ are the variances under H_0 for $\hat{\beta}$ and $\hat{\gamma}$, respectively. Using this measure and the ideas in Devarajan and Ebrahimi (2009), we create the statistical test

$$\text{Reject } H_0 \text{ if } \hat{\chi}_{YP}^2 = \frac{\hat{I}_{YP}^2}{\text{Var}(\hat{I}_{YP})} > \chi_1^2.$$

Despite the complexity of PO and YP models, these I measures have a computational advantage of not requiring an estimation of the baseline. We also note that parameter estimation in the YP model can only be done on

dichotomized covariates using currently available methods (Yang and Prentice 2005). Thus, for application purposes, $\hat{\beta}$ and $\hat{\gamma}$ for the YP model were obtained by fitting the model to covariates dichotomized by the median.

4.1.2 Special Cases: PH and PO

As described in §2.2.3, the YP model has two special cases, the PH and PO models. In Equations 2.2.12 and 2.2.13, setting $\gamma = \beta$ results in the Cox PH model and setting $\gamma = 0$ results in the PO model. Both of these models can be applied for quantitative gene expression and have one parameter, which we denote as β in their respective hazard functions. Thus, we are now looking at testing the null against the alternative hypotheses

$$H_0 : \beta = 0 \text{ against } H_1 : \beta \neq 0. \quad (4.1.12)$$

PO Case To obtain the test statistic for gene j in the PO case, we set $\hat{\gamma} = 0$ in Equation 4.1.8, which simplifies to

$$\hat{I}_{2,PO}^j = \sum_{i=1}^n \frac{\hat{\beta} z_{ij} \exp(\hat{\beta} z_{ij}) - 2 \exp(\hat{\beta} z_{ij}) + \hat{\beta} z_{ij} + 2}{\exp(\hat{\beta} z_{ij}) - 1}. \quad (4.1.13)$$

Here, $\hat{\beta}$ is estimated using the approach outlined in Martinussen and Scheike (2006) for the PO model. Using Equation 2.3.8 and the hazard function for the PO model, $\hat{I}_{1,PO}$ was calculated and found to have the same form as $\hat{I}_{2,PO}$. In other words, $I_{1,PO}(F : F_0|z) = I_{2,PO}(F_0 : F|z)$. This is not a required property of the two directed divergence measures. Because $\hat{I}_{1,PO} = \hat{I}_{2,PO}$, it

is sufficient to focus our attention on $\hat{I}_{2,PO}$, which will now be referred to as simply \hat{I}_{PO} . From Theorem 1, we know that \hat{I}_{PO} is a ML estimator and is asymptotically normal with mean I_{PO} .

The variance of \hat{I}_{PO} can be estimated by

$$\text{Var}(\hat{I}_{PO}) \approx \sum_{i=1}^n \sum_{j=1}^n \frac{1}{18} z_i^2 z_j^2 \sigma^4, \quad (4.1.14)$$

where σ is the variance under H_0 . This variance was estimated using the delta method and first three terms of the Taylor series expansion, and it is directly related to the $\sigma_\gamma = 0$ case in Equation 4.1.11. Using these measures and the ideas in Devarajan and Ebrahimi (2009), test statistics analogous to the generalized YP case can be formed,

$$\text{Reject } H_0 \text{ if } \hat{\chi}_{PO}^2 = \frac{\hat{I}_{PO}^2}{\widehat{\text{Var}}(\hat{I}_{PO})} > \chi_1^2.$$

PH Case To obtain the test statistic for the PH case, we first set $\gamma = \beta$ in Equation 4.1.8. This presents a form for $I_{2,PH}$ and a χ^2 test statistic can be created. It will result in the same statistical test described in Devarajan and Ebrahimi (2009), and although not explored in their paper, this measure could be used for gene selection. Their test is easy to compute; however, since it is based on the PH model, it could potentially over- or under-estimate gene effects if the PH assumption is violated. Thus, with the new YP and PO measures, we can provide the computational benefits of Devarajan and Ebrahimi (2009), while also addressing the issue of NPH.

4.1.3 \hat{I}_{PO} and \hat{I}_{YP} as Ranking Measures

In this section, we propose the use of \hat{I}_{PO} and \hat{I}_{YP} as ranking measures. As seen in Equation 4.1.6, \hat{I}_{PO} and \hat{I}_{YP} are measures of the divergence between the distribution functions F and F_0 from the PO and YP model, respectively, where $f(t|z)$ and $f_0(t)$ are the density functions under H_1 and H_0 , respectively. Thus, in each case, \hat{I} can be seen as a measure of the discrepancy between the two distribution functions, and also as the expected value of a log likelihood ratio. Because \hat{I} captures the divergence between H_1 and H_0 , we can use it to serve as a ranking measure. Essentially, \hat{I}_{PO} and \hat{I}_{YP} can be calculated for all p genes in a data set, and then these \hat{I} values can be ranked, greatest to least. The genes with higher \hat{I} values show a greater discrepancy between H_1 and H_0 and, thus, indicate a larger effect on survival based on the particular model chosen.

Similarly, we could compute respective test statistics $\hat{\chi}_{YP}^2 = \frac{\hat{I}_{YP}^2}{\hat{Var}(\hat{I}_{YP})}$ and $\hat{\chi}_{PO}^2 = \frac{\hat{I}_{PO}^2}{\hat{Var}(\hat{I}_{PO})}$ based on these models. These test statistics capture the divergence between the null and alternative hypotheses, while also taking the variance of each measure into account. Thus, we propose these test statistics as ranking measures or, equivalently, the p-values from these tests for gene selection. In §4.3, we will evaluate the performance of $\hat{\chi}_{YP}^2$ and $\hat{\chi}_{PO}^2$ in terms of gene selection. For easier readability, the test statistics in that section will be referred to as I_{YP} and I_{PO} .

4.1.4 R^2 Measures Using I_{PH} and I_{PO}

In this section, we utilize the tests for gene effect proposed in §4.1 and develop R^2 measures for the PH and PO models. First, recall that for the PO case, $I_{1,PO} = I_{2,PO} = I_{PO}$, and it has the form

$$I_{PO}(\beta|z) = \frac{\beta z \exp(\beta z) - 2 \exp(\beta z) + \beta z + 2}{\exp(\beta z) - 1} = \beta z - 2 + \frac{2\beta z}{\exp(\beta z) - 1}, \quad (4.1.15)$$

where z refers to gene expression for gene j , for $j = 1, \dots, p$. Similar measures were derived for the PH model in Devarajan and Ebrahimi (2009) and these are given by

$$I_{1,PH}(\beta|z) = \exp(-\beta z) + \beta z - 1 \quad \text{and} \quad I_{2,PH}(\beta|z) = \exp(\beta z) - \beta z - 1. \quad (4.1.16)$$

R_j^2 Measure Next, we develop an R^2 measure based on these I measures. First, we note that in our application, it is standard for \mathbf{z} to represent \log_2 gene expression. This normalizing log transformation stabilizes the variance, and then each gene is standardized by subtracting the mean expression and dividing by the standard deviation. Thus, we have $Z \equiv z_{ij} \stackrel{iid}{\sim} Normal(0, 1)$, where $i = 1, \dots, n$ and $j = 1, \dots, p$. We define \tilde{I} as the expectation of I with respect to the distribution of Z , which essentially integrates out the covariate. This differs from the approach for \hat{I} , where KL divergence is summed over the observations for each covariate. Here, \tilde{I} can be expressed as

$$\tilde{I} = \int_{-\infty}^{\infty} I(z) \times f_Z(z) dz \quad (4.1.17)$$

and is computed for the PH and PO models based on the respective I measures. Once \tilde{I} is calculated, R^2 is defined (Joe 1989; Soofi *et al.* 1995),

$$R_I^2 = 1 - \exp(-2\tilde{I}). \quad (4.1.18)$$

PO Case: $R_{I_{PO}}^2$ For the PO case, we first calculate its \tilde{I} as

$$\begin{aligned} \tilde{I}_{PO} &= \int_{-\infty}^{\infty} I_{PO}(z) f_Z(z) dz = \int_{-\infty}^{\infty} (\beta z - 2) f_Z(z) dz + \int_{-\infty}^{\infty} \frac{2\beta z}{\exp(\beta z) - 1} f_Z(z) dz \\ &= -2 + \int_{-\infty}^{\infty} \frac{2\beta z}{\exp(\beta z) - 1} f_Z(z) dz \\ &= -2 + E_f[g(z)], \end{aligned}$$

where $g(z) = \frac{2\beta z}{\exp(\beta z) - 1}$. Using the Taylor series expansion, we can estimate $E_f[g(z)]$ by

$$\begin{aligned} E_f[g(z)] &\approx g(\mu_z) + g'(\mu_z)E(z - \mu_z) + \frac{g''(\mu_z)}{2}E[(z - \mu_z)^2] \\ &= g(0) + \frac{g''(0)}{2}, \end{aligned}$$

where $g''(z) = \frac{2\beta^3 x \exp(2\beta x) - 4\beta^2 \exp(2\beta x) + 2\beta^3 x \exp(\beta x) + 4\beta^2 \exp(\beta x)}{[\exp(\beta x) - 1]^3}$. Now, take the $\lim_{z \rightarrow 0}$ for $g(z)$ and $g''(z)$ using L'Hopital's rule to get

$$\lim_{z \rightarrow 0} g(z) = 2 \text{ and } \lim_{z \rightarrow 0} g''(z) = \frac{1}{3}\beta^2.$$

Thus, $E[g(z)] \approx 2 + \frac{1}{2}(\frac{1}{3}\beta^2) = 2 + \frac{1}{6}\beta^2$, and therefore,

$$\tilde{I}_{PO} = -2 + 2 + \frac{1}{6}\beta^2 = \frac{1}{6}\beta^2. \quad (4.1.19)$$

The R^2 measure is then defined as

$$R_{\tilde{I}_{PO}}^2 = 1 - \exp(-2\tilde{I}_{PO}), \quad (4.1.20)$$

where β in \tilde{I}_{PO} is replaced by $\hat{\beta}$, the modified PL estimator in the PO model (Martinussen and Scheike 2006).

PH Case: $R_{\tilde{I}_{PH}}^2$ For the PH case, we first calculate \tilde{I} using $I_{2,PH}$ as

$$\begin{aligned} \tilde{I}_{2,PH} &= \int_{-\infty}^{\infty} I_{2,PH}(z) f_Z(z) dz = \int_{-\infty}^{\infty} \exp(\beta z) f_Z(z) dz + \int_{-\infty}^{\infty} (-\beta z - 1) f_Z(z) dz \\ &= \left[\int_{-\infty}^{\infty} \exp(\beta z) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz \right] - 1 \\ &= \exp\left(\frac{1}{2}\beta^2\right) - 1. \end{aligned} \quad (4.1.21)$$

Although not shown, $\tilde{I}_{1,PH}$ was found to have the same form as $\tilde{I}_{2,PH}$, and thus, $\tilde{I}_{1,PH} = \tilde{I}_{2,PH} = \tilde{I}_{PH}$. R^2 is defined as

$$R_{\tilde{I}_{PH}}^2 = 1 - \exp(-2\tilde{I}_{PH}). \quad (4.1.22)$$

where β in \tilde{I}_{PH} is replaced by $\hat{\beta}$, the PL estimator parameter in the PH model (Cox 1972).

These R^2 measures can be used to rank the genes from greatest to

least, where genes with larger R^2 values are interpreted as exhibiting larger effects on survival. These R^2 measures also offer a computational advantage. From Equations 4.1.19 and 4.1.21, we observe that both measures are simple functions of the model parameter β . In the next section, we propose alternative pseudo- R^2 type measures based on the PO and PLO models.

4.2 Pseudo- R^2 Measures

As described in §2.3.4, Rouam *et al.* (2010, 2011) developed pseudo- R^2 measures for gene selection based on the PH and CH models, using the partial likelihood for the respective models. These R^2 measures, denoted by R_{PH}^2 and R_{CH}^2 , can be interpreted in terms of the difference in gene expression between subjects experiencing and not experiencing the event of interest. Computationally, these methods are effective in that they do not require estimating model parameters. The measures can then be ranked from greatest to least, and genes with larger R^2 values are selected as having a significant effect on survival outcome. The issue with R_{PH}^2 is its use of the PH model, which relies on the proportional hazards assumption. In the CH model, the hazards ratio between two individuals with gene expression \mathbf{z} and \mathbf{z}^* cross over time (Rouam *et al.* 2011). Thus, while R_{CH}^2 does address the issue of proportional hazards in the PH model, it forces crossing hazards. Therefore, the method itself is specifically designed to identify crossing hazards.

In §4.2.1, we propose a generalized pseudo- R^2 measure that will encompass the measures created in Rouam *et al.* (2010, 2011), as well as our

new measures based on the PO and PLO models. The advantage of the PO model is that it allows for both proportional and non-proportional hazards, and therefore, is more applicable to the different hazard structures found in genomic data, as evidenced by the results shown in Table 3.1 where it was shown to fit genes that exhibit both PH and NPH. The PLO model generalizes the PO model and, thus, has a more versatile form. For this method, we will use a special case of the PLO model, where $\gamma = \beta$ in Equation 2.2.10. In §4.2.2, we also discuss a computational correction for R_{CH}^2 that helps improve its performance. Then in §4.2.3, we expand our work by developing new R^2 measures based on the likelihood ratio for the PO model, similar to what was done in Allison *et al.* (1995), O’Quigley *et al.* (2005), and Nagelkerke (1991).

4.2.1 Generalized Pseudo- R^2 Measure

In this section, we generalize the R^2 measures in Rouam *et al.* (2010, 2011) and show that the PH, CH, PO and PLO models are special cases. We begin by computing the derivative of the log partial likelihood with respect to the parameter β for each of the models above, which we will call the score function $U(\beta; t)$, evaluated at $\beta = 0$. R^2 for the PO and PLO models are derived below.

PO Case Given covariate z , the survival function and hazard function for the PO model can be written as

$$S(t|z) = \frac{S_0(t)}{S_0(t) - S_0(t) \exp(\beta z) + \exp(\beta z)} \quad (4.2.1)$$

and

$$\lambda(t|z) = -\frac{\partial \ln S(t|z)}{\partial t} = \frac{\lambda_0(t) \exp(\beta z)}{S_0(t) - S_0(t) \exp(\beta z) + \exp(\beta z)}, \quad (4.2.2)$$

respectively. The partial likelihood is written as

$$L(\beta) = \prod_{t \leq \tau} \prod_{i=1}^n \left\{ \frac{\lambda_i(t|z_i)}{\sum_{j=1}^n Y_j \lambda_j(t|z_j)} \right\}^{\Delta N_i(t)} = \prod_{t \leq \tau} \prod_{i=1}^n \left\{ \frac{\frac{\lambda_0(t) \exp(\beta z_i)}{S_0(t) - S_0(t) \exp(\beta z_i) + \exp(\beta z_i)}}{\sum_{j=1}^n \frac{Y_j \lambda_0(t) \exp(\beta z_j)}{S_0(t) - S_0(t) \exp(\beta z_j) + \exp(\beta z_j)}} \right\}^{\Delta N_i(t)},$$

where $Y_j = 1$ if the subject is at risk before time t , and z_i represents the gene expression for given gene j . For fixed t ,

$$\log L(\beta) = \sum_{i=1}^n \int_0^t \left\{ \beta z_i - \log [S_0(s) - S_0(s) \exp(\beta z_i) + \exp(\beta z_i)] - \log \left[\sum_{j=1}^n \frac{Y_j \exp(\beta z_j)}{S_0(s) - S_0(s) \exp(\beta z_j) + \exp(\beta z_j)} \right] \right\} dN_i(s).$$

Hence,

$$U(\beta; t) = \frac{d \log L(\beta)}{d\beta} = \sum_{i=1}^n \int_0^t \left\{ z_i - \frac{z_i \exp(\beta z_i) [1 - S_0(s)]}{S_0(s) - S_0(s) \exp(\beta z_i) + \exp(\beta z_i)} - \frac{\sum_{j=1}^n \frac{Y_j z_j \exp(\beta z_j) S_0(s)}{[S_0(s) - S_0(s) \exp(\beta z_j) + \exp(\beta z_j)]^2}}{\sum_{j=1}^n \frac{Y_j \exp(\beta z_j)}{S_0(s) - S_0(s) \exp(\beta z_j) + \exp(\beta z_j)}} \right\} dN_i(s), \quad (4.2.3)$$

and setting $\beta = 0$, we get

$$U(0; t) = \sum_{i=1}^n \int_0^t \left\{ z_i S_0(s) - \frac{\sum_{j=1}^n Y_j z_j S_0(t)}{\sum_{j=1}^n Y_j} \right\} dN_i(s) = \sum_{i=1}^n \int_0^t \left\{ w(s) \left[z_i - \frac{\sum_{j=1}^n Y_j z_j}{\sum_{j=1}^n Y_j} \right] \right\} dN_i(s), \quad (4.2.4)$$

where $w(s) = S_0(s)$.

PLO Case As seen in equation 2.2.10, the PLO model introduces another parameter into the PO model and has the form

$$\frac{1-S(t|z)}{S(t|z)} = \left[\frac{1-S_0(t)}{S_0(t)} \right]^{\exp(\gamma z)} \exp(\beta z).$$

It is useful because of this generalization, and it allows the hazard functions corresponding to two values of a covariate to cross. In this paper, we set $\gamma = \beta$ so we have

$$\frac{1-S(t|z)}{S(t|z)} = \left[\frac{1-S_0(t)}{S_0(t)} \right]^{\exp \beta z} \exp(\beta z).$$

Then, the survival function and hazard function for this special case of the PLO model can be written as

$$S(t|z) = \frac{1}{\exp(\beta z) \left[\frac{1-S_0(t)}{S_0(t)} \right]^{\exp(\beta z)} + 1} = \frac{S_0(t)^{\exp(\beta z)}}{\exp(\beta z) [1 - S_0(t)]^{\exp(\beta z)} + S_0(t)^{\exp(\beta z)}} \quad (4.2.5)$$

and

$$\lambda(t|z) = -\frac{\partial \ln S(t|z)}{\partial t} = \frac{\lambda_0(t) \exp(2\beta z)}{[1 - S_0(t)] \left[\exp(\beta z) + \left(\frac{S_0(t)}{1-S_0(t)} \right)^{\exp(\beta z)} \right]}, \quad (4.2.6)$$

respectively. The partial likelihood is written as

$$L(\beta) = \prod_{t \leq \tau} \prod_{i=1}^n \left\{ \frac{\lambda_i(t|z_i)}{\sum_{j=1}^n Y_j \lambda_j(t|z_j)} \right\}^{\Delta N_i(t)} = \prod_{t \leq \tau} \prod_{i=1}^n \left\{ \frac{\frac{\lambda_0(t) \exp(2\beta z_i)}{[1-S_0(t)] \left[\exp(\beta z_i) + \left(\frac{S_0(t)}{1-S_0(t)} \right)^{\exp(\beta z_i)} \right]}}{\sum_{j=1}^n \frac{Y_j \lambda_0(t) \exp(2\beta z_j)}{[1-S_0(t)] \left[\exp(\beta z_j) + \left(\frac{S_0(t)}{1-S_0(t)} \right)^{\exp(\beta z_j)} \right]}} \right\}^{\Delta N_i(t)},$$

where $Y_j = 1$ if the subject is at risk before time t , and for fixed t ,

$$\begin{aligned} \log L(\beta) = & \sum_{i=1}^n \int_0^t \left\{ 2\beta z_i - \log [1 - S_0(s)] - \log \left[\exp(\beta z_j) + \left(\frac{S_0(s)}{1-S_0(s)} \right)^{\exp(\beta z_j)} \right] \right. \\ & \left. - \log \left[\sum_{j=1}^n \frac{Y_j \exp(2\beta z_j)}{[1-S_0(s)] \left[\exp(\beta z_j) + \left(\frac{S_0(s)}{1-S_0(s)} \right)^{\exp(\beta z_j)} \right]} \right] \right\} dN_i(s). \end{aligned}$$

Hence,

$$\begin{aligned} U(\beta; t) = \frac{d \log L(\beta)}{d\beta} = & \sum_{i=1}^n \int_0^t \left\{ 2z_i - \frac{z_i \exp(\beta z_i) + z_i \exp(\beta z_i) \log \left(\frac{S_0(s)}{1-S_0(s)} \right) \left(\frac{S_0(s)}{1-S_0(s)} \right)^{\exp(\beta z_i)}}{\exp(\beta z_j) + \left(\frac{S_0(s)}{1-S_0(s)} \right)^{\exp(\beta z_j)}} \right. \\ & \left. - \frac{\sum_{j=1}^n \frac{Y_j z_j \exp(2\beta z_j) [1-S_0(s)] \left\{ 2 \left(\frac{S_0(s)}{1-S_0(s)} \right)^{\exp(\beta z_j)} - \log \left(\frac{S_0(s)}{1-S_0(s)} \right) \exp(\beta z_j) \left(\frac{S_0(s)}{1-S_0(s)} \right)^{\exp(\beta z_j)} + \exp(\beta z_j) \right\}}{\left\{ [1-S_0(s)] \left[\exp(\beta z_j) + \left(\frac{S_0(s)}{1-S_0(s)} \right)^{\exp(\beta z_j)} \right] \right\}^2}}{\sum_{j=1}^n \frac{Y_j \exp(2\beta z_j)}{[1-S_0(s)] \left[\exp(\beta z_j) + \left(\frac{S_0(s)}{1-S_0(s)} \right)^{\exp(\beta z_j)} \right]}} \right\} dN_i(s), \end{aligned} \tag{4.2.7}$$

and setting $\beta = 0$, we get

$$\begin{aligned}
U(0; t) &= \sum_{i=1}^n \int_0^t \left\{ 2z_i - \frac{z_i + z_i \log \left(\frac{S_0(s)}{1-S_0(s)} \right) \left(\frac{S_0(s)}{1-S_0(s)} \right)}{1 + \frac{S_0(s)}{1-S_0(s)}} \right. \\
&\quad \left. - \frac{Y_j z_j [1 - S_0(s)] \left\{ 2 \left(\frac{S_0(s)}{1-S_0(s)} \right) - \log \left(\frac{S_0(s)}{1-S_0(s)} \right) \left(\frac{S_0(s)}{1-S_0(s)} \right) + 1 \right\}}{\sum_{j=1}^n Y_j} \right\} dN_i(s) \\
&= \sum_{i=1}^n \int_0^t \left\{ z_i \left[1 + S_0(s) - S_0(s) \log \left(\frac{S_0(s)}{1-S_0(s)} \right) \right] - \frac{\sum_{j=1}^n Y_j z_j \left[1 + S_0(s) - S_0(s) \log \left(\frac{S_0(s)}{1-S_0(s)} \right) \right]}{\sum_{j=1}^n Y_j} \right\} dN_i(s) \\
&= \sum_{i=1}^n \int_0^t \left\{ w(s) \left[z_i - \frac{\sum_{j=1}^n Y_j z_j}{\sum_{j=1}^n Y_j} \right] \right\} dN_i(s), \tag{4.2.8}
\end{aligned}$$

where $w(s) = 1 + S_0(s) - S_0(s) \log \left(\frac{S_0(s)}{1-S_0(s)} \right)$.

General Form Equations 2.3.11, 4.2.4, and 4.2.8 suggest that the score function has the following generalized form

$$U(0; t) = \sum_{i=1}^n \int_0^t \left\{ w(s) \left[z_i - \frac{\sum_{j=1}^n Y_j z_j}{\sum_{j=1}^n Y_j} \right] \right\} dN_i(s) = \int_0^t \left\{ \frac{w(s)(\bar{Y} - 1)}{\bar{Y}} \left[z_i - \frac{\sum_{j \in R^*(t_i)} z_j}{\bar{Y} - 1} \right] \right\} dN_i(s), \tag{4.2.9}$$

where $w(s)$ is a weight function that specifies the model of interest, $\bar{Y} = \sum_{j=1}^n Y_j$ is the number of subjects at risk at time t_i and $R^*(t_i)$ is the set of individuals not experiencing the event at time t_i . Thus, it can be seen that this is a measure of the weighted difference in gene expression between subjects observed to experience the event of interest and those observed to not experience the event.

An estimate for $U_i(0; t)$ is given by

$$\hat{U}_i = \delta_i \hat{w}(t_i) \left(z_i - \frac{\sum_{j=1}^n Y_j z_j}{\sum_{j=1}^n Y_j} \right), \quad (4.2.10)$$

where the estimation of $\hat{w}(t_i)$ for each model is described in Table 4.1 and δ_i is the indicator of failure at time t_i .

It would be ideal to have scores that are independently distributed. For this reason, we introduce W_i , a robust score created by Lin and Wei (1989).

$$W_i(0; t) = \int_0^t \left\{ w(s) \left[z_i - \frac{s^{(0)}(t)}{s^{(1)}(t)} \right] \right\} dN_i(s), \quad (4.2.11)$$

where $s^{(r)}(t) = E[S^{(r)}(t)]$. This W_i can be estimated by

$$\hat{W}_i = \hat{U}_i - \hat{E}\hat{U}_i = \delta_i \hat{w}(t_i) \left(z_i - \frac{\sum_{j=1}^n Y_j z_j}{\sum_{j=1}^n Y_j} \right) - \sum_{j=1}^n \frac{\delta_j \hat{w}(t_j) Y_i}{\sum_{r=1}^n Y_r} \left(z_i - \frac{\sum_{r=1}^n Y_r z_r}{\sum_{r=1}^n Y_r} \right). \quad (4.2.12)$$

The sum of W_i is identical to the sum of U_i , but W_i are independent.

The index can then be written as

$$R^2 = \frac{1}{k} \frac{\left(\sum_{i=1}^n \hat{W}_i \right)^2}{\sum_{i=1}^n \hat{W}_i^2}, \quad (4.2.13)$$

where k is the number of uncensored failure times. It can be seen as the robust score statistic divided by the number of distinct uncensored failure times. This

index falls between 0 and 1 and can be interpreted in terms of the percentage of separability of subjects experiencing and not experiencing the event in relation to the gene expression.

This generalized index will result in indices corresponding to the PH, CH, PO and PLO models based on the choice of the weight, $w(t)$, determined by the respective log partial likelihood. The estimated weights, $\hat{w}(t)$, for each special case are shown in Table 4.1. These weights are then used in the calculation in Equation 4.2.10 to create the respective measure in Equation 4.2.13. $\hat{\Lambda}_0(t)$ is estimated by the left-continuous version of the Nelson’s estimator, and $\hat{S}_0(t)$ is estimated using the Kaplan-Meier (KM) estimator. We also note that R_{PH}^2 and R_{CH}^2 are identical to the measures described in Rouam *et al.* (2010, 2011), respectively, but R_{PO}^2 and R_{PLO}^2 are our newly proposed measures that allow for, but not necessarily enforce, NPH.

Table 4.1: R^2 Measure: Special Cases

Model	Measure	Weight, $\hat{w}(t)$
PH	R_{PH}^2	1
CH	R_{CH}^2	$1 + \log\{\hat{\Lambda}_0(t)\}$
PO	R_{PO}^2	$\hat{S}_0(t)$
PLO	R_{PLO}^2	$1 + \hat{S}_0(t) - \hat{S}_0(t) \log\left(\frac{\hat{S}_0(t)}{1 - \hat{S}_0(t)}\right)$

4.2.2 Computational Correction for R_{CH}^2 and R_{PLO}^2

The measure R_{CH}^2 has weight $\hat{w}(t_i) = (1 + \log \Lambda_0(t_i))$, where $\Lambda_0(t_i)$ is estimated by the left-continuous version of Nelson’s estimator. However, this weight has an inherent computational issue when $\Lambda_0(t_i) = 0$. To handle this error, Rouam *et al.* (2011) sets $\hat{w}(t_i) = 1$ if $\Lambda_0(t_i) = 0$, implying that $\log(\Lambda_0(t_i)) \rightarrow 0$ as

$\Lambda_0(t_i) \rightarrow 0$. This is unrealistic because $\log(\Lambda_0(t_i)) \rightarrow -\infty$ as $\Lambda_0(t_i) \rightarrow 0$. Thus, we propose an empirical computational correction for this error that uses a plot of the cumulative hazard versus the weight to obtain an approximation for the weight as the cumulative hazard approaches zero. In our algorithm, we set $\hat{w}(t_i)$ equal to this approximation when $\Lambda_0(t_i) = 0$. We call this modified measure R_{ModCH}^2 . Similarly, for R_{PLO}^2 , an empirical correction was made to account for the computational issue when $S_0(t_i) = 1$ by obtaining a graphical approximation for the weight as the survival approaches 1.

4.2.3 Likelihood Ratio (LR) Based R^2 Measures for the PO Model

Recall from §2.3.4, we discussed various other R^2 measures, each of which used the log-partial likelihood from the PH model as its foundation. In this section, we propose an extension of these measures to the PO model.

First, we note that the modified partial likelihood was calculated for the PO model in §4.2.1. Its log has the form

$$\log L(\beta) = \sum_{i=1}^n \int_0^t \left\{ \beta z_i - \log [S_0(s) - S_0(s) \exp(\beta z_i) + \exp(\beta z_i)] - \log \left[\sum_{j=1}^n \frac{Y_j \exp(\beta z_j)}{S_0(s) - S_0(s) \exp(\beta z_j) + \exp(\beta z_j)} \right] \right\} dN_i(s),$$

and its value under the null hypotheses is written as $\log L(0)$. β is estimated using the approach outlined in Martinussen and Scheike (2006). We then define three measures based on the differences between these two values. In other words, the measures are based on the log likelihood ratios, $\log \frac{L(\hat{\beta})}{L(0)} =$

$\log L(\hat{\beta}) - \log L(0)$.

The first measure is based on Allison's index (Allison *et al.* 1995), which uses a transformation of the log partial likelihood ratio test. It has the form

$$R_{LR,A}^2 = 1 - \exp\left(-\frac{2}{N} \times [\log L(\hat{\beta}) - \log L(0)]\right). \quad (4.2.14)$$

Here, N is the number of subjects. The second measure is a modified version of Allison's index based on the work in O'Quigley *et al.* (2005). It divides the log-likelihood by k , the number of failures, and is less sensitive to censoring, which is beneficial to our application where most real datasets have a high fraction of censored observations. It is given by

$$R_{LR,O}^2 = 1 - \exp\left(-\frac{2}{k} \times [\log L(\hat{\beta}) - \log L(0)]\right). \quad (4.2.15)$$

The last measure is based on Nagelkerke's index in Nagelkerke (1991). It also modifies Allison's index by dividing the initial index by its maximum possible value. It has the form

$$R_{LR,N}^2 = \frac{R_{LR,A}^2}{R_{max}^2}, \quad (4.2.16)$$

where $R_{max}^2 = 1 - \exp\left(\frac{2}{N} \times \log L(0)\right)$.

While these measures result in different values and ranges for a specified data set, we note from empirical observation that their rankings are the same. In our simulated study, we found that although the values differ, $R_{LR,A}^2$, $R_{LR,O}^2$, and $R_{LR,N}^2$ ranked the genes in the same exact order. Thus, because our focus is on ranking genes, we will choose to use $R_{LR,O}^2$ for the remainder of the analysis. For simplicity of notation, we will refer to $R_{LR,O}^2$ as R_{LR}^2 going

forward.

4.3 Simulations and Examples

In the previous two sections, we proposed various measures for variable selection and ranking. The first two measures, I_{PO} and I_{YP} , were based on a test for gene effect using KL divergence that essentially quantifies the separation between the null (no gene effect) and alternative (gene effect) hypotheses. $R_{I_{PO}}^2$ and $R_{I_{PH}}^2$ then used these tests for gene effect to create respective R^2 measures. Then, several R^2 -type measures were proposed, which we defined as R_{PO}^2 , R_{PLO}^2 , R_{ModCH}^2 , and R_{LR}^2 . First, R_{PO}^2 and R_{PLO}^2 used the score function for the PO and PLO models, respectively, to create R^2 measures that can be interpreted as a percentage of separability between covariates of those experiencing and not experiencing the events of interest. Then, R_{ModCH}^2 proposed a computational correction for the R_{CH}^2 measure in Rouam *et al.* (2011). Lastly, R_{LR}^2 was based on the likelihood ratio of the PO model. In this section, we apply these measures to simulated data sets with varying types of hazards and compare them to existing variable selection measures, such as R_{PH}^2 from Rouam *et al.* (2010), R_{CH}^2 from Rouam *et al.* (2011) and \hat{c}'_+ based on the concordance regression from Dunkler *et al.* (2010) described in §2.3.2.

4.3.1 Simulation Schemes

We use two different simulation algorithms to create various survival data sets based on a method described in Dunkler *et al.* (2010). To account

for various types of hazards, survival times, Y were generated from 5 different distributions: standard log-normal (LN), log-logistic (LL1, LL2), and Weibull (W1, W2). LL1 and W1 refer to the case where the shape parameters are the same but the scale parameters differ, and LL2 and W2 refer to the cases where both the shape and scale parameters differ. We use a more informed, broad approach compared to Dunkler *et al.* (2010), who only considered survival times coming from the W1 model. Here, the LN, LL2 and W2 cases are of particular interest because of their ability to simulate crossing hazards. To simulate censoring, we drew random samples with uniform follow-up times X from $U(0, \tau)$ and defined the observed survival time as $T = \min(Y, X)$ with censoring indicator $I(Y < X)$. We chose τ to get censoring proportions of 0, 33, and 67%.

For each case, we simulated censored survival times and gene expression data for $N = 200$ subjects and $p = 5000$ mock genes using the algorithms outlined below, where gene expression is linked to the survival time in an algorithm based on the log hazard ratio, $\beta_g(t) = \beta_0 \log(HR)$. This $\log(HR)$ is calculated based on the respective model of interest using the distributions described in Klein and Moeschberger (2003). Because of the complexity in the LN distribution, we use $\beta_g(t) = \beta_0(t^2 - 1)$ for this case to simulate crossing hazards, similar to what was done in Dunkler *et al.* (2010). Then, β_0 was chosen so that only the first 400 genes were assumed to have an effect on survival time, with 200 having a large effect and 200 having a small effect. We considered two simulation schemes. Scheme 1 is a univariate approach where gene expression is linked to survival one gene at a time, and Scheme 2

takes a multivariate approach that includes various correlations between gene expressions. These schemes are outlined in detail below.

Simulation Scheme 1: Univariate approach

1. Using the approach in Klein and Moeschberger (2003) and Dunkler *et al.* (2010), generate N survival times, $y_i, i = 1, \dots, N$, from one of the following:

- LN ($\mu = 0, \sigma = 1$)
- LL1 ($\alpha_1 = 2, \lambda_1 = 2, \lambda_2 = 4$)
- LL2 ($\alpha_1 = 3, \alpha_2 = 4, \lambda_1 = 1, \lambda_2 = 2$)
- W1 ($\alpha_1 = 1, \lambda_1 = \frac{1}{2}$)
- W2 ($\alpha_1 = 3, \alpha_2 = 2, \lambda_1 = 1, \lambda_2 = \frac{1}{2}$)

where (α_1, α_2) and (λ_1, λ_2) are the shape and scale parameters, respectively. In the LN model, μ and σ are the location and scale parameters, respectively.

2. Generate N follow-up times, $z_i, i = 1, \dots, N$, from $\text{Uniform}(0, \tau)$. Note, τ was chosen to produce censoring proportions of 0%, 33%, and 67%.
3. Let $d_i = I(y_i < z_i)$ and $t_i = \min(y_i, z_i)$, and sort the N tuples (t_i, d_i) so that $t_i < t_{i+1}$.
4. For each subject, $j = 1, \dots, N$, draw a gene expression value a_j from $\text{Normal}(0, 1)$.

5. In this step, we assign each tuple (t_i, d_i) a gene expression x_i , which is sampled from the $a_j, j = 1, \dots, N$. We start with the smallest observed survival time t_1 and the corresponding risk set $R_1 = 1, \dots, N$. For each $i = 1, \dots, N$, sample a subject $j^* \in R_i$, whose gene expression vector a_{j^*} will be assigned to x_i as follows:
 - (a) If $d_i = 0$, randomly sample j^* from the risk set R_i . Remove that subject from the risk set.
 - (b) If $d_i = 1$, sample j^* from the risk set R_i by assigning sample probabilities proportional to $\exp[x_j \beta_g(t)]$ to the subjects $j \in R_i$. Remove that subject from the risk set.
6. Repeat steps (4) and (5) p times, using the appropriate $\beta_g(t)$ each time to get the desired amount of small, large, and zero effect genes.

Simulation Scheme 2: Multivariate approach This algorithm is similar to Scheme 1, but includes various correlations between covariates.

1. Follow steps (1), (2) and (3) from Simulation Scheme 1.
2. For each subject, $j = 1, \dots, N$, draw gene expression values a_{jg} as follows:

$$a_{jg} = \begin{cases} -1 + \varepsilon_{jg} & \text{if } j \leq 0.5n, g \leq 0.05p & .50 \\ 1 + \varepsilon_{jg} & \text{if } j > 0.5n, g \leq 0.05p & .50 \\ 1.5 \times I(u_{j1} < 0.4) + \varepsilon_{jg} & \text{if } 0.05 < g \leq 0.1p & .35 \\ 0.5 \times I(u_{j2} < 0.7) + \varepsilon_{jg} & \text{if } 0.1 < g \leq 0.2p & .05 \\ 1.5 \times I(u_{j3} < 0.3) + \varepsilon_{jg} & \text{if } 0.2 < g \leq 0.3p & .32 \\ \varepsilon_{jg} & \text{if } g > 0.3p & 0 \end{cases}$$

, where the third column represents the approximate correlations that result, $\varepsilon_{jg} \sim N(0, 1)$ denotes a standard normally distributed error term, u_{jg} is a uniform random variable in the range $[0, 1]$, and I is the indicator function that assumes a value of 1 (argument is true) or 0 (argument is false).

3. In this step, we assign each tuple (t_i, d_i) a gene expression x_i , which is sampled from the $a_j, j = 1, \dots, N$. We start with the smallest observed survival time t_1 and the corresponding risk set $R_1 = 1, \dots, N$. For each $i = 1, \dots, N$, sample a subject $j^* \in R_i$, whose gene expression vector a_{j^*} will be assigned to x_i as follows:

- (a) If $d_i = 0$, randomly sample j^* from the risk set R_i . Remove that subject from the risk set.
- (b) If $d_i = 1$, sample j^* from the risk set R_i by assigning sample probabilities proportional to $\exp \left[\sum_{g=1}^p x_{jg} \beta_g(t) \right]$ to the subjects $j \in R_i$. Remove that subject from the risk set.

For each scheme and censoring combination, 200 data sets were generated and assessed. Each data set was analyzed using the assessment measures

discussed below, ordering the genes based on the values produced by each method from greatest to least. Average AUC, its 95% confidence interval, sensitivity and specificity are calculated across the 200 simulations in each case and used to compare the methods. We let P be the number of positive genes (the 400 genes having an effect on survival), N be the number of negative genes (the 4600 not having an effect on survival), and TP and TN be the number of true positive and true negative genes selected, respectively. Then, these measures are described as follows:

- Sensitivity: true positive rate (TPR), where $TPR = TP/P$
- Specificity: true negative rate (TNR), where $TNR = TN/N$
- ROC Curve: plot TPR verses false positive rate ($FPR = 1-TNR$) at various thresholds
- AUC: area under the ROC curve

We also use the Youden index (Youden 1950) to compare all of the methods. This measure, also known as Youden’s J statistic, is a single value that captures the performance of each method and is calculated as $J = sensitivity + specificity - 1$.

4.3.2 Simulation Results

In this section, we analyze the performance of our variable selection methods using the simulation schemes outlined in §4.3.1. First, we look at the results for scheme 1, the univariate approach. Table 4.2 reports the AUCs

for each method across the five simulation schemes, LN, LL1, LL2, W1 and W2. We begin by noting that although not included, the standard deviation of the AUCs were all very small, ranging from 0.01 to 0.28 (data not shown). We observe several scenarios in which one or more of the proposed methods outperform existing methods.

First, we look at the performance of our three *PO*-based R^2 measures, R_{PO}^2 , R_{LR}^2 , and $R_{\bar{I}PO}^2$. In each scenario and across all censoring schemes, these three measures perform almost identically. In some instances, such as LL2 0% and 33%, we do see a slight improvement in $R_{\bar{I}PO}^2$. Next, we compare the *PH* based measures $R_{\bar{I}PH}^2$ and R_{PH}^2 . From Table 4.2, we observe that the AUCs are almost identical for the two measures, with a slight improvement in some cases for $R_{\bar{I}PH}^2$. In fact, looking at the Youden indices in Table 4.4, we note that $R_{\bar{I}PH}^2$ outperforms R_{PH}^2 for all three LL2 cases and for the 0% censoring case of LN. This also holds true for the AUCs.

Next, we examine the LN case, where we recognize R_{PH}^2 and \hat{c}'_+ being outperformed by various measures. Specifically, R_{CH}^2 and R_{ModCH}^2 perform similarly and outperform the *PO* and *PH*-based R^2 measures, except for the 67% censoring case where R_{PO}^2 , R_{LR}^2 , and $R_{\bar{I}PO}^2$ perform the best. This is not surprising since the LN model allows for crossing hazards. R_{PLO}^2 performs well for lower censoring, but its AUC decreases as censoring increases. As censoring increases, the *PO*-based R^2 measures also outperform the *PH*-based measures, specifically in terms of their Youden indices seen in Table 4.4. I_{YP} outperforms I_{PO} and \hat{c}'_+ in the 0 and 33% censoring cases, but its performance decreases as censoring increases. In its application, I_{YP} is only capable of being applied

Table 4.2: Scheme 1 Results, AUCs

		LN	LL1	LL2	W1	W2
0%	R_{PO}^2	0.81	1.00	0.73	1.00	1.00
	R_{LR}^2	0.81	1.00	0.73	1.00	1.00
	R_{IPO}^2	0.81	1.00	0.75	1.00	1.00
	R_{PLO}^2	0.99	0.86	0.84	0.93	1.00
	R_{CH}^2	1.00	0.68	0.87	0.65	0.98
	R_{ModCH}^2	1.00	0.73	0.89	0.69	0.98
	R_{IPH}^2	0.92	0.99	0.54	1.00	1.00
	R_{PH}^2	0.87	0.99	0.50	1.00	1.00
	I_{PO}	0.80	1.00	0.72	1.00	1.00
	I_{YP}	1.00	0.97	0.82	1.00	1.00
	\hat{c}'_+	0.81	1.00	0.73	1.00	1.00
	33%	R_{PO}^2	0.89	0.99	0.77	0.99
R_{LR}^2		0.88	0.99	0.77	1.00	1.00
R_{IPO}^2		0.88	0.99	0.78	1.00	1.00
R_{PLO}^2		0.82	0.69	0.80	0.82	0.87
R_{CH}^2		0.98	0.82	0.86	0.70	0.76
R_{ModCH}^2		0.97	0.84	0.87	0.74	0.82
R_{IPH}^2		0.77	0.99	0.64	1.00	1.00
R_{PH}^2		0.77	0.99	0.61	1.00	1.00
I_{PO}		0.88	0.99	0.76	0.99	0.99
I_{YP}		0.95	0.91	0.76	0.98	1.00
\hat{c}'_+		0.82	0.99	0.71	0.99	1.00
67%		R_{PO}^2	0.93	0.97	0.81	0.99
	R_{LR}^2	0.93	0.98	0.81	0.99	0.95
	R_{IPO}^2	0.93	0.98	0.81	0.99	0.95
	R_{PLO}^2	0.68	0.63	0.74	0.51	0.89
	R_{CH}^2	0.88	0.90	0.83	0.87	0.57
	R_{ModCH}^2	0.87	0.90	0.84	0.87	0.71
	R_{IPH}^2	0.92	0.98	0.78	0.99	0.97
	R_{PH}^2	0.92	0.97	0.78	0.99	0.97
	I_{PO}	0.93	0.97	0.81	0.99	0.94
	I_{YP}	0.83	0.78	0.62	0.86	0.94
	\hat{c}'_+	0.88	0.96	0.71	0.98	0.98

to dichotomized covariates. Thus, I_{YP} 's performance may be affected by its inability to accommodate actual gene expression. I_{PO} also outperforms \hat{c}'_+ as censoring increases. This is also evident in the Youden index in Table 4.4.

For the LL1 case, we expect PO-based measures to perform well since the log logistic model is related to the PO model, and our results provide evidence in support. While R_{CH}^2 and R_{ModCH}^2 do improve as censoring increases, R_{PO}^2 , R_{LR}^2 , and $R_{I_{PO}}^2$ still outperform R_{CH}^2 and R_{ModCH}^2 at each censoring level. Similar to the LN case, the performance of R_{PLO}^2 and I_{YP} decreases as censoring increases. In the LL2 case, R_{PH}^2 and $R_{I_{PH}}^2$ perform significantly worse than the other R^2 measures, but their AUCs do increase from approximately .50 – .54 in the 0% censoring case to .78 in the 67% censoring case. Thus, they do show improvement with higher censoring, but they are still consistently outperformed. This is also evident in Table 4.4, where R_{PH}^2 and $R_{I_{PH}}^2$ have the lowest Youden index of all the reported measures. This log logistic model allows for crossing hazards and is related to the PLO model, so not surprisingly, R_{PLO}^2 outperforms the *PO*-based R^2 measures in the 0 and 33% censoring cases, but similar to the LN and LL1 case, we see its AUCs drop in the 67% case. \hat{c}'_+ is outperformed by either I_{PO} or I_{YP} , as well as R_{PLO}^2 , at each censoring level, with I_{YP} performing better for lower censoring and I_{PO} performing better for higher censoring. In this case, we also observe that the *PO*-based R^2 measures outperform the *PH*-based methods, which emphasizes the PO model's ability to handle potential crossing hazards.

For the W1 case, we observe R_{PO}^2 , R_{LR}^2 , and $R_{I_{PO}}^2$ outperform R_{CH}^2 and R_{ModCH}^2 at all censoring levels. This result is intuitive because this Weibull

model is related to the PH model and the PO model does allow for PH. For the W2 case, we see R_{PLO}^2 outperform R_{CH}^2 and R_{ModCH}^2 at all censoring levels, especially for higher censoring. This W2 case allows for crossing hazards, and yet here, we observe a clear advantage for our PO and PLO based measures over R_{CH}^2 which was purposefully designed to handle crossing hazards. Also, in this case, we observe that R_{ModCH}^2 , the proposed modification to R_{CH}^2 , performs significantly better than R_{CH}^2 as the censoring proportion increases.

Next, we look at the results for scheme 2 shown in Table 4.3. In general, the AUCs are observed to be slightly lower than those in Table 4.2, but this is due to the complexity of the scheme itself, where correlations are introduced between genes. The trend of the results, however, mimic what was seen in scheme 1. We note that the standard deviation of the AUCs were also very small, ranging from 0.01 to 0.28 (data not shown).

Similar to scheme 1, the *PO*-based R^2 measures, R_{PO}^2 , R_{LR}^2 , and $R_{I_{PO}}^2$, perform almost identically across all scenarios and censoring levels. The *PH*-based measures, $R_{I_{PH}}^2$ and R_{PH}^2 , also perform similarly. In the LN case, which allows for crossing hazards, R_{PH}^2 and $R_{I_{PH}}^2$ are outperformed by the proposed *PO*-based R^2 measures, R_{CH}^2 and R_{ModCH}^2 for the 0 and 33% censoring cases, but for the 67% censoring case, the *PO* and *PH*-based measures perform similarly and better than the other R^2 measures. The AUCs for R_{PLO}^2 , R_{CH}^2 , and R_{ModCH}^2 decrease as censoring increases, with the *PO* and *PH*-based measures having a clear advantage in the 67% censoring case. \hat{c}'_+ is also outperformed by I_{PO} and I_{YP} in the 0 and 33% censoring cases and by I_{PO} in the 67% censoring case. These differences are also evident in the Youden values shown in Table

Table 4.3: Scheme 2 Results, AUCs

		LN	LL1	LL2	W1	W2
0%	R_{PO}^2	0.84	0.91	0.55	0.91	0.90
	R_{LR}^2	0.84	0.91	0.56	0.91	0.90
	R_{IPO}^2	0.82	0.91	0.57	0.91	0.90
	R_{PLO}^2	0.76	0.51	0.88	0.51	0.88
	R_{CH}^2	0.89	0.86	0.89	0.86	0.79
	R_{ModCH}^2	0.89	0.87	0.89	0.86	0.81
	R_{IPH}^2	0.55	0.91	0.63	0.91	0.90
	R_{PH}^2	0.55	0.90	0.66	0.91	0.89
	I_{PO}	0.84	0.91	0.53	0.91	0.90
	I_{YP}	0.86	0.90	0.76	0.90	0.89
	\hat{c}'_+	0.82	0.91	0.50	0.91	0.90
	33%	R_{PO}^2	0.87	0.90	0.63	0.90
R_{LR}^2		0.88	0.91	0.61	0.91	0.90
R_{IPO}^2		0.86	0.91	0.64	0.91	0.90
R_{PLO}^2		0.59	0.51	0.87	0.53	0.88
R_{CH}^2		0.89	0.87	0.90	0.87	0.70
R_{ModCH}^2		0.88	0.87	0.89	0.87	0.74
R_{IPH}^2		0.72	0.90	0.57	0.90	0.90
R_{PH}^2		0.75	0.90	0.54	0.90	0.89
I_{PO}		0.87	0.90	0.61	0.90	0.89
I_{YP}		0.84	0.87	0.82	0.89	0.88
\hat{c}'_+		0.80	0.91	0.50	0.91	0.90
67%		R_{PO}^2	0.89	0.90	0.81	0.90
	R_{LR}^2	0.89	0.90	0.80	0.90	0.89
	R_{IPO}^2	0.88	0.90	0.80	0.90	0.89
	R_{PLO}^2	0.56	0.50	0.71	0.50	0.87
	R_{CH}^2	0.81	0.88	0.88	0.88	0.53
	R_{ModCH}^2	0.81	0.88	0.87	0.88	0.55
	R_{IPH}^2	0.87	0.90	0.70	0.90	0.89
	R_{PH}^2	0.89	0.90	0.71	0.90	0.87
	I_{PO}	0.89	0.90	0.80	0.90	0.85
	I_{YP}	0.69	0.75	0.67	0.77	0.74
	\hat{c}'_+	0.84	0.90	0.53	0.90	0.87

4.4. For the LL1 case, the *PO* and *PH*-based R^2 measures perform similarly and produce slightly higher AUCs than R_{CH}^2 and R_{ModCH}^2 . This result is also intuitive since the log-logistic model is related to the PO model which allows for both proportional and non-proportional hazards. They also significantly outperform R_{PLO}^2 at each censoring level. I_{PO} and \hat{c}'_+ perform similarly and outperform I_{YP} , whose AUCs decrease as censoring increases. For the LL2 case, which allows for crossing hazards, R_{PLO}^2 performs similarly to R_{CH}^2 and R_{ModCH}^2 , and they all significantly outperform the *PO* and *PH*-based R^2 measures. This results is expected since the LL2 case allows for crossing hazards. However, similar to Scheme 1, R_{PLO}^2 's performance falls in the 67% censoring case. In this case, we observe that the performance of \hat{c}'_+ is poor and no better than a coin toss. I_{YP} outperforms I_{PO} in the 0 and 33% censoring cases, but consistent with previous results, its AUC drops in the 67% censoring case. However, its performance may be affected by its inability to accommodate continuous gene expression. For the W1 case, we observe results similar to scheme 1. R_{PO}^2 , R_{LR}^2 , and $R_{I_{PO}}^2$ outperform R_{CH}^2 and R_{ModCH}^2 at all censoring levels, which is intuitive since the PO model allows for PH and this Weibull model is related to the PH model. For the W2 case, R_{PLO}^2 once again outperforms R_{CH}^2 and R_{ModCH}^2 at all censoring levels, especially for higher censoring. In this case, we also observe that R_{ModCH}^2 performs slightly better than R_{CH}^2 .

Overall, throughout different scenarios (LN, LL1, LL2, W1, W2) and censoring schemes (0, 33, 67%), we observe that the proposed methods outperform and, in some cases, perform as well as existing methods. In most cases, we notice some form of improvement. Our PO-based methods are strong, and

Table 4.4: Youden Index, Scheme 1 & 2

	Scheme 1						Scheme 2					
	0%		33%		67%		0%		33%		67%	
	LN	LL2	LN	LL2	LN	LL2	LN	LL2	LN	LL2	LN	LL2
R_{PO}^2	0.53	0.31	0.62	0.39	0.68	0.48	0.23	0.04	0.25	0.10	0.26	0.24
R_{LR}^2	0.53	0.33	0.60	0.41	0.68	0.49	0.23	0.05	0.25	0.07	0.26	0.22
$R_{\bar{I}^{PO}}^2$	0.52	0.36	0.59	0.43	0.68	0.50	0.20	0.06	0.23	0.11	0.25	0.19
R_{PLO}^2	0.89	0.53	0.41	0.43	0.24	0.26	0.17	0.23	0.06	0.22	0.07	0.14
R_{CH}^2	0.99	0.59	0.82	0.56	0.60	0.52	0.26	0.28	0.25	0.28	0.23	0.28
R_{ModCH}^2	0.99	0.61	0.80	0.58	0.60	0.54	0.26	0.28	0.24	0.28	0.23	0.28
$R_{\bar{I}^{PH}}^2$	0.66	0.04	0.49	0.18	0.66	0.44	0.02	0.06	0.13	0.03	0.24	0.13
R_{PH}^2	0.46	0	0.49	0.12	0.66	0.41	0.04	0.09	0.17	0.02	0.25	0.16
I_{PO}	0.51	0.29	0.60	0.37	0.67	0.48	0.22	0.02	0.24	0.07	0.25	0.23
I_{YP}	0.99	0.51	0.44	0.25	0.38	0.01	0.21	0.16	0.11	0.15	0.07	0.05
\hat{c}'_+	0.53	0.33	0.54	0.28	0.61	0.29	0.21	0.02	0.19	0.00	0.22	0.01

I_{YP} performs particularly well for lower censoring. R_{PLO}^2 performs similarly to R_{CH}^2 in many cases, but it is important to note that our modification, R_{ModCH}^2 , performs slightly better than R_{CH}^2 in many cases. Overall, depending on the simulation scheme and type of NPH present, we can find benefits for each of our measures. Most importantly, the proposed variable selection methods are more flexible and generalize existing methods.

4.3.3 Examples

In the previous section, we studied how our variable selection methods, specifically R_{PO}^2 , R_{LR}^2 , R_{PLO}^2 , $R_{\bar{I}^{PO}}^2$, R_{ModCH}^2 , $R_{\bar{I}^{PH}}^2$, I_{PO} and I_{YP} , performed in terms of mock gene selection on various simulated data sets, where we knew which mock genes had a significant impact on survival. In this section, we will apply and compare the performance of our variable selection methods on the oral and ovarian data described in Chapter 3. Because we do not know the number of significant genes in real data, the approach will differ from the simulations

in the previous section where we looked at AUC, ROC and Youden values. Here, we will instead rank the genes based on each method and compare a preselected number of top genes across methods. We applied each method on the 12,776 and 32,575 genes in the oral and ovarian data sets, respectively, where each gene then has a value for each measure.

R^2 Measure Performance To help compare the methods, we first examine the actual values produced by the R^2 methods for a randomly selected 15 genes in each dataset. Table 4.5 shows the R^2 values for the six proposed R^2 measures, R_{PO}^2 , R_{LR}^2 , R_{PLO}^2 , R_{IPO}^2 , R_{ModCH}^2 , and R_{IPH}^2 , as well as existing measures R_{CH}^2 and R_{PH}^2 , for the oral dataset. From this, it is clear that each method performs differently. In most scenarios, R_{PLO}^2 , R_{CH}^2 and R_{ModCH}^2 produce much lower values compared to the other measures. R_{PO}^2 , R_{LR}^2 , and R_{IPO}^2 also appear to be producing much different values; however, we observe that the ranking of genes 1-15 based on those values would be similar. The values for R_{IPO}^2 and R_{IPH}^2 are particularly high compared to the others.

Table 4.6 shows a similar set of results for the ovarian dataset, depicting the R^2 values for 15 genes. Again, there are clear differences between the values being produced by each method. Here, many of the values are lower than what was observed in the oral dataset, but that could be due to the significance of the genes in this particular subset. If we look particularly at Gene 4, we see that R_{IPO}^2 and R_{IPH}^2 are still producing higher values compared to the other methods, with R_{PLO}^2 , R_{CH}^2 and R_{ModCH}^2 being the lowest.

Next, we look at the ranges for the R^2 values across each dataset in

Table 4.5: R^2 Examples in the Oral Dataset

Gene	R_{PO}^2	R_{LR}^2	R_{IPO}^2	R_{PLO}^2	R_{CH}^2	R_{ModCH}^2	R_{IPH}^2	R_{PH}^2
1	0.124	0.042	0.795	0.008	0.016	0.053	1	0.129
2	0.124	0.042	0.795	0.008	0.016	0.053	1	0.129
3	0.103	0.038	0.733	0.01	0.172	0.017	0.94	0.088
4	0.377	0.132	0.981	0.014	0.066	0.101	1	0.381
5	0.094	0.033	0.065	0.003	0.018	0.05	0.093	0.094
6	0.082	0.027	0.589	0.022	0.062	0.103	0.917	0.067
7	0.134	0.045	0.994	0.014	0.069	0.09	1	0.118
8	0.168	0.054	0.764	0.001	0.182	0.186	0.994	0.177
9	0.081	0.036	0.216	0.001	0.06	0.085	0.354	0.071
10	0.081	0.036	0.216	0.001	0.06	0.085	0.354	0.071
11	0.046	0.018	0.054	0.058	0.095	0.137	0.08	0.037
12	0.122	0.047	0.769	0.094	0.054	0.001	0.997	0.146
13	0.12	0.047	0.527	0	0.011	0.052	0.954	0.126
14	0.122	0.03	0.509	0.097	0.003	0.001	0.976	0.165
15	0.232	0.083	0.869	0.005	0.016	0.053	1	0.265

Table 4.6: R^2 Examples in the Ovarian Dataset

Gene	R_{PO}^2	R_{LR}^2	R_{IPO}^2	R_{PLO}^2	R_{CH}^2	R_{ModCH}^2	R_{IPH}^2	R_{PH}^2
1	0.058	0.008	0.048	0.035	0.058	0.099	0.064	0.048
2	0.003	0	0.001	0	0.002	0.002	0	0
3	0	0	0	0.002	0.002	0.001	0	0
4	0.119	0.013	0.371	0.001	0.068	0.105	0.629	0.101
5	0.018	0.002	0.007	0.065	0	0.011	0.017	0.028
6	0.003	0.001	0.013	0.029	0.003	0.028	0.002	0.001
7	0.155	0.016	0.008	0.013	0.065	0.073	0.014	0.139
8	0.164	0.016	0.105	0.056	0.346	0.364	0.134	0.109
9	0.127	0.012	0.022	0.082	0	0.003	0.039	0.117
10	0.024	0.003	0.013	0.009	0.006	0.032	0.02	0.016
11	0.02	0.002	0.008	0.01	0.01	0.001	0.014	0.018
12	0.299	0.033	0.124	0.002	0.087	0.128	0.212	0.212
13	0.021	0.001	0.012	0.039	0.009	0.006	0.043	0.027
14	0.042	0.004	0.005	0.018	0.018	0.005	0.01	0.029
15	0.064	0.006	0.015	0.036	0.008	0.004	0.037	0.056

Table 4.7. While there are some differences across both data sets, we see that the R_{PO}^2 , R_{IPO}^2 , and R_{IPH}^2 values consistently range from approximately 0 to 1. Other methods, particularly R_{LR}^2 , R_{PLO}^2 , R_{CH}^2 , and R_{ModCH}^2 , have a much lower upper limit. For the R_{PO}^2 method, this range is beneficial since it is interpreted as the percentage of separability in gene expression between those experiencing and not experiencing the event of interest. Thus, for each gene, we are seeing somewhere from 0 to almost 100% separation.

Table 4.7: R^2 Ranges in the Oral and Ovarian Datasets

Method	Oral	Ovarian
R_{PO}^2	(0, 0.93)	(0, 1.00)
R_{LR}^2	(0, 0.39)	(0, 0.11)
R_{IPO}^2	(0, 1.00)	(0, 0.94)
R_{PLO}^2	(0, 0.43)	(0, 0.79)
R_{CH}^2	(0, 0.45)	(0, 0.67)
R_{ModCH}^2	(0, 0.38)	(0, 0.69)
R_{IPH}^2	(0, 1.00)	(0, 1.00)
R_{PH}^2	(0, 0.91)	(0, 0.55)

Next, we examine the differences between the genes selected by the R^2 measures, selecting the top 500 genes in each case. Figure 4.1 shows the overlaps between the PO-based R^2 measures, R_{PO}^2 , R_{LR}^2 , and R_{IPO}^2 . In both the oral and ovarian data, we observe a significant amount of genes common to all three measures. In fact, they select approximately 70% of common genes, which is not surprising since all three measures are based on the PO model and performed similarly in the simulations. Next, we look only at R_{PO}^2 , R_{PLO}^2 , and R_{ModCH}^2 in Figure 4.2. There are minor overlaps between the three measures, with the largest overlap falling between R_{PO}^2 and R_{ModCH}^2 . They have approximately 27% of genes in common in both cases. However, there

are no genes common to all three measures in the ovarian dataset. This is not surprising since each of these methods is based on a different model, so we would expect them to select some different genes. Although not shown here, we note that R_{CH}^2 and R_{ModCH}^2 selected approximately 80% of the same genes.

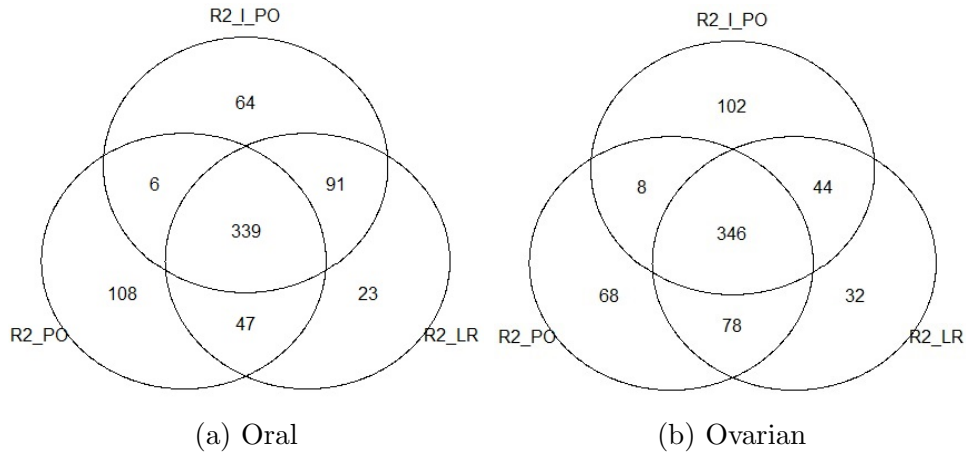


Figure 4.1: Top 500 Selected Genes, R^2 PO-based Measures

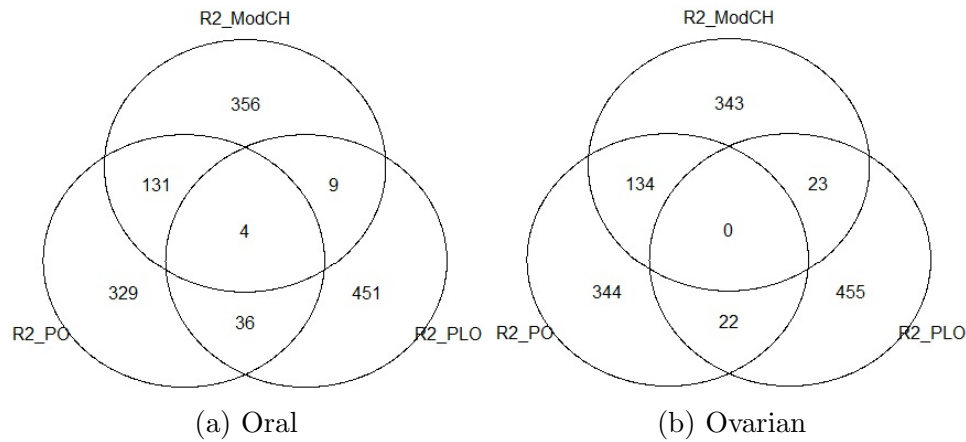


Figure 4.2: Top 500 Selected Genes, Other R^2 Measures

The differences between the measures is most likely related to the presence of NPH genes. We speculated that R_{CH}^2 and R_{PLO}^2 's occasional poor

performance in the simulations was due to the fact that they force NPH and do not allow for PH. On the other hand, R_{PO}^2 and the other PO-based measures allow for both. In fact, 69% of the genes selected in the ovarian data by R_{CH}^2 do not fit the PH model, while this number is only 20% for R_{PO}^2 . Recall, R_{CH}^2 's purpose is to specifically determine genes with crossing hazards. Thus, it is not surprising that such a large number of genes selected by R_{CH}^2 do not fit the PH model. R_{PLO}^2 also selects a large proportion of genes, 79%, for which PH does not fit in the ovarian data. If the goal is to only determine NPH, this is a positive characteristic. However, if the ultimate goal is to select genes having a significant impact on survival regardless of the type of hazard, PH or NPH, then this could have a negative effect. R_{CH}^2 essentially ignores potentially significant genes exhibiting PH. On the other hand, only 20% of the genes selected by R_{PO}^2 do not fit the PH model, which is showing less favoritism to NPH, but is still able to identify genes with the NPH characteristic. Also, in the ovarian data, 95% and 63% of the genes selected fit the PO model for R_{PO}^2 and R_{CH}^2 , respectively. Thus, R_{PO}^2 has an advantage in selecting genes where PO fits, which is useful because we saw in Table 3.1 that PO fits a large proportion of genes for which the PH model does not fit. Thus, R_{PO}^2 is able to correctly identify genes regardless of the type of hazard, which may be an advantage. It does not favor nor neglect any gene based on hazard type, which is a positive attribute. In the end, each measure essentially selects a different subset of genes, so the appropriate measure can be chosen depending on the variable selection goal.

Although each R^2 measure appears to select a different subset of genes,

we recommend using R_{PO}^2 or another PO-based R^2 measure because of their inherent versatility. These PO-based measures demonstrated their ability to handle PH and various forms of NPH throughout the simulation results, while R_{ModCH}^2 only performed well in detecting crossing hazards. In §4.2.1, we also showed that R_{PO}^2 , as well as R_{PLO}^2 , provide an added benefit of being easily interpreted as the percentage of separation in gene expression between those experiencing and not experiencing the event of interest. However, R_{PO}^2 outperformed R_{PLO}^2 in most scenarios. Thus, using the genes selected with R_{PO}^2 or another PO-based R^2 measure would produce the most accurate results.

Comparing I_{PO} , I_{YP} , and $Concreg$ We will now perform a similar analysis on the remaining three measures, I_{PO} , I_{YP} , and c'_+ . Recall, we can calculate a p -value for the test based on I_{PO} and I_{YP} , and then use a standard $p < \alpha$ approach to select significant genes. However, for the sake of a direct comparison with c'_+ , we will begin by selecting the top 500 genes based on our ranking approach. The overlaps between the genes selected using these measures can be seen in Figure 4.3. We note a similar divide between the genes selected using the I measures and c'_+ , just as we saw with the R^2 measures. Note, the I measures were applied using the χ^2 test statistics described in §4.1. The three measures found 26 and 1 common genes for the oral and ovarian sets, respectively, and there is also a small overlap between I_{PO} and c'_+ . I_{YP} and c'_+ also have a very small overlap in the ovarian dataset, but they share over 100 common genes in the oral dataset.

Again, these differences are most likely attributed to the types of haz-

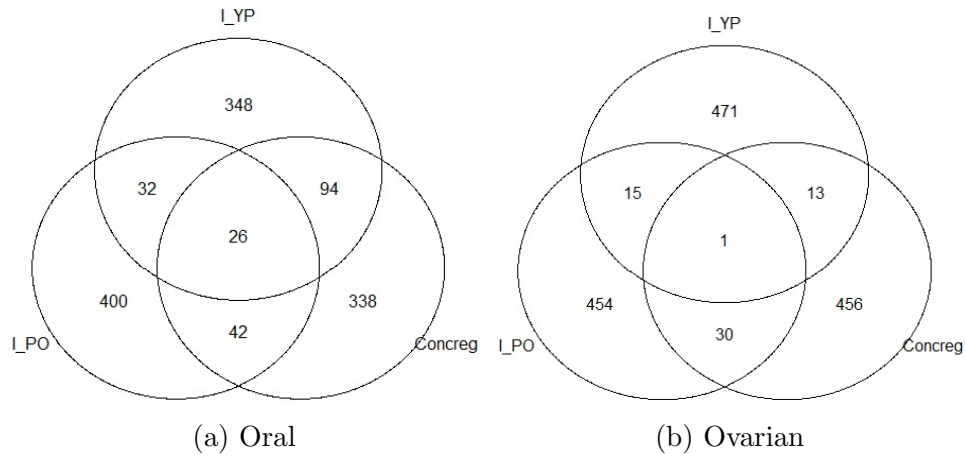


Figure 4.3: Top 500 Selected Genes, I_1 , I_2 and Concreg Measures

ards present. To understand the differences between gene selection, we look at the PH model GOF for the selected genes for each method. In the ovarian data, 11% of the genes selected by I_{PO} violate the PH assumption, and I_{YP} and Concreg have a slightly larger proportion of 17% and 18%, respectively. Thus, the PO measure is able to handle both PH and NPH without favoring either type of hazard. Also, I_{PO} picks up a slightly larger proportion of genes for which PO fits, which is not surprising since it is based on that model. In general, each method has proven to be useful and the appropriate method can be chosen based on the variable selection problem.

We recommend the use of I_{YP} because of its inherent versatility. It is able to handle various types of hazards and retains both I_{PO} and I_{PH} as special cases. However, the performance of this measure could be improved by developing methods to estimate its parameters using continuous covariates. Thus, for now, I_{PO} appears to be the most useful. This measure performed similarly or better than \hat{c}'_+ in every simulated scenario. It is also easy to

calculate and its statistical test produces a p -value that can be used for simple variable selection at a pre-defined level of significance.

***I*-type Measures: A Visual Representation** We now focus our attention on the application of the *I*-type measures, I_{PO} and I_{YP} . First, using the χ^2 test described in §4.1, we calculate p -values and select genes by controlling the FDR at the $\alpha = .05$ level using the Benjamini-Hochberg approach (Benjamini and Hochberg 1995). I_{PO} selected 802 and 5,469 genes, and I_{YP} selected 8,559 and 10,636 genes, for the oral and ovarian data, respectively.

First, we create a weighted average of gene expression based on the selected subset of genes using the technique described in §3.2. Here, the χ^2 values based on the I_{PO} and I_{YP} themselves are used as the weights. Goodness of fit tests are then applied for the PH, PO and YP models, and the weighted averages are dichotomized by the median so survival curves can be plotted. As a comparison, curves based on I_{PH} are also created. The curves and GOF results can be seen in Figure 4.4. Looking at the oral curves, we observe physical evidence of diverging and converging survival curves, but in each case, all three models fit (GOF $p > .05$). Recall, the PO model allows for PH and both the PH and PO models are special cases of the YP model. Thus, the YP and PO models are demonstrating their ability to handle PH. On top of GOF, model significance was also checked, and in each case, the model of interest is significant ($p < .05$) for its weighted average.

Next, we look at the ovarian curves, where we recognize evidence of crossing and converging survival curves for I_{PO} and I_{YP} . Not surprisingly, I_{PH}

shows evidence of proportional survival curves. However, the GOF results will give us a more direct interpretation of the hazards. We observe in plot (d) and (f) that the PH, PO and YP models fit (GOF $p > .05$). On the other hand, for plot (e), we see that neither the PH nor PO model fit the weighted average. However, the YP model fits in this case, which further demonstrates its versatility. On top of GOF, model significance was also checked, and in each case, the model of interest is significant ($p < .05$) for its weighted average.

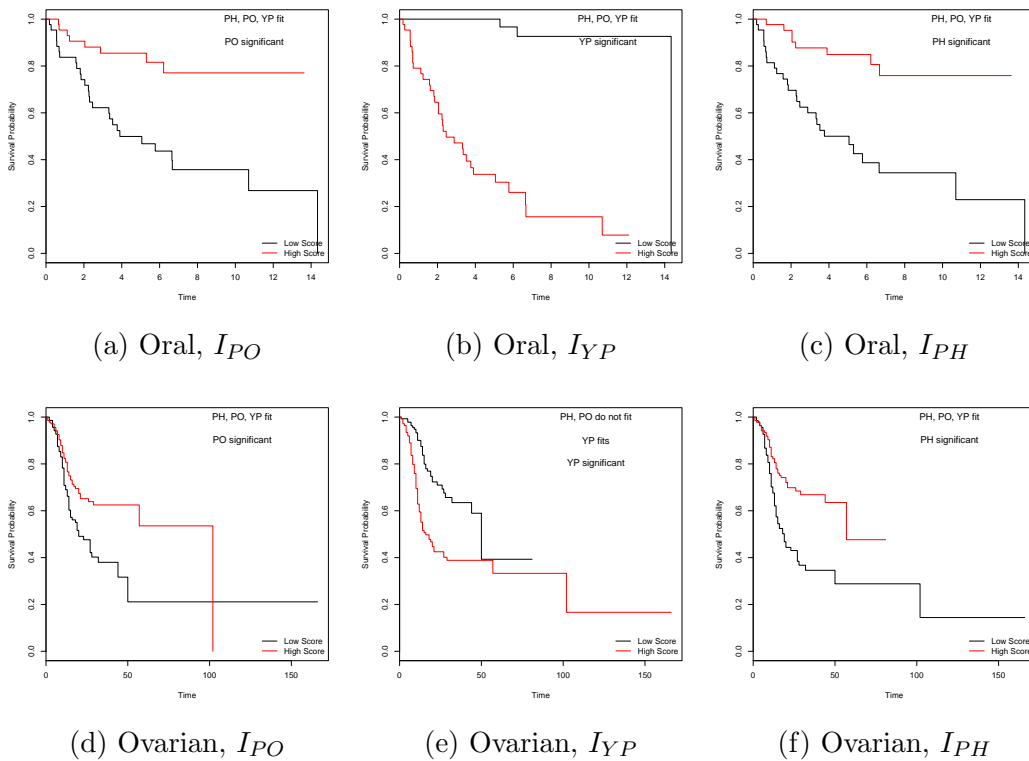


Figure 4.4: Survival Curves for Dichotomized Weighted Average using I_{PO} , I_{YP} and I_{PH}

Next, we examine the genes selected using the BH FDR approach in more detail. Table 4.8 shows the number of selected genes at the $\alpha = 0.05$

level, as well as the various proportions of genes selected that do not fit the PH model, fit the PO model, and fit the YP model. For the oral data, the FDR approach selected 802 and 8,559 genes using the I_{PO} and I_{YP} measures, respectively, of which 3% and 20% do not fit the PH model. This number is particularly high for the I_{YP} set. Also note, a large proportion of the genes selected by both measures fit the PO and YP model. In fact, for the genes selected by the I_{PO} measure, 97% fit the PO model, and for the genes selected by the I_{YP} measure, 90% fit the YP model. For the ovarian data, which selects 5,469 and 10,636 genes using the I_{PO} and I_{YP} measures, respectively, we also discover a high presence of NPH genes. 13% and 16% do not fit the PH model, and a large proportion of the genes selected in both measures fit the PO and YP model. Thus, these measures are demonstrating their ability to select genes with varying types of hazards.

Table 4.8: GOF for Genes Selected using I Measures

	Method $p < .05$	# Genes Selected	% Cox PH not fit	% PO not fit	% Cox PH & PO not fit	% PO fit	% YP fit
Oral	I_{PO}	802	3%	3%	1%	97%	87%
	I_{YP}	8,559	20%	12%	9%	88%	90%
Ovarian	I_{PO}	5,469	13%	7%	4%	93%	89%
	I_{YP}	10,636	16%	12%	7%	88%	95%

PO Measures and Individual Gene Analysis We now compare the genes selected by two variable selection methods based on the PO model, R_{PO}^2 and I_{PO} . Figure 4.5 shows the overlap between the 500 selected genes for these two methods. They found 181 and 209 common genes for the oral and ovarian data, respectively, which is approximately 36% and 42% of the total selected.

Based on the simulations and examples in this section, R_{PO}^2 and I_{PO} seem to be performing the best out of the new measures. Thus, we now will focus our attention on the 181 and 209 common genes selected by both measures. First, we will pinpoint individual genes in that group showing some form of NPH and then examine whether there is an interaction effect between multiple genes.

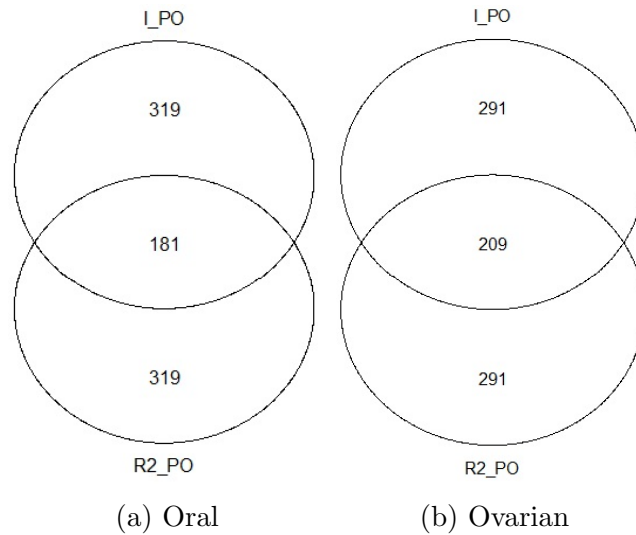


Figure 4.5: Top 500 Selected Genes, PO Measures

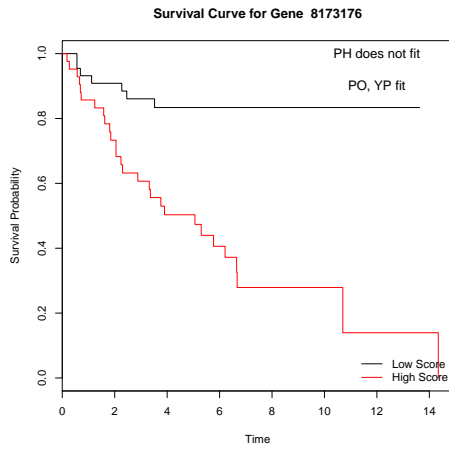
As a demonstration, we select 2 genes from the group of 181 and 209 and dichotomize their gene expression by the median. Doing so will allow us to plot the survival curves for each gene separately and examine if NPH is present. Figure 4.6 shows the survival curves of two of those genes for each set. In the oral examples, we observe evidence of diverging hazards in plot (a) and converging hazards in plot (b). We also included the PH, PO and YP model goodness-of-fit (GOF) results for each gene. In both cases, the PH GOF p -value is less than $\alpha = .05$, reaffirming that the PH model does not fit. Another interesting finding is that the PO GOF p -value for both genes is large,

indicating that the PO model fits for both genes. Thus, we see that when the PO measures are used to reduce the data set, they are able to select genes that fit the PO model but also exhibit NPH. In both cases, the YP model fits as well.

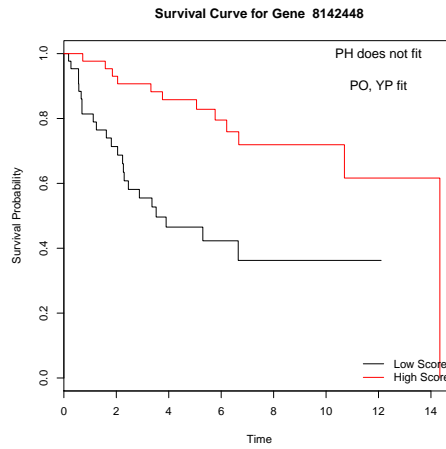
In the ovarian examples, we observe evidence of crossing hazards for both plots (c) and (d). In both cases, those with a “High” gene expression have lower survival until just after 50 months, where the hazards cross and those with a “Low” gene expression have a lower survival probability. In the second gene, we also discover a second crossing at 100 months. Similar to the oral examples, the PH GOF p -values are less than $\alpha = .05$, reaffirming that the PH model does not fit, but the PO and YP models have large p -values, indicating that they fit the gene well.

Gene Interaction We will now examine the interaction between the two genes in each data set seen above to check if some sort of modulating effect is present. We looked at the dichotomized values for the two genes and broke the subjects into four sub-groups: Genes 1 & 2, High; Gene 1, High & Gene 2, low; Gene 1, Low & Gene 2, High; and Genes 1 & 2, Low. The survival curves for these four sub-groups are plotted on one axes to compare in Figure 4.7.

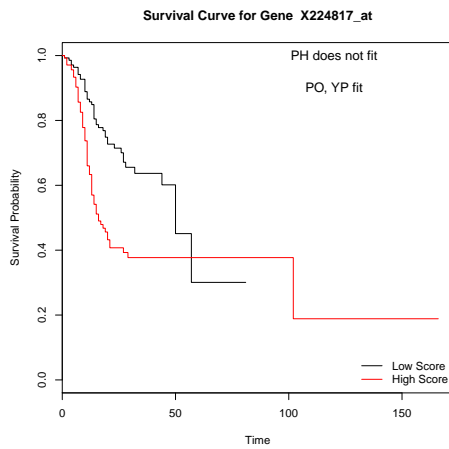
Based on the oral survival plot, it appears as though when gene 1 is lowly expressed and gene 2 is highly expressed, patients have the best prognosis. Alternatively, when the expression levels are switched, patients appear to have the worst prognosis. When both genes are highly expressed, we observe increased survival over the group of patients for which both genes are



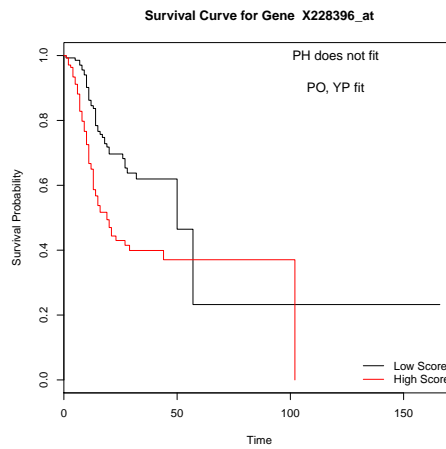
(a) Oral



(b) Oral



(c) Ovarian



(d) Ovarian

Figure 4.6: Individual Genes Selected by PO Measures

lowly expressed, until approximately the 6 year mark. At that point, there is a crossing effect and the lowly expressed group has a better prognosis than the highly expressed group. In the ovarian plot, it appears as though when both genes are highly expressed, the survival is the lowest. Until 50 months, when both genes are lowly expressed, patients have the best prognosis. When gene 1 is highly expressed and gene 2 is lowly expressed and vice-versa, we

observe many crossing effects over time. In general, we could hypothesize that the crossing effects seen in the univariate plots in Figure 4.6 are due in part to some sort of interaction between both genes.

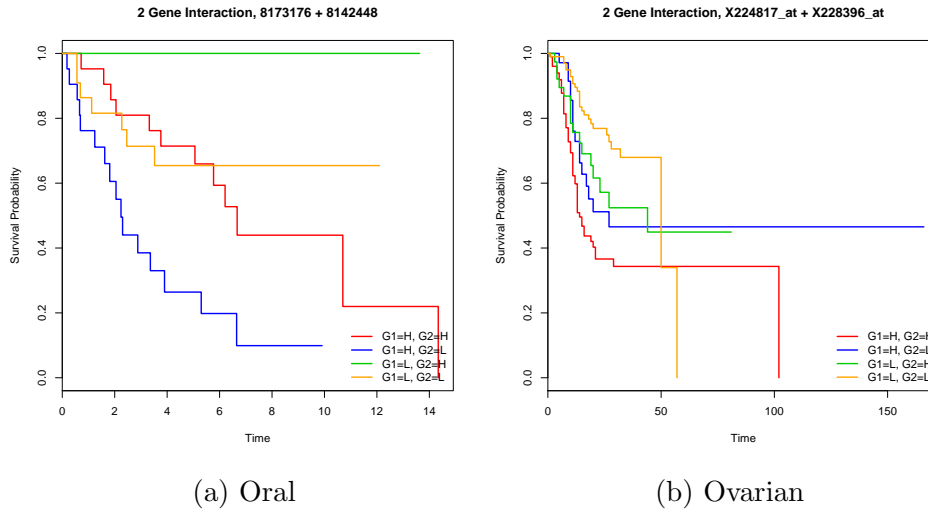


Figure 4.7: Two Gene Interaction

In summary, through simulations and data example we were able to demonstrate the usefulness of our proposed measures, and depending on the types of hazards present, some measures may perform better than others. For example, R_{CH}^2 is specifically designed to identify crossing hazards, but a measure based on PO or YP will provide more flexibility as it allows for varying types of hazards. Thus, the appropriate measure can be chosen based on the variable selection goal and type of hazards present. However, we stress that our proposed measures provide a more versatile and general framework that allow for various types of hazards, both proportional and non-proportional.

CHAPTER 5

PROPOSED METHODS FOR SUPERVISED DIMENSION REDUCTION

In Chapter 4, we developed variable selection methods based on the PO, PLO and YP models that addressed the issue of NPH using marginal screening approaches. In this chapter, we will develop supervised dimension reduction methods for effectively handling high-dimensional genomic data with censored survival outcomes. In §2.4, we discussed numerous dimension reduction techniques based on principal components regression (PCR) and partial least squares (PLS), both of which are special cases of continuum power regression (CPR) (de Jong *et al.* 2001; Sundberg 2002; Spitzner 2004). Li and Gui (2004) proposed a method combining the ideas of PLS and the Cox PH model. While their method addresses the issue of high-dimensionality, it neglects the

potential failure of the PH assumption. It also offers no approach to account for a large proportion of censored observations. In this section, we propose a method that will perform supervised dimension reduction within a framework that allows for NPH and directly accounts for censored observations.

In §5.1, we propose two methods that combine CPR for dimensionality reduction with the accelerated failure time (AFT) model defined in §2.2.3. The first method will be referred to as CPRAFT, and the second method, which directly adjusts the censored observations, will be referred to as ACPRAFT. We will show that our proposed method also has a special case using PLS, named (A)PLSAFT, described in Devarajan *et al.* (2010). In their approach, they consider the log-normal and Weibull models. In §5.2, we extend the (A)CPRAFT method to the generalized F model defined in Equation ?? and show that it has many special cases, including gamma, log-normal, Weibull, log-logistic and Burr. In §5.3, we further extend the (A)CPRAFT method to the semiparametric AFT model, a more general model where the error is unspecified. In §5.4, we discuss the use of (A)CPRAFT for gene ranking and selection, and we show examples using both simulations and real genomic data. In §5.5, we develop an algorithm using (A)CPRAFT for survival prediction and evaluate its performance using large-scale genomic data. Using the methods outlined in §5.4 and 5.5, we demonstrate that (A)CPRAFT has the ability to reduce the number of covariates used in survival prediction, while simultaneously ranking genes.

5.1 (A)CPRAFT

(A)CPRAFT are supervised dimension reduction methods for censored survival data that can be used for survival prediction and variable selection. The ACPRAFT method specifically has a clear advantage over other methods in the literature because it directly addresses the three main issues with the application of survival analysis to genomic data. First, it addresses the issue of high-dimensionality using CPR, reducing the number of genes into a smaller number of CPR components that are linear combinations of genes. Second, it addresses the issue of NPH by using the AFT model, a model that does not assume PH but partly overlaps with the PH model. Lastly, it addresses the issue of censoring by imputing the censored observations using the extracted CPR components and the fitted AFT model. In §2.4, we discussed other methods for dimension reduction, and while some used either PLS or the AFT model as part of their procedure, none of the methods addressed the issue of censoring directly. In Table 3.1, the two data sets examined had a proportion of censored observations of approximately 59%. Thus, in this application, having a method like ACPRAFT that adjusts for censored data is highly beneficial. We also explore the use of the CPR coefficients in developing a survival prediction model.

In the rest of this section, we will discuss the (A)CPRAFT procedure, as well as its special case using PLS from Devarajan and Ebrahimi (2010).

CPR The first step in the (A)CPRAFT algorithm is to apply CPR, an approach outlined in 2.4.2. The goal of CPR is to find weight vectors, \mathbf{w} , such

that the linear combination of gene expression, $\mathbf{Z}\mathbf{w}$, maximize the objective function $R^2(\mathbf{v}, \mathbf{y})\text{Var}(\mathbf{v})^\gamma$, where $\mathbf{v} = \mathbf{Z}^{(\gamma)}\mathbf{w}$ and $\gamma \equiv \alpha/(1 - \alpha)$, subject to the constraints outlined in §2.4.2. $\mathbf{Z}^{(\gamma)}$ is found via the spectral value decomposition of \mathbf{Z} as explained in de Jong *et al.* (2001). After this step, standard PLS can be applied to $\mathbf{Z}^{(\gamma)}$. The parameter α can be chosen, providing a more general form with OLS ($\alpha = 0$), PLS ($\alpha = 1/2$), and PCR ($\alpha = 1$) as special cases. Let $u_{ik} = z_i w_k, i = 1, \dots, N, k = 1, \dots, K$, denote these linear combinations of gene expressions selected by CPR. These represent the components, and the number of components $K < p$ is chosen based on leave-one-out cross validation (LOOCV) to minimize the predicted residual sum of squares (PRESS) statistic. In the next section, we explain how (A)CPRAFT is used and selects both K , the optimal number of components, and α , the CPR parameter.

PLS PLS is a special case of CPR, so it maximizes the objective function seen above with $\gamma = 1$ or $\alpha = 1/2$, where $\gamma \equiv \alpha/(1 - \alpha)$. In other words, PLS essentially maximizes the function $R^2(\mathbf{v}, \mathbf{y})\text{Var}(\mathbf{v})^\gamma$, where $\mathbf{v} = \mathbf{Z}^\gamma\mathbf{w}$, when $\gamma = 1$. Therefore, it maximizes the covariance between $\mathbf{Z}\mathbf{w}$ and $\log(t)$. Thus, the first step in the (A)PLSAFT algorithm is to apply PLS to the data to create a weight vector, w_k , such that the linear combinations of the gene expressions $\mathbf{Z}w_k$ maximize the objective function $w_k = \arg_{w w' = 1} = \max \text{Cov}^2(\mathbf{Z}w, \log(t))$ subject to zw_k being orthogonal to the already found linear combinations $zw_j, 1 \leq j < k$. Let $u_{ik} = z_i w_k, i = 1, \dots, N, k = 1, \dots, K$, denote the linear combinations of gene expressions selected by PLS. These represent the PLS components. The number of PLS components $K < p$ is chosen based on leave-

one-out cross validation (LOOCV) to minimize the predicted residual sum of squares (PRESS) statistic.

AFT Model Fitting The next step in the (A)CPRAFT methods involve fitting the AFT model described in Equation 2.2.14. Here, we use the notation

$$\log Y_i = \beta' u_i + \sigma^* W_i, i = 1, \dots, N, \quad (5.1.1)$$

where Y_i is the survival time for the i -th subject, $u'_i = (u_{i0}, \dots, u_{iK})$ is a $(K + 1)$ vector for the i -th subject, $\beta = (\beta_0, \beta_1, \dots, \beta_K)$ is the $(K + 1)$ vector of unknown regression parameters, W_i are independent error terms with a common distribution F_W and σ^* is the scale parameter. In this setting, u_{ik} represent the CPR components for the i -th subject.

From Equation 5.1.1, given a gene expression vector \mathbf{z} and PLS component \mathbf{u} , the survival function of Y is

$$\begin{aligned} S(y|\mathbf{z}) &= P(Y > y|\mathbf{z}) \approx P(\log Y > \log y|\mathbf{u}) \\ &= P(\beta' \mathbf{u} + \sigma^* W > \log y|\mathbf{u}) \\ &= P(W > (\log y - \beta' \mathbf{u})/\sigma^*) \\ &= 1 - F_W \left(\frac{\log y - \beta' \mathbf{u}}{\sigma^*} \right). \end{aligned}$$

$S(y|\mathbf{z})$ is estimated by replacing the unknown parameters with their maximum likelihood estimates.

The AFT model has both parametric and semi-parametric cases. In the semi-parametric case, the distribution of the error W is unspecified. An

(A)CPRAFT extension using this semi-parametric case will be discussed in §5.3. In the parametric cases, the model is specified by the distribution of the error term. A flexible parametric AFT model is described below.

Parametric Generalized F (GF) Model From Equation 5.1.2, we know that if $Y \sim$ Generalized F, then the distribution of Y is

$$f_Y(y|\mu, \sigma, n, m) = \frac{1}{\sigma\beta(n, m)} e^{-\frac{\mu n}{\sigma}} y^{\frac{n}{\sigma}-1} \left(\frac{n}{m}\right)^n \left[1 + \left(\frac{n}{m}\right) (e^{-\mu} y)^{\frac{1}{\sigma}}\right]^{-(n+m)}, y > 0. \quad (5.1.2)$$

As shown in Ciampi *et al.* (1986) and Cox (2008), this model serves as an umbrella distribution with many special cases, where choosing specific parameter values will produce estimates for a particular model of interest. A few special cases are described in Table 5.1.

Table 5.1: GF Model: Special Cases

Model	Parameters
Log-Normal	$m \rightarrow \infty, n \rightarrow \infty$
Weibull	$m \rightarrow \infty, n = 1$
Log Logistic	$m = 1, n = 1$

Many other models fall under the umbrella, including gamma, exponential, Burr and so forth. Thus, extending (A)CPRAFT to this generalized distribution essentially extends the method to all of these distributions, making the method broadly applicable. This extension will be discussed in more detail in §5.2.

5.1.1 CPRAFT Algorithm

The CPRAFT algorithm has the following steps:

1. Choose α parameter.
2. Perform CPR using desired α to obtain weight vectors w_k , using the PRESS statistic based on LOOCV to choose K , the number of components.
3. Use time to event data (t_i, δ_i) and covariate $u_i = (u_{i1}, \dots, u_{ik})$ (components), to model the effect of the covariate on Y using Equation 5.1.1.
4. Fit the AFT model (Equation 5.1.1) via maximization of the likelihood function based on the specific distribution of choice.

5.1.2 ACPRAFT Algorithm

The ACPRAFT algorithm has the following steps:

1. Use uncensored data to obtain components where the number of components is chosen as specified above. Use these components as covariates for the model in Equation 5.1.1 and obtain estimates for β and σ^* .
2. Let $v_i = \log t_i$. Replace v_i by $v_i^* = \delta_i v_i + (1 - \delta_i) \hat{E}(\log Y_i | \log Y_i > v_i)$, $i = 1, \dots, N$. Under (5.1.1), $\hat{E}(\log Y_i | \log Y_i > v_i) = \hat{\beta} u_i + \hat{\sigma}^* E\left(W | W > \frac{v_i - \hat{\beta} u_i}{\hat{\sigma}^*}\right)$, where $\hat{\beta}$ and $\hat{\sigma}^*$ are obtained in step 1 (Buckley and James 1979; Devarajan *et al.* 2010) and W is the error term in Equation 5.1.1. The calculation of $E\left(W | W > \frac{v_i - \hat{\beta} u_i}{\hat{\sigma}^*}\right)$ can be computed for various models.

3. Use v_1^*, \dots, v_N^* from step 2 to construct new CPR components. The number of components K for this adjusted survival data is determined using PRESS based on LOOCV as described earlier.
 - (a) Use these components as covariates in Equation 5.1.1 and follow steps 3 and 4 of CPRAFT above.
 - (b) Alternatively, use the CPR coefficients to build the final model.

The algorithms above fully demonstrate the ability of CPRAFT and ACPRAFT to reduce the data into K components using CPR. We will show in §5.3 that it has the advantage of simultaneously being able to rank the variables of interest, in this case genes. Also, ACPRAFT has a significant advantage over CPRAFT because it adjusts for the censored observations. This is highly beneficial because genomic data often have a large proportion of censored observations.

In step 1 of the algorithm, choosing parameter $\alpha = .5$ results in the special case (A)PLSAFT described in Devarajan *et al.* (2010). In §5.4, we examine the performance of this special case of (A)CPRAFT using $\alpha = .5$. In future work, we could compare this to (A)CPRAFT at varying α levels. Then in §5.5, we apply (A)CPRAFT, choosing the optimal α using PRESS based on LOOCV, to develop a survival prediction model using the CPR coefficients. We also discuss the use and limitations of the CPR components for survival prediction.

5.2 A Flexible Parametric Approach to (A)CPRAFT

While (A)CPRAFT has proved to be useful, it requires the fitting of specific AFT models, and Devarajan *et al.* (2010) only considered the AFT log-normal and Weibull cases in their (A)PLSAFT approach. We propose a generalization of (A)CPRAFT using the GF model, which encompasses many special cases and is, therefore, a flexible alternative for modeling censored survival data in conjunction with large-scale genomic data.

5.2.1 GF-based (A)CPRAFT Algorithm

To extend (A)CPRAFT to the GF model, we first obtain an expression for $E\left(W|W > \frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}\right)$ for this model, seen in §5.1.2, where $W = \frac{\log Y - \mu}{\sigma}$. Using the distribution of Y defined in Equation 5.1.2, we find the distribution of W .

Let $W = g(Y) = \frac{\log Y - \mu}{\sigma}$. Then, $g^{-1}(W) = e^{\sigma W + \mu}$ and $\frac{d}{dW}g^{-1}(W) = \sigma e^{\sigma W + \mu}$.

Thus,

$$\begin{aligned} f_W(w) &= f_Y(g^{-1}(w)) \left| \frac{d}{dw}g^{-1}(w) \right| \\ &= \frac{1}{\beta(n, m)} \left(\frac{n}{m}\right)^n e^{nw} \left[1 + \left(\frac{n}{m}\right)e^w\right]^{-(n+m)}, \quad -\infty < w < \infty. \end{aligned}$$

Therefore, if $Y \sim \text{GF}$, then $W \sim f_W(w)$ as seen above.

Now, we compute $E(W|W > z) = \frac{\int_w^\infty w f_W(w) dw}{1-F(z)}$, which is given by

$$E(W|W > z) = \frac{\int_0^1 \ln \left[\frac{m}{n} \left(\frac{v}{1-v} \right) \right] f_V(v) dv}{\frac{\left(\frac{n}{m}\right)e^z}{1+\left(\frac{n}{m}\right)e^z} - \int_0^{\frac{\left(\frac{n}{m}\right)e^z}{1+\frac{n}{m}e^z}} f_V(v) dv}, \quad (5.2.1)$$

where $f_V(v) \sim \text{Beta}(n, m) = \frac{1}{\beta(n, m)} v^{n-1} (1-v)^{m-1}$ and z represents $\frac{v_i - \hat{\beta} u_i}{\hat{\sigma}^*}$.

To apply this extension in the CPRAFT algorithm described in §5.1.1, the GF model would be used for model fitting in step 4. For the ACPRAFT algorithm in §5.1.2, use the expression in Equation 5.2.1 in step 2 and fit a GF model to estimate the parameters in steps 1 and 3. The GF model is fitted using the R package `flexsurvreg` described in Cox (2008).

Special Cases While the GF model is the most flexible, we will be interested in comparing its performance to its special cases, particularly log-normal, log logistic, and Weibull. To obtain the expected values for these special cases, replace the parameters in Equation 5.2.1 with the parameters specified in Table 5.1. The expressions for log-normal and Weibull can be seen in Devarajan *et al.* (2010), and the new log-logistic extension has the form

$$E(W|W > z) = \frac{\int_w^\infty w \frac{e^{-w}}{(1+e^{-w})^2} dw}{\frac{e^{-z}}{1+e^{-z}}}, \quad (5.2.2)$$

where $z = \frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}$.

In this thesis, we will examine (A)CPRAFT using the parametric GF model, as well as its separate cases log-normal, Weibull and log-logistic. In the next section, we discuss using the semiparametric AFT model, where the error term is unspecified.

5.3 A Semi-Parametric Approach to (A)CPRAFT

Although the generalization based on the GF model is parametric in nature, it offers tremendous modeling flexibility unlike its special cases. Here, we further extend our (A)CPRAFT approach using the semiparametric AFT model. This approach has no distribution assumption for the error term. This is an attractive property because it will not force the choice of a specific model, thus providing optimum flexibility in the application of the method.

We begin by noting that the semiparametric AFT model has the form

$$\log Y_i = \beta' u_i + W_i, i = 1, \dots, N, \quad (5.3.1)$$

where Y_i is the survival time for the i -th subject, $u_i' = (u_{i0}, \dots, u_{iK})$ is a $(K + 1)$ vector for the i -th subject, $\beta = (\beta_0, \beta_1, \dots, \beta_K)$ is the $(K + 1)$ vector of unknown regression parameters, and W_i are independent error terms with unknown distribution F . In this setting, u_{ik} represent the CPR components for the i -th subject.

Referring to the CPRAFT algorithms described in 5.1.1, step 1 remains the same. In steps 2 and 3, use the time to event data and covariates (components) to model the effect of the covariate X . Fitting the semiparametric AFT model is based on the Buckley-James (BJ) type estimator developed by Jin *et al.* (2006).

Referring to the ACPRAFT algorithm described in 5.1.2, step 1 remains the same. In step 2, set $v_i = \log t_i$. Replace v_i by

$$\begin{aligned} v_i^* &= \delta_i v_i + (1 - \delta_i) \hat{E}(\log Y_i | \log Y_i > v_i), i = 1, \dots, N \\ &= \delta_i v_i + (1 - \delta_i) \left[\hat{\beta} u_i + E \left(W | W > v_i - \hat{\beta} u_i \right) \right] \\ &= \delta_i v_i + (1 - \delta_i) \left[\hat{\beta} u_i + \frac{\int_{e_i(\beta)}^{\infty} u d\hat{F}_\beta(u)}{1 - \hat{F}_\beta\{e_i(\beta)\}} \right], \end{aligned}$$

where $e_i(\beta) = v_i - \hat{\beta} u_i$ and $\hat{F}_\beta(t) = 1 - \prod_{i: e_i(\beta) < t} \left(1 - \frac{\delta_i}{\sum_{j=1}^n I\{e_j(\beta) \geq e_i(\beta)\}} \right)$ is the Kaplan Meier estimator of F based on $\{e_i(\beta), \delta_i\}$.

In step 3, use v_1^*, \dots, v_N^* from step 2 to construct new components, as described in 5.1.2. Then, use these components as covariates in the semiparametric AFT model and fit as described above using the BJ type method in Jin *et al.* (2006).

5.4 (A)CPRAFT as a Ranking Method

In this section, we discuss how (A)CPRAFT can be used for variable selection, which in our application is gene selection. We compare the performance of (A)CPRAFT algorithms based on log-normal (LN), log-logistic (LL), Weibull

(W), generalized F (GF), and semiparametric AFT (sAFT) models. To implement the CPR portion of the (A)CPRAFT algorithms, we use the generalized CPR algorithm defined in de Jong *et al.* (2001). In this application, we use $\alpha = 1/2$, which corresponds to the (A)PLSAFT special case based on PLS. In future work, other α values could be applied.

Some methods for variable selection based on PLS and CPR have been cited in the literature. For example, Mehmood *et al.* (2012) discuss two possible measures:

1. Regression coefficients (β)
2. Variable importance projection (VIP)

When using the regression coefficients, genes are ranked based on the absolute value of their coefficients, and genes with a higher value are considered to have a larger effect on survival. Because the values of the coefficients typically tend to be very small, we found VIP to be more useful. VIP was first introduced by Wold *et al.* (1993).

The concept behind VIP is to accumulate the importance of each gene being reflected by the weight \mathbf{w} from each component. Essentially, it is a measure of the contribution of each gene according to the variance explained by each component. The VIP value, ν , for gene j is calculated as

$$\nu_j = \sqrt{p \sum_{k=1}^K [SS_k (w_{kj}/\|\mathbf{w}_k\|)^2] / \sum_{k=1}^K SS_k}, \quad (5.4.1)$$

where p is the number of genes, K is the number of components, \mathbf{w}_k is the

weight vector for the k -th component, and SS_k is the sum of squares explained by the k -th component. In other words, CPR produces K vectors of weights, each of which has p elements corresponding to the p genes. In the VIP calculation, $(w_{kj}/\|\mathbf{w}_k\|)^2$ represents the importance of the j -th component.

Thus, ν_j can be calculated for all $j = 1 \dots p$, and then the genes are selected when they are larger than some threshold set by the user. A popular threshold is 1 and is discussed in Mehmood *et al.* (2012). For our application, we start by ranking the genes based on their VIP value, highest to lowest. Once genes are ranked, specificity, sensitivity and AUCs can be computed so that the various (A)CPRAFT methods (log-normal, log-logistic, Weibull, generalized-F, semiparametric AFT) can be compared. In a similar fashion, specificity, sensitivity, and AUCs can be computed using the ranked absolute values of the CPR coefficients (β). We also compare ACPRAFT to CPRAFT, where no adjustment is made based on censoring. We focus on cases with a higher proportion of censoring ($> 0\%$) because that is where we expect to see a difference between ACPRAFT and CPRAFT. In this analyses, we look only at the $\alpha = .5$ case, but in future work, other α values could be chosen and evaluated.

5.4.1 Simulation Results

Simulations were performed using the approach described in §4.3.1. For each scheme and censoring combination, 200 data sets were generated and assessed. (A)CPRAFT using $\alpha = .5$ (namely, (A)PLSAFT) was run on each simulated data set and the absolute value of the PLS coefficients and VIP values (un-

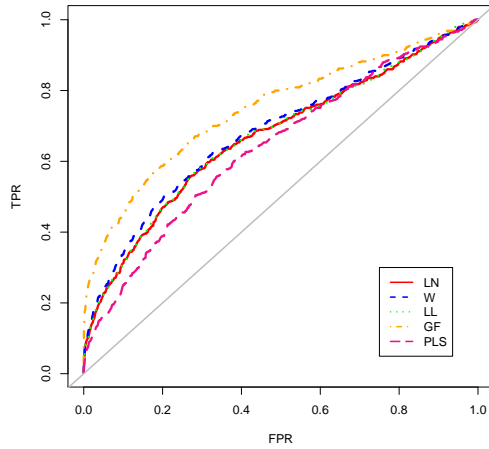
adjusted for CPRAFT and adjusted based on each model for ACPRAFT) were calculated and ranked. For CPRAFT, the PLS coefficients and VIP are calculated in step 1 of the algorithm in §5.1.1, and therefore, are not model specific. Alternatively, ACPRAFT has the advantage of calculating these measures in step 3, taking each specific model into account. Average AUCs and confidence intervals, and specificity and sensitivity measures, were calculated across the 200 simulations in each case and used to compare the methods below. The following abbreviations are used to denote the various models for ACPRAFT and unadjusted CPRAFT: LN, W, LL, GF, sAFT, PLS_{unadj} (unadjusted CPRAFT).

Since ACPRAFT adjusts for censored observations, we expect it to perform well, and better than CPRAFT, in settings where the fraction of censored observations is high. In §4.3.1, we defined our simulation schemes that had censoring proportions of 0, 33, and 67%. Here, to fully demonstrate ACPRAFT's capabilities, we include simulated data at the 80% censoring level as well. High censoring appears to be a common theme among genomic data with censored survival outcomes (Bhattacharjee *et al.* (2001); Beer *et al.* 2002; Tothill *et al.* 2008; Saintigny *et al.* 2011; TCGA Research Network), so it will be interesting to see how ACPRAFT and CPRAFT perform in this setting. The purpose of this analysis is two-fold. First, we want to compare the two ranking methods - PLS coefficients and VIP - and second, we want to compare the performance across CPRAFT and ACPRAFT. Comparisons will be done using AUCs, sensitivities, specificities, and Youden indices defined in §4.3.1.

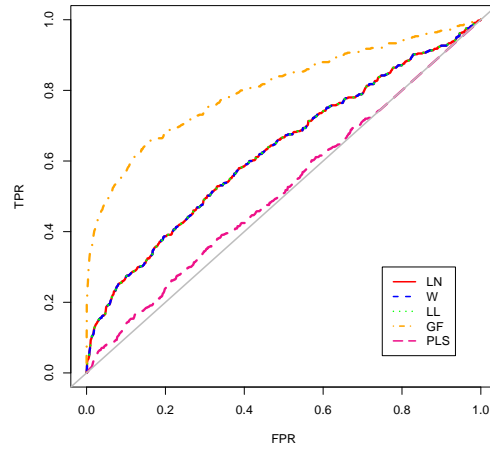
Initial Analyses: GF vs. Special Cases

In initial analyses, we compared the performance of GF with LN, W, and LL. Recall that, the three latter models are special cases of GF. For these analyses, we examined simulated data sets generated using Schemes 1 and 2 in §4.3.1 based on the approach outlined in Dunkler *et al.* (2010), and used ROC curves to compare the performance between methods. The main purpose of this analysis is to compare the performance of GF to its special cases. A similar performance was observed for multiple simulated datasets. Hence, results are reported only for a single, representative simulated dataset. Figure 5.1 depicts the ROC curves for the 33% and 80% censoring cases in scheme 1 using both the PLS coefficients and VIP. In all four cases, we observe that the GF case outperforms LN, W, and LL. Also, the unadjusted method performs significantly worse than GF, and the VIP approach appears to produce higher ROC curves than the PLS coefficients. These differences for scheme 1 are also made apparent in Table 5.2, which shows both the Youden index and AUCs for the 33 and 80% censoring cases. The AUCs for GF are higher than its special cases in each case, and the differences increase as censoring increases. We also note that the unadjusted method produces the lowest Youden index and AUC in every scenario. From this table, we also observe that the Youden index and AUCs are significantly higher for VIP than the PLS coefficients, indicating that VIP is a better tool for variable selection.

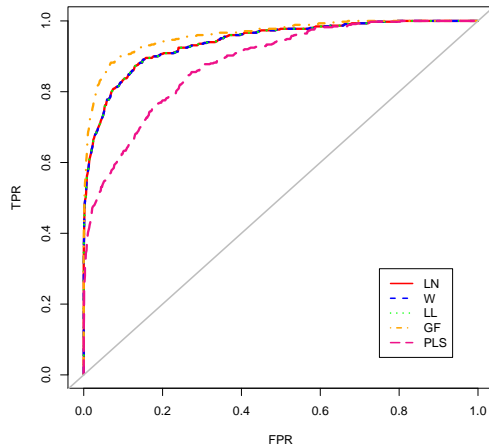
Next, we evaluate the performance using Scheme 2. Figure 5.2 shows the ROC curves for GF comparing VIP and the PLS coefficients. Similar to scheme 1, we observe that VIP significantly outperforms the PLS coefficients.



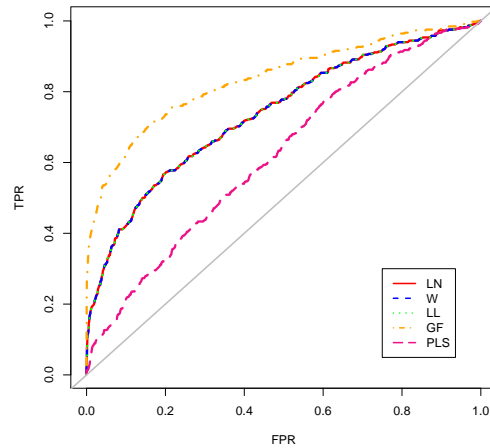
(a) 33% Censoring, PLS Coeff



(b) 80% Censoring, PLS Coeff



(c) 33% Censoring, VIP



(d) 80% Censoring, VIP

Figure 5.1: ROC Curves for Simulation Scheme 1

Figure 5.3 shows the ROC curves for GF and unadjusted PLSAFT based on VIP. Here, it is again evident that GF outperforms the unadjusted case. These differences will be evaluated in more detail in the next section. Although not shown, scheme 2 showed similar results for GF compared to its special cases

Table 5.2: Simulation Scheme 1

		PLS Coeff		VIP	
		Youden	AUC	Youden	AUC
33 % cens.	LN	.19	.67	.65	.94
	W	.21	.68	.65	.94
	LL	.20	.67	.65	.94
	GF	.33	.75	.74	.96
	PLS _{unadj}	.12	.64	.48	.88
80% cens.	LN	.16	.63	.30	.74
	W	.16	.63	.30	.74
	LL	.16	.63	.30	.74
	GF	.44	.80	.49	.84
	PLS _{unadj}	.03	.51	.09	.62

Note: LN, W, LL and GF are the adjusted methods in ACPRAFT.

LN, LL and W. At each censoring level, GF had the largest AUC and Youden index, similar to what was observed for scheme 1 in Figure 5.1 and Table 5.2. Because GF outperforms its special cases, we will focus only on GF, sAFT and the unadjusted case for the remainder of the analyses.

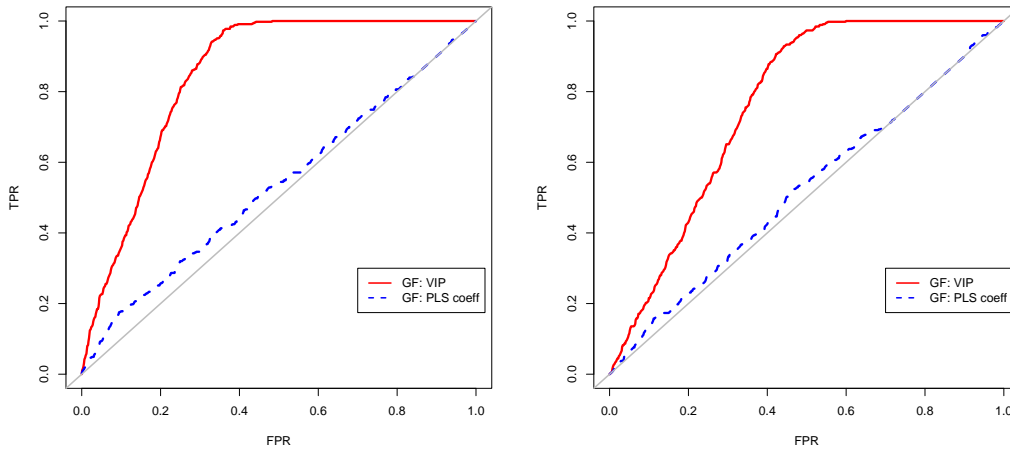


Figure 5.2: ROC Curves for Simulation Scheme 2, VIP vs. PLS Coeff for GF

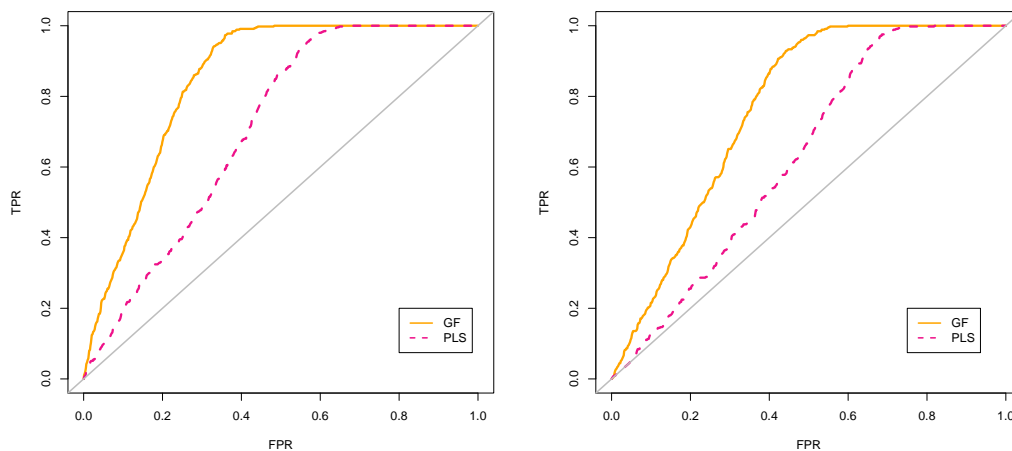


Figure 5.3: ROC Curves for Simulation Scheme 2, GF vs. $\text{PLS}_{\text{unadj}}$ using VIP

Simulation Scheme 1: Full Analyses

In this section, we evaluate the performance of our methods using data simulated from scheme 1 described in §4.3.1. In each case, we examine average AUCs, sensitivities and specificities across 200 simulations. Table 5.3 shows the AUCs for simulation scheme 1 using VIP as the variable selection tool. Across each scheme and censoring level, we discover that our adjusted methods (GF and sAFT) have a clear advantage over the unadjusted case, and as the censoring increases, this difference becomes even more apparent. In fact, $\text{PLS}_{\text{unadj}}$ has the lowest AUC in every single scheme for the 33 and 67% censoring cases. Thus, we observe a clear benefit when imputing censored observations using ACPRAFT. Next, we examine differences between GF- and sAFT-based ACPRAFT. Both GF and sAFT perform similarly in many instances, but there are particular schemes where one outperforms the other. For example, GF's AUCs are higher for the LN, LL1 and LL2 cases, and sAFT's

AUCs are higher in the W1 and W2 cases. In Table 5.4, AUCs using the PLS coefficients are shown and observed to be lower than VIP in each case. Thus, VIP is the optimal choice for variable selection. However, in this case, we still observe GF and sAFT outperforming the unadjusted CPRAFT method.

Table 5.3: AUC using VIP, Simulation Scheme 1

		LN	LL1	LL2	W1	W2
33 % cens.	GF	.91	.92	.69	.95	.89
	sAFT	.91	.92	.76	.97	.92
	PLS _{unadj}	.88	.89	.74	.93	.89
67 % cens.	GF	.87	.89	.72	.90	.76
	sAFT	.86	.84	.71	.95	.84
	PLS _{unadj}	.80	.78	.65	.85	.72
80 % cens.	GF	.85	.84	.75	.87	.64
	sAFT	.81	.77	.68	.92	.70
	PLS _{unadj}	.72	.66	.59	.77	.56

Note: GF and sAFT are the adjusted methods in ACPRAFT.

Table 5.4: AUC using PLS Coeff, Simulation Scheme 1

		LN	LL1	LL2	W1	W2
33 % cens.	GF	.67	.64	.52	.66	.56
	sAFT	.68	.64	.55	.71	.54
	PLS _{unadj}	.60	.57	.53	.62	.54
67 % cens.	GF	.71	.63	.57	.62	.52
	sAFT	.66	.55	.56	.73	.55
	PLS _{unadj}	.54	.51	.53	.54	.50
80 % cens.	GF	.68	.63	.60	.62	.51
	sAFT	.59	.54	.54	.75	.53
	PLS _{unadj}	.51	.51	.51	.53	.50

Note: GF and sAFT are the adjusted methods in ACPRAFT.

Next, we focus our attention on the Youden index displayed in Table 5.5 for each scheme and censoring level using both PLS coefficients and VIP as ranking methods. First, we note that Youden index values based on PLS

coefficients are lower than those based on VIP in every single case. Thus, as expected, VIP significantly outperforms the PLS coefficients in ranking genes in every scenario. Furthermore, we observe that GF and sAFT outperform the unadjusted method in almost every scenario, and just as we had observed with the AUCs, this difference becomes more apparent as the censoring increases. Thus, we notice a clear improvement when censored observations are adjusted. GF and sAFT perform similarly in many cases, but as seen in the AUC results, one occasionally shows a clear advantage over the other.

Table 5.5: Youden Index, Simulation Scheme 1

	Youden Index	33% cens.		67% cens.		80% cens.	
		PLS Coeff	VIP	PLS Coeff	VIP	PLS Coeff	VIP
LN	GF	.21	.52	.29	.55	.22	.52
	sAFT	.21	.52	.19	.52	.10	.48
	PLS _{unadj}	.10	.50	.03	.41	.01	.17
LL1	GF	.15	.56	.15	.54	.14	.45
	sAFT	.15	.56	.05	.43	.04	.24
	PLS _{unadj}	.06	.50	.01	.21	.01	.07
LL2	GF	.02	.20	.07	.23	.11	.31
	sAFT	.05	.29	.06	.20	.04	.14
	PLS _{unadj}	.03	.23	.03	.10	.01	.05
W1	GF	.18	.62	.12	.52	.14	.50
	sAFT	.24	.70	.31	.65	.34	.61
	PLS _{unadj}	.11	.56	.04	.40	.03	.21
W2	GF	.04	.45	.02	.22	.01	.10
	sAFT	.07	.52	.05	.39	.04	.18
	PLS _{unadj}	.03	.43	.01	.12	.01	.02

Note: GF and sAFT are the adjusted methods in ACPRAFT.

Simulation Scheme 2: Full Analyses

We now examine simulation scheme 2, which introduces correlations between genes. Recall that in simulation scheme 1, GF and sAFT performed similarly

and better than the unadjusted approach, and VIP outperformed the PLS coefficients. In this section, we will show that the same conclusions can be made for scheme 2. However, although the results have a similar trend, the AUC and Youden values for scheme 2 are lower than those in scheme 1 due to the complexity in the data introduced by the correlation structure between genes. The AUCs for scheme 2 can be seen in Table 5.6. Similar to scheme 1, we observe that our adjusted methods have a clear advantage over the unadjusted case and this difference becomes greater as the censoring increases. Looking at GF and sAFT, we notice that GF results in higher AUCs for LN and LL2, which differs slightly from the scheme 1 results, but GF and sAFT perform very similarly in the remaining schemes. In Table 5.7, AUCs using the PLS coefficients are shown and observed to be much lower than VIP in each case. In fact, in most cases, the AUC is approximately .50, which means that its selection capability is similar to a coin flip. Thus, VIP is the optimal choice for variable selection.

Table 5.6: AUC using VIP, Simulation Scheme 2

		LN	LL1	LL2	W1	W2
33 % cens.	GF	.58	.82	.66	.68	.61
	sAFT	.57	.89	.66	.67	.60
	PLS _{unadj}	.54	.84	.67	.67	.62
67 % cens.	GF	.82	.88	.70	.79	.67
	sAFT	.75	.88	.69	.69	.68
	PLS _{unadj}	.63	.80	.70	.69	.66
80 % cens.	GF	.84	.86	.84	.83	.70
	sAFT	.76	.88	.76	.74	.71
	PLS _{unadj}	.68	.78	.72	.69	.67

Note: GF and sAFT are the adjusted methods in ACPRAFT.

Table 5.7: AUC using PLS Coeff, Simulation Scheme 2

		LN	LL1	LL2	W1	W2
33 % cens.	GF	.50	.55	.50	.51	.51
	sAFT	.51	.56	.50	.51	.51
	PLS _{unadj}	.50	.52	.50	.51	.50
67 % cens.	GF	.50	.56	.50	.52	.50
	sAFT	.50	.54	.50	.51	.50
	PLS _{unadj}	.50	.50	.50	.51	.50
80 % cens.	GF	.50	.57	.50	.52	.50
	sAFT	.50	.54	.50	.51	.50
	PLS _{unadj}	.50	.50	.50	.51	.50

Note: GF and sAFT are the adjusted methods in ACPRAFT.

Next, we examine the Youden index displayed in Table 5.8. First, we note that, similar to scheme 1, the Youden index values based on PLS coefficients are lower than those based on VIP in every single case. In fact, in many cases, the Youden index for PLS coefficients is 0. In these results, we also observe GF and sAFT outperform the unadjusted case in most scenarios, specifically for higher censoring. GF and sAFT perform similarly in many cases, but as seen in the AUC results, one occasionally shows a clear advantage over the other.

VIP vs. PLS Coefficients Through data simulation, we were able to compare two approaches for variable selection with (A)CPRAFT. The first method was based on the absolute value of the PLS coefficients and the second was based on VIP. We found that VIP showed a significant improvement over PLS coefficients in mock gene selection. It has also been shown in the literature that using a threshold of 1 (i.e., selecting variables with a $VIP > 1$) produces more relevant variables (Mehmood *et al.* 2012). Using this threshold is one

Table 5.8: Youden Index, Simulation Scheme 2

	Youden Index	33% cens.		67% cens.		80% cens.	
		PLS Coeff	VIP	PLS Coeff	VIP	PLS Coeff	VIP
LN	GF	0	.01	0	.11	0	.13
	sAFT	.01	.01	.01	.09	0	.09
	PLS _{unadj}	0	.01	0	.01	0	.03
LL1	GF	.05	.19	.06	.24	.06	.20
	sAFT	.06	.24	.04	.23	.04	.22
	PLS _{unadj}	.02	.16	.01	.10	0	.08
LL2	GF	0	.03	0	.04	0	.13
	sAFT	0	.03	0	.03	0	.06
	PLS _{unadj}	0	.03	0	.03	0	.05
W1	GF	.01	.04	.01	.09	.02	.15
	sAFT	.01	.03	.02	.05	.01	.07
	PLS _{unadj}	0	.03	.01	.04	0	.04
W2	GF	.01	.02	0	.03	0	.04
	sAFT	.01	.02	0	.04	0	.03
	PLS _{unadj}	0	.02	0	.02	0	.02

Note: GF and sAFT are the adjusted methods in ACPRAFT.

way to avoid selecting arbitrary cut-off points. We plan to investigate the VIP threshold in future work.

CPRAFT vs. ACPRAFT In this section, we were also able to compare the performance of CPRAFT and ACPRAFT using extensive simulations, specifically for its special case using PLS ($\alpha = .5$). These methods were evaluated and compared using AUCs and Youden index across 200 simulated data sets generated by our two simulation schemes. While some variation in the results was observed, there were some common findings. First, when comparing the performance of ACPRAFT to CPRAFT, we observed that all ACPRAFT methods outperformed CPRAFT in terms of mock gene selection, particularly as the censoring increased. Recall that ACPRAFT adjusts the

censored observations based on various pre-specified models, while CPRAFT makes no such adjustment. Thus, we can conclude that this adjustment has a positive effect on gene selection, and the improvements become even more apparent as the censoring proportion increases. As for ACPRAFT methods, as a whole they performed fairly similar across all the data sets. In the initial analyses, we observed that GF outperformed LN, LL and W, and in general, GF has an advantage because it is a broad model that incorporates these models and others as special cases. Thus, for application purposes, it provides a nice umbrella structure for all the models of potential interest. The sAFT approach is also based on a flexible model, as it assumes no distribution for the error term. It performed similarly, and better in some cases, to GF.

5.4.2 Examples

In the previous section, we compared ACPRAFT and CPRAFT using $\alpha = .5$ in terms of mock gene selection using VIP on various simulated data sets. In this section, we apply this (A)PLSAFT to the oral and ovarian data sets and explore using VIP for gene selection. We also look at survival curves using the CPR coefficients and components and discuss their usefulness and limitations in predicting survival. The goal in this section is both to explain how the dimension of the data can be reduced, but more importantly how gene selection can be accomplished.

We observed in our simulations that VIP produced more accurate results than PLS coefficients and, hence, we focus on this measure in this section. PLSAFT extracted 5 and 2 components from the oral and ovarian datasets,

respectively, which explained 84% and 70% of the variation in the data. Next, APLSAFT was applied to these datasets based on the GF and AFT models. The simulation results showed GF either outperformed or matched the performance of LN, LL and W, so we focus only on GF. Using GF, the final number of PLS components was 9 for oral and 15 for ovarian, and the proportion of variation explained was approximately 100% for both. Using sAFT, the final number of PLS components was 5 for oral and 15 for ovarian, and the proportion of variation explained was approximately 97% for oral and 100% for ovarian. For the oral data, the proportion of variation explained for 5 components was 97% using GF and sAFT, which is higher than PLSAFT. For the ovarian data, the proportion of variation explained for 2 components was 95% for GF and 93% for sAFT, which are both higher than PLSAFT.

VIP Analysis In the simulations, VIP was used as a ranking measure, where we selected the top 400 mock genes known to have a significant association with survival. In the analysis of real data, the threshold is unclear. Thus, while we note that its gene selection function was investigated and determined to perform well in the simulations, we will focus on other methods of evaluation here. We investigate thresholding VIP in the next section.

Here, we employ the weighted average technique discussed in §3.2. We use all the genes and weight their expressions by their actual VIP values. Essentially, this will put a heavier weight on genes deemed significant. Then, we dichotomize this weighted average and plot the survival curves for the “High” and “Low” weighted average values to get a visual representation of

the data. Goodness of fit (GOF) tests are also run for the PH (Grambsch and Therneau 1994), PO (Martinussen and Scheike 2006) and semiparametric AFT (Novak 2010) models on the continuous weighted averages. Figure 5.4 shows the resulting survival curves and GOF for each method, and the results are interesting. The oral results can be seen in plots (a) - (c). All three have a similar shape and show fairly proportional hazards. This is emphasized by looking at GOF, where we see PH fitting in each case (GOF $p > .05$). We note that in these cases, the PO and AFT models also fit. This is not surprising since the PH and PO models intersect with AFT. For the ovarian data, all three curves have a very similar shape and show crossing hazards, which is evidence of NPH. To formally check for NPH, we run a GOF test for the PH model on the continuous weighted averages and show the results with each curve. In all three cases, the GOF tests reaffirm that the PH model does not fit (GOF $p < .05$), and therefore, NPH is present. In addition, we performed formal GOF tests on the PO and AFT models (Martinussen and Scheike 2006; Novak 2010). Our results, shown in the figures below, suggest that the PO model does not fit the ovarian data but the more general AFT model provides a good fit. On top of GOF, model significance was also checked, and in most cases, the model of interest is significant ($p < .05$) for its respective weighted average.

VIP Thresholding In this section, we evaluate VIP as a variable selection measure via thresholding. As discussed in Mehmood *et al.* (2012), it is common practice to select a variable whose VIP value, ν_j , is greater than 1. Thus,

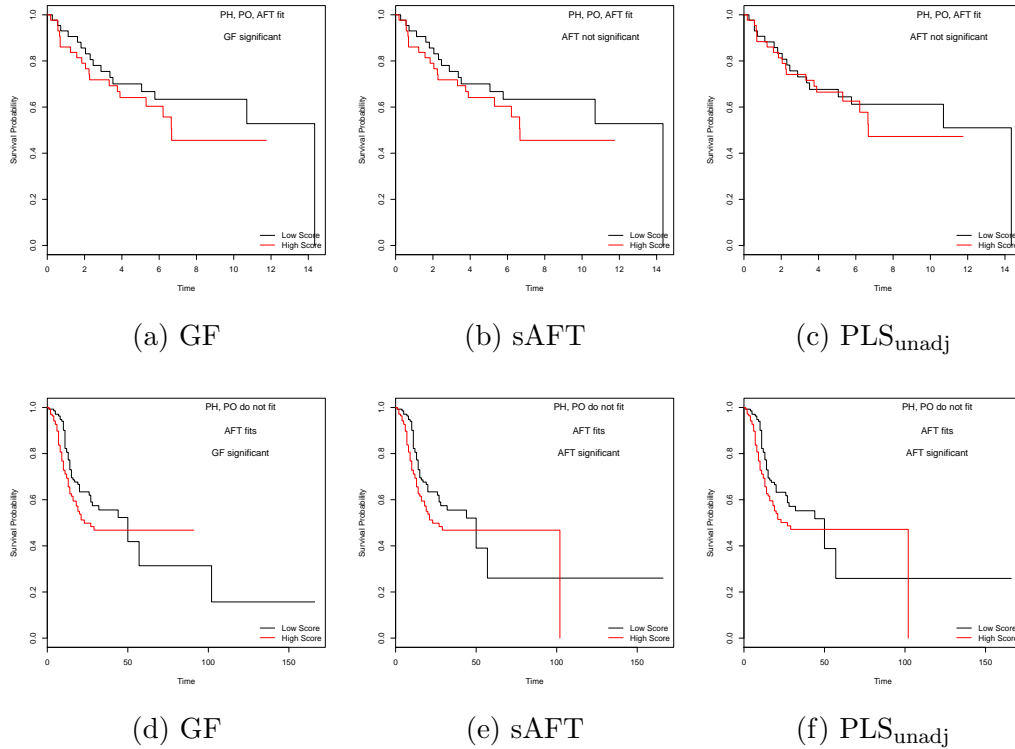


Figure 5.4: Survival Curves for Dichotomized VIP Weighted Average
(a)-(c) Oral, (d)-(f) Ovarian

in this section, the VIP values are calculated for each gene, so $\nu = \nu_1, \dots, \nu_p$, and genes with $\nu_j > 1$ are selected as having a relationship with survival and evaluated. We first note that this is a preliminary and purely exploratory analysis on thresholding, but we hope to examine this more in future work.

The VIP thresholding analysis was performed for both the oral and ovarian datasets. The selected genes using the adjusted GF and sAFT methods can be seen in Figure 5.5. For the oral data, both methods select slightly over 4,700 genes, which is approximately 37% of the total genes in the data set. For the ovarian data, both methods select over 13,500 genes, which is

approximately 42% of the total genes in the data. In future work, we hope to examine a more conservative threshold that would reduce this number even farther. In each case, we also note that the GF and sAFT methods select a majority of common genes. In fact, approximately 95% and 98% of the genes selected were common to both GF and sAFT for oral and ovarian, respectively.

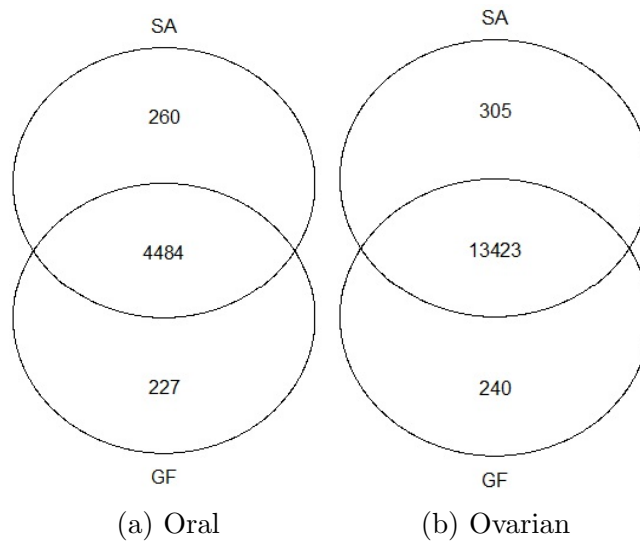


Figure 5.5: Selected Genes using $VIP > 1$

Next, we examine the types of hazards present in the genes selected in an effort to determine if both methods are able to select genes exhibiting NPH. To do this, we look at the proportion of genes for which the PH model does not fit, PO fits, and AFT fits. The results can be seen in Table 5.9. In each case, we discover that over 10% of the genes selected do not fit the PH model, demonstrating their ability to handle NPH. We also point out that in the oral data, the PO model fits 95% of the selected genes, and in the ovarian data, both the PO and AFT models fit over 90% of genes.

Table 5.9: GOF for Genes Selected using $VIP > 1$

	Method VIP > 1	# Genes Selected	% Cox PH not fit	% PO not fit	% Cox PH & PO not fit	% PO fit	% AFT fit
Oral	GF	4,711	10%	5%	3%	95%	78%
	sAFT	4,744	11%	5%	3%	95%	76%
Ovarian	GF	13,663	12%	10%	6%	90%	92%
	sAFT	13,728	12%	10%	5%	90%	92%

Note: GF and sAFT are the adjusted methods in ACPRAFT.

CPR Coefficient Analysis In the simulations, the CPR coefficients were also used as a ranking measure, where we selected the top 400 mock genes known to have a significant association with survival. While we showed that VIP outperformed the CPR coefficients in terms of variable selection, the CPR coefficients have an inherent benefit in calculating the weighted average discussed in §3.2. By calculating $\boldsymbol{\eta} = \mathbf{Z}\hat{\mathbf{w}}$, where $\hat{\mathbf{w}}$ is the CPR coefficient vector, we can interpret $\boldsymbol{\eta}$ as a linear predictor. In fact, in §5.5, we will see that this can be used to build a prediction model, where $\boldsymbol{\eta}$ is the prognostic index used to directly predict log survival time. In this analysis, we begin by plotting the KM survival curves for dichotomized $\boldsymbol{\eta}$, similar to what was done using VIP in Figure 5.4. The results using the CPR coefficients can be seen in Figure 5.6. GOF tests are also run for the PH (Grambsch and Therneau 1994), PO (Martinussen and Scheike 2006) and semiparametric AFT (Novak 2010) models on the continuous weighted averages. For GF, all three models fit and the GF model is significant. For sAFT and the unadjusted method, the PH model does not fit, but the PO and AFT models do fit and the AFT model shows significance. We also note that these figures have a different form than Figure 5.4, and later, we will show that they are more useful for survival prediction.

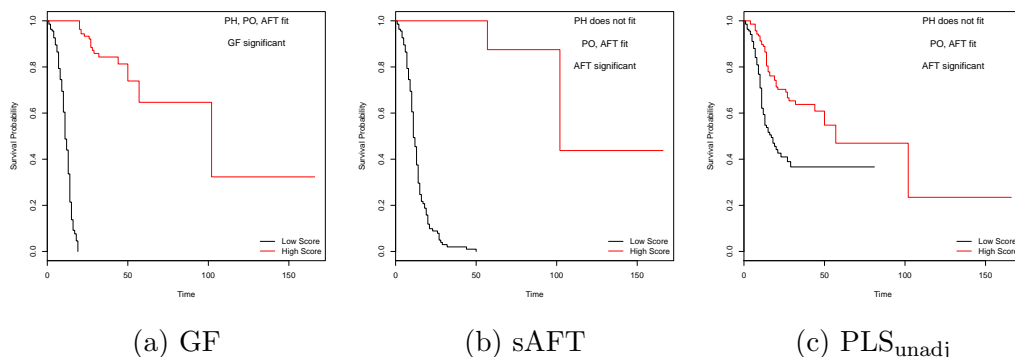


Figure 5.6: Survival Curves for Dichotomized CPR Coeff Weighted Average, Ovarian

Next, we plot the predictive KM survival curves for $\boldsymbol{\eta}$ using the CPR coefficients coming from APLSAFT GF and sAFT. The GOF results are the same as what we saw in Figure 5.7. The difference here is that we plot a single survival curve, where we treat $\boldsymbol{\eta}$ as the predicted survival time (i.e., no censoring). The purpose of this visualization is to emphasize the potential for using these CPR coefficients to build a predictive survival model. These weighted averages and their ability to predict survival time will be evaluated in more detail in §5.5.

Analysis of Reduced Components from APLSAFT We look at the PLS components resulting from the use of GF or sAFT in conjunction with APLSAFT. For the ovarian dataset, GF and sAFT extracted 15 components each and for the oral dataset, these methods extracted 9 and 5 components, respectively. To start this analysis, we focus only on the first two components for illustrative purposes. We dichotomized each component using the median split and plotted KM survival curves in order to visually represent the dimen-

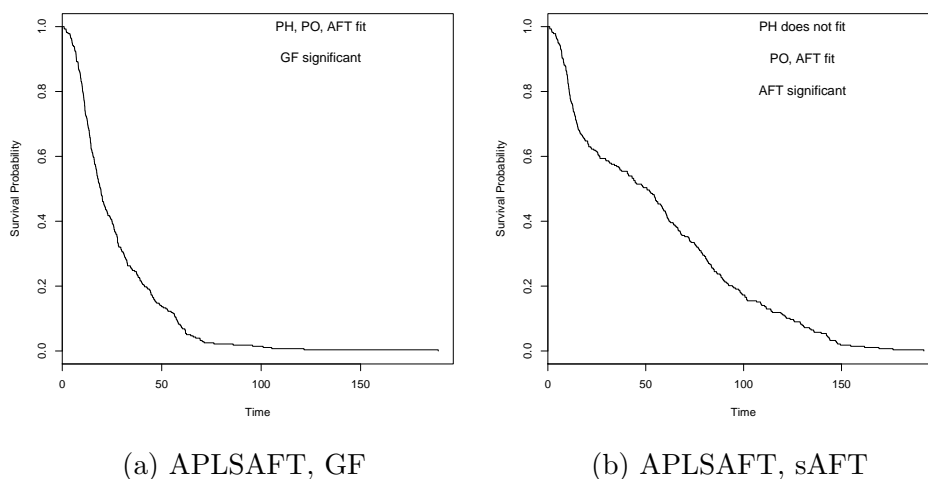
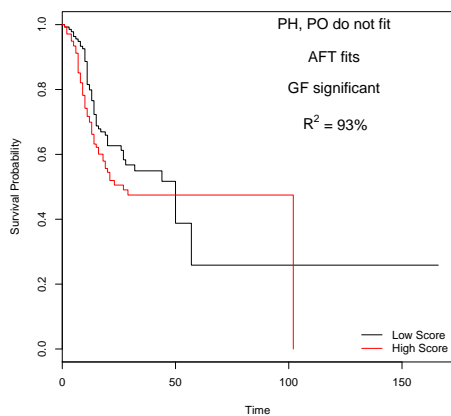


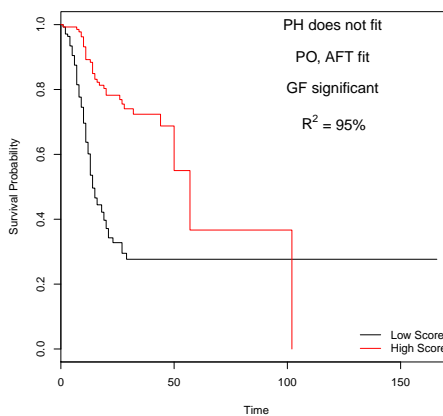
Figure 5.7: Predicted Survival Curves for Weighted Averages using CPR Coefficients

sion reduction. We also performed a GOF test of the PH, PO and AFT models on each component.

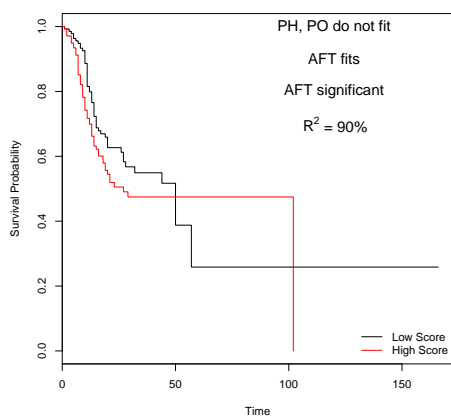
Figure 5.8 shows the survival curves and GOF results for each component for the ovarian dataset. For both GF and sAFT, it is clear that for each component there is evidence of NPH because PH does not fit. For GF, we observe evidence of crossing hazards in both components, and for sAFT, we see crossing hazards in the first component and converging hazards in the second. The PO GOF results also show us something interesting. The PO model fits well for the second component (GOF $p > .05$), but not for the first (GOF $p < .05$). To get a visual representation, the odds curves were plotted in Figure 5.9 for GF. We note that the odds appear to be proportional in the second component but not in the first component, which supports our PO GOF results. On top of GOF, model significance was also checked, and in each



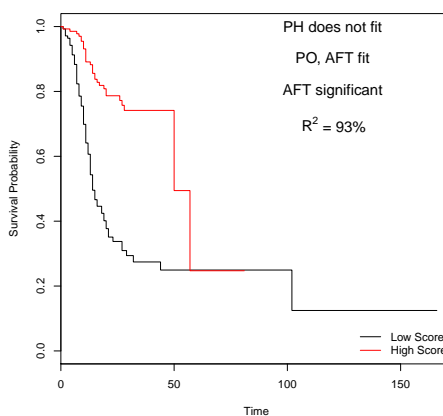
(a) GF, Comp 1



(b) GF, Comp 2



(c) sAFT, Comp 1



(d) sAFT, Comp 2

Figure 5.8: Ovarian: Survival Curves for Dichotomized PLS Components

case, the model of interest is significant ($p < .05$) for its respective component.

Next, we look at similar survival curves for the oral data in Figure 5.10. In these cases, we observe a different result than the ovarian data. For the first component of GF and sAFT, there is clear evidence of PH. However, we note that the PO and AFT models still both fit in these cases. In the second component, we discover evidence of diverging and converging survival curves

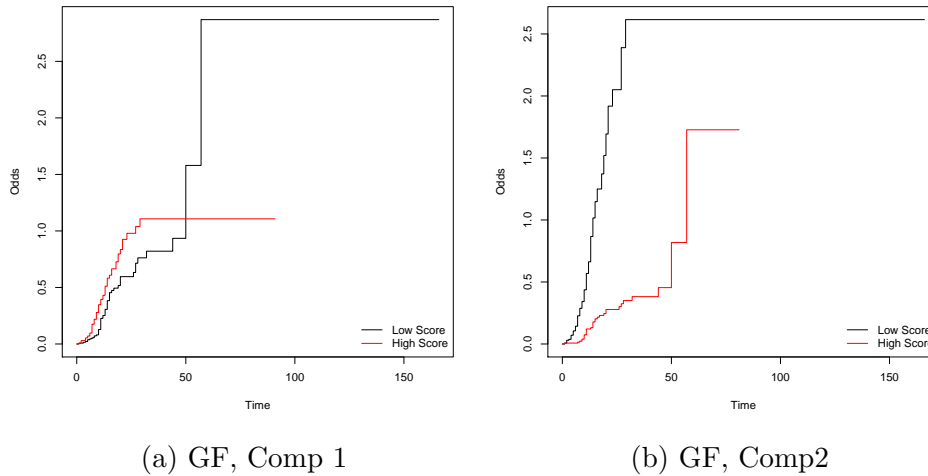
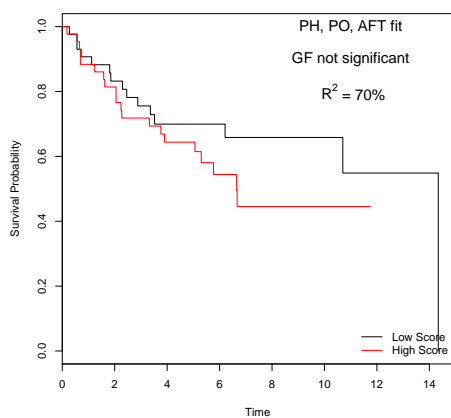


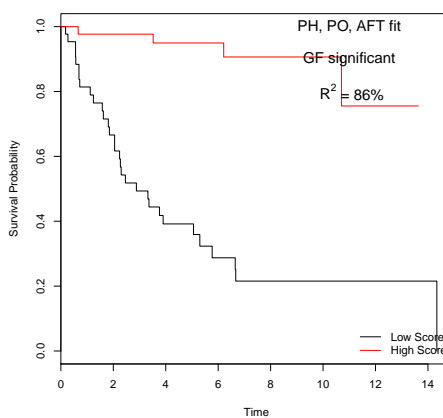
Figure 5.9: Ovarian: Odds Curves for Dichotomized PLS Components

for GF and sAFT, respectively. However, the PH model still fits, indicating that the PH assumption is not violated. The PO and AFT models also fit this case, which is not surprising since they do overlap with the PH model. Thus, they still offer a potentially more useful alternative.

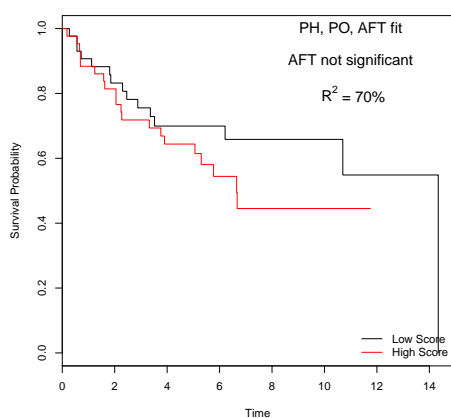
Next, we examine a weighted average using the PLS components, created using step 3a in the ACPRAFT algorithm in §5.1.2. Here, the GF and AFT models are fit to the final GF and sAFT components, respectively. We then calculate $\boldsymbol{\eta} = \mathbf{u}\boldsymbol{\beta}$, where \mathbf{u} are the components and $\boldsymbol{\beta}$ are the coefficients coming from the GF and AFT model fits. Similar to the weighted average calculated using the CPR coefficients, this $\boldsymbol{\eta}$ is a linear predictor that can be seen as a prognostic index used to directly predict survival time. In this analysis, we plot the KM survival curves for $\boldsymbol{\eta}$ using the components and model fit coming from APLSAFT GF and sAFT. GOF tests are also run on the continuous weighted averages, and the results for the ovarian data can be seen in Figure



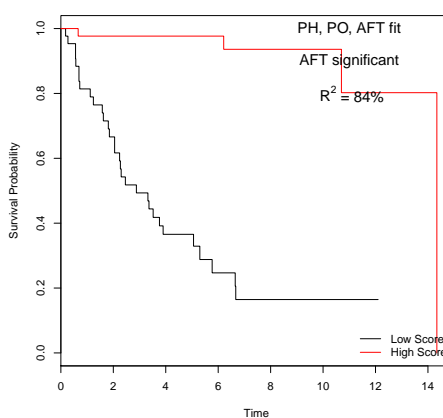
(a) GF, Comp 1



(b) GF, Comp 2



(c) SA, Comp 1



(d) SA, Comp 2

Figure 5.10: Oral: Survival Curves for Dichotomized PLS Components

5.11. For both GF and sAFT, all three models fit and the GF and AFT models are significant for their respective methods. Because the first two components account for a large fraction of the variation in survival time (Figure 5.8), it would also be useful to develop a similar weighted average using only the first two components. Unfortunately, these approaches have a major disadvantage in terms of survival prediction because component information is not available

for new subjects. Alternatively, gene expression is available for new subjects, and thus, the CPR coefficients are more useful in building a prediction model. In §5.5, we discuss this in more detail.

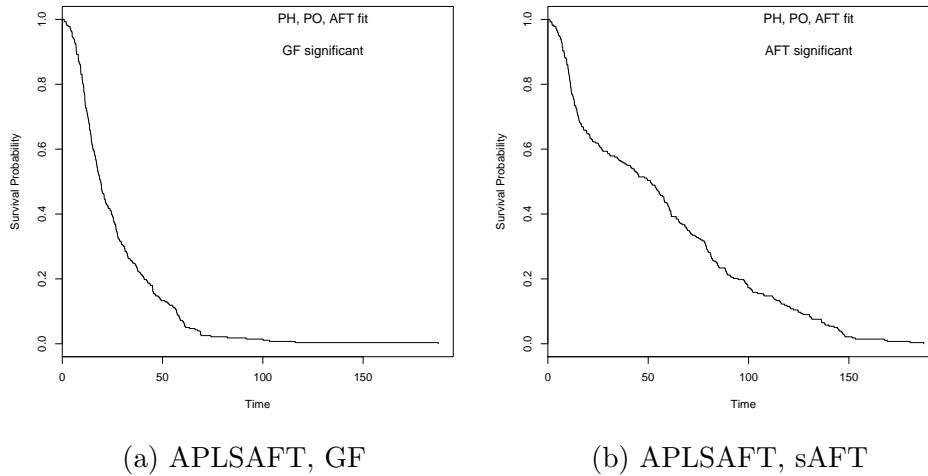


Figure 5.11: Predicted Survival Curves for Weighted Averages using PLS Components

In this section, we proposed a supervised dimension reduction approach based on the AFT model that directly addresses the issue of censored observations. Our proposed approach was based on CPR and the parametric AFT model using GF and the semiparametric AFT model, and it was shown to outperform special cases of parametric AFT. In general, ACPRAFT outperformed CPRAFT in all scenarios for a fixed $\alpha = .5$ in terms of gene selection. We also examined three types of weighted averages. The first and second approaches used VIP and the CPR coefficients to weight gene expression, respectively, and the third approach used the GF and AFT model fits on the final CPR components coming from APLSAFT GF and sAFT. Looking

at the individual component plots in Figure 5.8 and VIP weighted averages in Figure 5.4, we can see that the plots have similar shapes, and thus, using the first two components produces very similar information as VIP. We also see that the R^2 is larger than 90% for the first two components. For this reason, it would be useful to develop a prediction model using the first two or all of the components, similar to what was done in Figure 5.11. However, as mentioned before, this has an inherent practical issue. The ultimate goal is to develop a survival prediction model that will predict a new subject's survival time, but there is no component information for new subjects. Alternatively, gene expression information is available for each new subject, and thus, VIP or the CPR coefficients are a more useful alternative. While a weighted average using VIP could serve as a prognostic index, it does not directly predict survival time. Thus, in the next section, we focus our analysis on using the CPR coefficients to build a prediction model, similar to what was done in Figure 5.7. Its use for survival prediction will be evaluated in the next section at varying α levels.

5.5 Development of a Survival Prediction Model using (A)CPRAFT

In the previous section, we evaluated the performance of (A)CPRAFT as a variable selection tool for a fixed $\alpha = .5$. Through simulations, we showed that VIP is a superior ranking measure over PLS coefficients and that the GF and sAFT cases outperformed the unadjusted method. In this section, our goal is

to build a survival prediction model, and therefore, the CPR coefficients are more useful because they will allow us to directly predict log survival time. We will then compare the performance of our prediction model coming from GF and sAFT to the unadjusted method. Here, we will use the more flexible approach, choosing various optimal α in (A)CPRAFT, after first reducing the data via marginal screening procedures described in Chapters 2 and 4. We develop a survival prediction model and evaluate it using the prognostic index (PI) defined below. We apply this method to large-scale genomic data and evaluate the performance of the GF, sAFT and unadjusted methods using various measures of prediction accuracy.

5.5.1 (A)CPRAFT and Survival Prediction

In §5.1, we noted that CPR maximizes the function $R^2(\mathbf{v}, \mathbf{y})\text{Var}(\mathbf{v})^\gamma$, where $\mathbf{v} = \mathbf{Z}^{(\gamma)}\mathbf{w}$ and $\gamma \equiv \alpha/(1 - \alpha)$, subject to the constraints outlined in §2.6.2. CPR has special cases OLS ($\alpha = 0$), PLS ($\alpha = 1/2$), and PCR ($\alpha = 1$). In this setting, we propose running the (A)CPRAFT procedure for the following choices of α : $\alpha = .25, .5, .75$ and $.95$. Because $p \gg n$, OLS is not useful here. Thus, we select the α corresponding to PLS and a value falling between OLS and PLS, as well as two values between PLS and PCR. Then, we select the (α, K) combination that produces the lowest PRESS based on LOOCV after imputing the censored observations in ACPRAFT. This procedure is performed separately for GF and sAFT, which could potentially result in a different optimal (α, K) combination in each. The optimal (α, K) is used to build a prediction model using CPR coefficients. The α used in the unadjusted

CPRAFT cannot be optimally chosen, so we will consider two unadjusted cases, one for each of the α values chosen for the GF and sAFT methods.

Before applying (A)CPRAFT, we use marginal screening procedures discussed in Chapters 2 and 4 to narrow down the number of genes. In Chapter 4, we showed that our proposed methods performed well in terms of gene selection. By using these as a pre-filter, we ensure that the genes being used for prediction demonstrate some type of relationship with survival, and it also significantly reduces the computation time. This smaller subset of genes is chosen in three ways. First, we apply a univariate semiparametric AFT model fitting approach from §2.2.3 and select genes that are both significant and fit the model at a 0.05 level of significance. Next, we examine a second subset of genes chosen based on our versatile I_{YP} measure in §4.1.1. Genes that are both significant for our measure and fit the univariate YP model at the 0.05 level of significance are selected. Lastly, we select a third subset of genes using concordance regression (CON) defined in §2.3.2, choosing only those genes significant at the 0.05 level. The univariate analyses based on AFT and CON filters are adjusted for age and stage.

Once a subset is selected, (A)CPRAFT is applied and the optimal (α, K) is chosen. The CPR regression coefficients, $\hat{\mathbf{w}}$, are retained for both the adjusted methods (GF and sAFT) as well as for the initial unadjusted step. In each case, $\hat{\mathbf{w}}$ is used to predict the log survival time of subject i given their gene expression profile \mathbf{Z}_i using the prognostic index, $PI = \mathbf{Z}_i \hat{\mathbf{w}}$.

For this application, the data will first be split into training and test sets, where $\hat{\mathbf{w}}_{TR}$ will be found using the training set and used to predict the

log survival time in the test set. Thus, $PI = \mathbf{Z}\hat{\mathbf{w}}_{TR}$, where \mathbf{Z} will be from the test set, and this PI can be evaluated for its prediction accuracy using the measures described in the following section.

In summary, our (A)CPRAFT survival prediction algorithm consists of the following steps:

1. Filter data using three separate approaches.
 - Univariate model fitting: Semiparametric AFT and YP models
 - Univariate concordance regression (CON) from Dunkler *et al.* (2010)
2. Randomly split filtered data into training (67% of subjects) and test set (33% of subjects). This process is repeated $N = 25$ times and the median of the prediction accuracy measures in step 5 are reported.
3. Apply (A)CPRAFT (unadjusted, GF and sAFT) to training set using $\alpha = (.25, .5, .75, .95)$.
 - Choose optimal (α, K) combination.
 - Retain the CPR regression coefficients, $\hat{\mathbf{w}}_{TR}$.
4. Use $\hat{\mathbf{w}}_{TR}$ from Step 3 to predict log survival times in the test set.
 - Calculate $PI = \mathbf{Z}\hat{\mathbf{w}}_{TR}$, where \mathbf{Z} is from the test set.
5. Evaluate PI 's ability to predict log survival time using the measures of prediction accuracy described in the next section.

5.5.2 Measures of Prediction Accuracy

The *PI* measure calculated above is essentially \hat{Y} , the predicted log survival times. Thus, the method is effective and useful if *PI* explains the actual survival times. In this section, we discuss various measures of prediction accuracy that will be used to evaluate the performance of these methods.

R_{Comp}^2 This measure reports the percentage of variation in Y that is being explained by the CPR components coming from (A)CPRAFT.

Mean Squared Error (MSE) This measure is given by $MSE = \frac{1}{n^*} \sum_{i=1}^n \delta_i (\hat{Y}_i - \log T_i)^2$, where $n^* = \sum_{i=1}^n \delta_i$ and $\delta_i = 1$ implies the event was observed. This will be calculated for both the training set, MSE_{TR} , and the test set, MSE_{TE} .

Time Dependent AUC This method is discussed in Blanche *et al.* (2013). It calculates time dependent ROC curves and AUCs at varying time points, specifically $t = 3$ years and $t = 5$ years. Essentially, this measure quantifies the methods' ability to predict 3 and 5 year survival, with values closer to 1 indicating better prediction accuracy.

All of these measures are used in the following section to evaluate and compare the predictive performance of ACPRAFT using GF and sAFT models to the unadjusted CPRAFT.

5.5.3 Examples

The (A)CPRAFT prediction algorithm and analyses described in 5.5.1 and 5.5.2 were applied to the oral, ovarian and RNA-seq data described in Chapter

2. The oral data has 86 subjects and 12,766 genes, the ovarian data has 276 subjects and 32,575 genes, and the RNA-seq data has 221 subjects and 19,341 genes. Both the oral and ovarian data have 59% censored observations, and the RNA-seq data has 62%. The high censoring proportion and the large number of genes make each set a perfect candidate for our (A)CPRAFT prediction algorithm, which will attempt to reduce the data and simultaneously adjust the censored observations in an effort to accurately predict survival time. We also note that the RNA-seq analysis was performed using linear gene expression, unlike the oral and ovarian data which was done on the \log_2 scale.

After the initial filtering step, the (A)CPRAFT algorithm was run 25 times and the resulting median values of prediction accuracy are shown in Table 5.10 for the ovarian dataset. Although not reported, the measures performed consistently with most having standard deviations less than 0.1 across the 25 runs. The filtering mechanisms each reduce the data by a significant fraction, particularly for the AFT and CON filters. The R_{Comp}^2 values, which reflect the percentage of variation in survival time being explained by the CPR components, is significantly higher for GF and sAFT compared to both unadjusted cases. In fact, in the majority of cases, the GF and sAFT values are greater than 95%, while the unadjusted R^2 values range mostly from 69-78%. Also, we observe larger MSE for the test data compared to the training data. However, in both cases, we observe consistently smaller values for the ACPRAFT methods compared to CPRAFT, indicating that GF and sAFT result in more accurate predictions. Lastly, we examine the AUCs calculated for both 3 and 5 year survival. Not surprisingly, the training set AUCs are

larger than those of the test set in each case, but in both the training and test sets and across all three filtering mechanisms, we observe larger AUCs for GF and sAFT compared to the unadjusted methods. For example, using the CON filter, we observe an AUC range of .82-.93 for GF and sAFT, while the unadjusted methods range only from .69-.76. Although the results are not shown here, a hypothesis test was also performed to test whether there is a statistically significant difference between the AUCs of each method. At the $\alpha = .05$ level, majority of the tests showed that there was a significant difference between the ACPRAFT and CPRAFT AUCs. For example, for the AFT filter, 80% of the 25 sets showed significance in each case, indicating that GF and sAFT outperform their unadjusted methods in terms of 3 and 5 year prediction.

A similar analyses was performed on the oral dataset and the median results are shown in Table 5.11 . Similar to the ovarian results, the filtering mechanisms each reduce the data by a significant fraction. Although the differences are less extreme than what was observed for the ovarian dataset, the R_{Comp}^2 values are still consistently higher for GF and sAFT compared to the corresponding unadjusted cases. Looking at the MSE values, we observe significantly smaller values for ACPRAFT compared to CPRAFT in each case, implying that GF and sAFT outperform the unadjusted methods. For AUCs, however, the differences are not as extreme. We still observe larger AUCs for the training set compared to the test set, but the differences between the methods are much smaller. For example, in the AFT filtered results, the AUCs range from .86-1 for GF and sAFT and from .86-.98 for the unadjusted meth-

Table 5.10: Ovarian: Measures of Prediction Accuracy, Median Results

			GF	sAFT	Unadj (GF α)	Unadj (sAFT α)
Filter: AFT 3,550 genes	Training Set Results	R_{Comp}^2	95.5	99.8	72.3	71.4
		MSE_{TR}	0.07	0.02	0.44	0.47
		AUC, $t = 3$	0.85	1	0.8	0.84
		AUC, $t = 5$	0.88	1	0.76	0.83
	Test Set Results	MSE_{TE}	0.53	0.5	0.58	0.51
		AUC, $t = 3$	0.77	0.86	0.71	0.73
AUC, $t = 5$		0.75	0.83	0.66	0.66	
Filter: CON 1,858 genes	Training Set Results	R_{Comp}^2	94.3	96	69.2	69.6
		MSE_{TR}	0.24	0.48	0.54	0.58
		AUC, $t = 3$	0.89	0.93	0.76	0.76
		AUC, $t = 5$	0.87	0.93	0.75	0.76
	Test Set Results	MSE_{TE}	0.29	0.66	0.67	0.73
		AUC, $t = 3$	0.84	0.82	0.69	0.69
AUC, $t = 5$		0.82	0.83	0.7	0.72	
Filter: YP 11,034 genes	Training Set Results	R_{Comp}^2	98.1	99.9	78.4	93.9
		MSE_{TR}	0.02	0.01	0.49	0.46
		AUC, $t = 3$	0.9	1	0.76	0.75
		AUC, $t = 5$	0.85	0.99	0.76	0.75
	Test Set Results	MSE_{TE}	0.78	0.73	0.85	1.19
		AUC, $t = 3$	0.72	0.75	0.55	0.57
AUC, $t = 5$		0.71	0.74	0.55	0.56	

Note: GF and sAFT are the adjusted methods in ACPRAFT.

ods. However, in each individual case, we still observe an improvement in ACPRAFT over CPRAFT.

A similar analyses was performed on the RNA-seq dataset and the median results are shown in Table 5.12 . This analysis was applied to linear gene expression, but we plan to apply the same analysis on \log_2 expression in our future work. Similar to the oral and ovarian results, the filtering mechanisms each reduce the data by a significant fraction. The R_{Comp}^2 values are still consistently higher for GF and sAFT compared to the corresponding unadjusted

Table 5.11: Oral: Measures of Prediction Accuracy, Median Results

			GF	sAFT	Unadj (GF α)	Unadj (sAFT α)
Filter: AFT 3,065 genes	Training Set Results	R_{COMP}^2	94.6	99	93.7	95.1
		MSE_{TR}	0.03	0.02	0.21	0.14
		AUC, $t = 3$	0.97	1	0.96	0.98
		AUC, $t = 5$	0.96	1	0.97	0.98
	Test Set Results	MSE_{TE}	0.1	0.34	1.08	0.99
		AUC, $t = 3$	0.89	0.92	0.89	0.89
		AUC, $t = 5$	0.94	0.95	0.93	0.94
Filter: CON 4,652 genes	Training Set Results	R_{COMP}^2	95.6	95.6	89.3	88.3
		MSE_{TR}	0.01	0.12	0.4	0.4
		AUC, $t = 3$	0.98	0.99	0.97	0.97
		AUC, $t = 5$	0.97	0.99	0.96	0.98
	Test Set Results	MSE_{TE}	0.36	0.44	1.2	1.17
		AUC, $t = 3$	0.86	0.88	0.86	0.86
		AUC, $t = 5$	0.9	0.92	0.9	0.91
Filter: YP 8,411 genes	Training Set Results	R_{COMP}^2	94.3	100	89.5	96.4
		MSE_{TR}	0.04	0.08	0.38	0.15
		AUC, $t = 3$	0.98	1	0.97	1
		AUC, $t = 5$	0.99	1	0.97	1
	Test Set Results	MSE_{TE}	0.11	0.51	1.2	1.2
		AUC, $t = 3$	0.86	0.87	0.83	0.85
		AUC, $t = 5$	0.89	0.93	0.88	0.9

Note: GF and sAFT are the adjusted methods in ACPRAFT.

Table 5.12: RNA-Seq: Measures of Prediction Accuracy, Median Results

			GF	sAFT	Unadj (GF α)	Unadj (sAFT α)
Filter: AFT 3,065 genes	Training Set Results	R_{COMP}^2	98.1	99	89	90.6
		MSE_{TR}	1.08	0.70	2.07	1.38
		AUC, t = 3	0.65	0.96	0.58	0.66
		AUC, t = 5	0.67	0.85	0.63	0.69
	Test Set Results	MSE_{TE}	1.72	1.72	2.9	2.03
		AUC, t = 3	0.63	0.79	0.53	0.60
AUC, t = 5		0.61	0.63	0.53	0.58	
Filter: CON 4,652 genes	Training Set Results	R_{COMP}^2	97.7	98.5	88.9	90.4
		MSE_{TR}	1.20	1.06	1.79	1.24
		AUC, t = 3	0.62	0.88	0.59	0.61
		AUC, t = 5	0.69	0.79	0.72	0.71
	Test Set Results	MSE_{TE}	2.09	1.79	2.34	1.95
		AUC, t = 3	0.59	0.70	0.56	0.57
AUC, t = 5		0.69	0.60	0.68	0.62	
Filter: YP 8,411 genes	Training Set Results	R_{COMP}^2	98.1	97.6	90.2	89.7
		MSE_{TR}	1.21	1.52	1.51	1.63
		AUC, t = 3	0.61	0.63	0.52	0.52
		AUC, t = 5	0.67	0.57	0.52	0.51
	Test Set Results	MSE_{TE}	2.00	2.45	2.95	2.52
		AUC, t = 3	0.57	0.53	0.51	0.52
AUC, t = 5		0.61	0.53	0.50	0.51	

Note: GF and sAFT are the adjusted methods in ACPRAFT.

cases. We also observe significantly smaller MSE values for ACPRAFT compared to CPRAFT in each case, implying that GF and sAFT outperform the unadjusted methods. We observe larger AUCs for the training set compared to the test set, and we also observe larger AUCs for GF and sAFT compared to the unadjusted methods in most cases. Thus, in each filtering case, we observe an improvement in ACPRAFT over CPRAFT.

In this section, we developed a prediction algorithm using (A)CPRAFT where the optimal α can be chosen to minimize PRESS based on LOOCV.

Using real-life examples, we showed that our prognostic index, $PI = \mathbf{Z}\hat{\mathbf{w}}_{TR}$, computed using the CPR coefficients, from (A)CPRAFT performed well in predicting patient survival. Specifically, we observed a significant improvement in prediction accuracy when the censored times are imputed using ACPRAFT (GF and sAFT) compared to the unadjusted CPRAFT method. As a comparison, we also ran our prediction algorithm using fixed $\alpha = 0.5$, the case corresponding to PLS, on the ovarian dataset. Although the results are not shown here, we note that its performance mimicked the results from our optimal α algorithm and, thus, offers a competitive alternative with a shorter computation time. Also, as discussed in §5.4.2, the CPR components could be used to build a prediction model using the model fitting in step 3a in the ACPRAFT algorithm in §5.1.2, but the limitation of this approach is that new subjects lack component information. In our future work, we hope to examine this further.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this thesis, we proposed methods for variable selection and supervised dimension reduction for large-scale genomic data with censored survival outcomes. We used different types of genomic data from cancer studies to demonstrate some of the potential statistical challenges in this type of data, specifically the issues of high-dimensionality, censoring, and the presence of NPH. To address the issue of NPH, we proposed gene selection and ranking methods based on models that relax the PH assumption and demonstrated their ability to correctly identify genes with a time-varying effect.

First, we proposed the I_{YP} measure based on KL divergence for the YP model to develop a test for gene effect. This measure contained two important special cases involving the PH and PO models. An advantage of these measures is that they do not require an estimate of the baseline hazard and,

instead, are simple functions of model parameters. Using I_{YP} , we developed a statistical test for gene effect in the YP model, where the test-statistic or p -value could be used as a ranking measure. This test was extended to the PO model and a similar measure, I_{PO} , was developed to test for gene effect in that model. These methods performed well in terms of gene selection and were able to select genes exhibiting NPH in real-life genomic data. However, one limitation of I_{YP} was that it can only be applied to the dichotomized case, but I_{PO} can be applied to continuous covariates. We propose to extend I_{YP} to continuous covariates as part of our future work. Next, we proposed the measures R_{PO}^2 and R_{PLO}^2 that are derived from the partial likelihood of their respective models. These measures have the computational advantage of not requiring the calculation of model parameters and the intuitive advantage of being interpreted as the percentage of separability in gene expression between the event and non-event groups. R_{PO}^2 proved to be a useful measure, outperforming other R^2 measures in the identification of significant genes. Following the approaches in Nagelkerke *et al.* (1991), Allison *et al.* (1995), and O’Quigley *et al.* (2005), we developed alternative R^2 measures for the PO model that utilized the likelihood ratio. These measures extend several currently existing R^2 measures for the PH model to the PO model. However, a detailed study of the properties of these measures will form part of our future work on this topic. These measures performed similarly to that of R_{PO}^2 . Last but not least, we developed R^2 measures for the PH and PO models that are specifically applicable to large-scale genomic data. A significant advantage of these measures is that they only require an estimate of the regression

coefficient β from the respective models.

The variable selection methods address the issue of NPH; however, they do not specifically address the issue of censoring and high-dimensionality. Therefore, we developed supervised dimension reduction methods for censored survival data, termed (A)CPRAFT, that combines CPR with the AFT model and adjusts for censored observations. In particular, we developed a survival prediction algorithm using ACPRAFT based on the generalized F and semi-parametric AFT models. The proposed methods address all three challenges in large-scale genomic data. We then proposed the use of CPR coefficients and VIP for variable selection. Through simulations, we found VIP to be more suitable for gene selection, and we showed that ACPRAFT performs well especially when the proportion of censored observations is high. We also showed that ACPRAFT performed well in terms of predicting survival time.

6.1 Future Work

While our simulations and data examples demonstrated the applicability and usefulness of the proposed variable selection and supervised dimension reduction methods, we note that there is potential for extensions. In this section, we will now discuss our plans for future work on this topic.

6.1.1 Applications

In this proposal, we analyzed the performance of our variable selection measures, R^2 and I , and our supervised dimension reduction approach, (A)CPRAFT,

separately. In our prediction analysis in Chapter 5, we combined the approaches by using I_{YP} as a filtering mechanism before applying (A)CPRAFT. We also are interested in using our R^2 -type measures as a filtering mechanism. In addition, we could apply our R^2 measures to the average gene expression created using VIP and the PLS components in order to quantify the combined effect of all selected genes. This has the potential to help highlight the different model fits. In other words, it would be interesting to see the value of R_{PO}^2 when the PO model fits versus when it does not fit, and the same for R_{PH}^2 and R_{PLO}^2 .

6.1.2 Censoring Variable and Covariate Dependence

In this study, we assumed independence between the censoring variable, C , and gene expression, Z . However, in real data, it is possible that the level of gene expression could have an effect on whether a subject has been censored. Thus, we plan to explore the dependence between C and Z in future work. We also plan to look at the effects of other variables, such as age and stage at diagnosis. Preliminary work was done in this area in the data examples in Chapter 3 and the filtering before the prediction analysis in §5.5, where each univariate model was fit adjusting for age and stage. Adjusting for these confounder variables is expected to produce more accurate results.

6.1.3 Supervised Dimension Reduction Methods

(A)CPRAFT for Variable Selection In this study, we evaluated the variable selection performance of a special case of (A)CPRAFT using $\alpha = .5$,

which uses PLS as the dimension reduction tool. In future work, we plan to evaluate the performance at various α levels. We also hope to investigate more conservative thresholding options for VIP.

(A)CPRAFT for Clustering In this study, we used (A)CPRAFT for variable selection using VIP, but another important application of genomic data is the classification of subjects into categories that could represent phenotypes or genotypes with different survival profiles. For example, we might want to classify subjects into various groups such as tumor type or high/low risk. As part of our future work, we plan to explore the applicability of (A)CPRAFT for patient clustering using the CPR components. In the first stage, (A)CPRAFT would be employed to construct CPR components, and in the subsequent step, a clustering method utilizing these components could be developed. Essentially, given the genomic profile for a new patient, we hope to develop a method for classifying that patient into one of the identified groups.

BIBLIOGRAPHY

- [1] Allison, P.D. *et al.* (1995) *Survival Analysis Using SAS: A Practical Guide*, SAS Publishing.
- [2] Belle, V. V. *et al.* (2011) Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics*, **21(1)**, 87-94.
- [3] Bair, E. *et al.* (2006) Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, **101(473)**, 119-137.
- [4] Bastien, P. *et al.* (2015) Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics*, **31(3)**, 397-404.
- [5] Beer, D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, **8**.
- [6] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, **57(1)**, 289-300.

- [7] Bhattacharjee, A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, **98(24)**, 13790-13795.
- [8] Blanche, H.M. *et al.* (2013) Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*.
- [9] Boulesteix, A.L. and Strimmer, K. (2007) Partial least squares: a versatile tool for the analysis of high dimensional genomic data. *Bioinformatics*, **8**, 32-44.
- [10] Bovelstad, H.M. *et al.* (2007) Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23**, 2080-2087.
- [11] Broet, P. *et al.* (2009) Prediction of Clinical Outcome in Multiple Lung Cancer Cohorts by Integrative Genomics: Implications for Chemotherapy Selection. *Cancer Research*, **69(3)**, 1055-1062.
- [12] Buckley, J. and James, I. *et al.* (1979) Linear Regression with Censored Data. *Biometrika*, **66(3)**, 429-436.
- [13] Cai, T. *et al.* (2009) Regularized estimation for the accelerated failure time model. *Biometrics*, **65(2)**, 394-404.
- [14] Ciampi, A. *et al.* (1986) Regression Analysis of Censored Survival Data with the Generalized F Family—An Alternative to the Proportional Hazards Model. *Statistics in Medicine*, **5**, 85-96.

- [15] Chen, K. *et al.* (2002) Semiparametric Analysis of Transformation Models with Censored Data. *Biometrika*, **89(3)**, 659-668.
- [16] Cheng, S. C. *et al.* (1995) Analysis of Transformation Models with Censored Data. *Biometrika*, **82(4)**, 835-845.
- [17] Cheng, S. C. *et al.* (1997) Survival Probabilities With Semiparametric Transformation Models. *Biometrika*, **92(437)**, 227-235.
- [18] Cox, C. (2008) The generalized F distribution: An umbrella for parametric survival analysis. *Statistics in Medicine*, **27**, 4301-4312.
- [19] Cox, D.R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society*, **34**, 187-220.
- [20] Datta, S. *et al.* (2007) Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Bioinformatics*, **63**, 259-271.
- [21] de Jong, S. *et al.* (2001) Canonical partial least squares and continuum power regression. *Journal of Chemometrics*, **15**, 85-100.
- [22] Dempster, N. M. *et al.* (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39(1)**, 1-38.
- [23] Devarajan, K. and Ebrahimi, N. (2002) Goodness-of-Fit Testing for the Cox Proportional Hazards Model. *Goodness-of-Fit Tests and Model Validity*, **Birkhäuser Boston**, 237-254.

- [24] Devarajan, K. and Ebrahimi, N. (2009) Testing for Covariate Effect in the Cox Proportional Hazards Regression Model. *Communications in Statistics - Theory and Methods*, **38(14)**, 2333-2347.
- [25] Devarajan, K. *et al.* (2010) A supervised approach for predicting patient survival with gene expression data. *Proc IEEE Int Symp Bioinformatics Bioeng*, **5521718**, 26-31.
- [26] Devarajan, K. and Ebrahimi, N. (2011) A semi-parametric generalization of the Cox proportional hazards regression model: Inference and applications. *Computational Statistics and Data Analysis*, **55**, 667-676.
- [27] Devarajan, K. and Ebrahimi, N. (2013) On penalized likelihood estimation for a non-proportional hazards regression model. *Statistics and Probability Letters*, **83**, 1703-1710.
- [28] Dudoit, S. *et al.* (2003) Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, **18(1)**, 71-103.
- [29] Dunkler, D. *et al.* (2010) Gene selection in microarray survival studies under possibly non-proportional hazards. *Bioinformatics*, **26**, 784-790.
- [30] Elston, R. C. and Spence, M. A. (2006) Advances in statistical human genetics over the last 25 years. *Statistics in Medicine*, **25**, 3049-3080.
- [31] Eng, K. H. and Hanlon, B. M. (2012) Discrete mixture regression models for heterogeneous time-to-event data: Cox Assisted Clustering. *Bioinformatics*, **30(12)**, 1690-1697.

- [32] Engler, D. and Li, Yi. (2009) Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies. *Stat Appl Genet Mol Biol*, **8(1)**, 1544-6115.
- [33] Fan *et al.* (2010) High-dimensional variable selection for Coxs proportional hazards model. *Institute of Mathematical Statistics*, **6**, 70-86.
- [34] Geisler, S.A. *et al.* (2002) p16 and p53 protein expression as prognostic indicators of survival and disease recurrence from head and neck cancer. *Clinical Cancer Research*, **8**, 3445-3453.
- [35] Geng, Y. *et al.* (2012) A Model-Free Machine Learning Method for Risk Classification and Survival Probability Prediction. *Stat*, **3(1)**, 337-350.
- [36] Gerds, T.A. and Schumacher, M. (2006) Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, **48(6)**, 1029-1040.
- [37] Goeman, J. J. *et al.* (2005) Testing association of a pathway with survival using gene expression data. *Bioinformatics*, **21(9)**, 1950-1957.
- [38] Grambsch, P. and Therneau, T. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.
- [39] Gui, J. and Li, H. (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21(13)**, 3001-3008.

- [40] Harrell, F. E. *et al.* (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361-387.
- [41] Hastie, T. *et al.* (2001) Supervised harvesting of expression trees. *Genome Biology*, **2**, 1-12.
- [42] Heagerty, P. J. *et al.* (2000) Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*, **56**, 337-344.
- [43] Hong, F. *et al.* (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825-2827.
- [44] Huang, J. and Harrington, D. (2004) Dimension reduction in the linear model for right-censored data: Predicting the change of HIV-I RNA levels using clinical and protease gene mutation data. *Lifetime Data Analysis*, **10**, 425-443.
- [45] Huang, J. *et al.* (2006) Regularized Estimation in the Accelerated Failure Time Model with High-Dimensional Covariates. *Biometrics*, **62(3)**, 813-820.
- [46] Jin, Z. *et al.* (2003) Rank-based inference for accelerated failure time model. *Biometrika*, **90**, 341-353.
- [47] Jin, Z. *et al.* (2006) On least-squares regression with censored data. *Biometrika*, **93**, 147-161.

- [48] Joe, H. (1989) Relative Entropy Measures of Multivariate Dependence. *Journal of the American Statistical Association*, **84**, 157-164.
- [49] Johnson, B. A. (2009) On lasso for censored data. *Electronic Journal of Statistics*, **3**, 485-506.
- [50] Johnson, B. A. (2008b) Estimation in the L_1 -Regularized Accelerated Failure Time Model. *Technical Report, Emory University*.
- [51] Johnson, B. A. *et al.* (2008) Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *Journal of the American Statistical Association*, **103**, 672-680.
- [52] Kaneko, S. *et al.* (2012) Gene Selection using a High-Dimensional Regression Model with Microarrays in Cancer Prognostic Studies. *Cancer Informatics*, **11**, 29-39.
- [53] Klein, P. J. and Moeschberger, L. M. (2003) Survival Analysis: Techniques for censored and truncated data. *New York: Springer*.
- [54] Laimighofer, M. *et al.* (2016) Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression. *Journal of Computational Biology*, **23(4)**, 1-12.
- [55] Leng, C. and Ma, S. (2007) Accelerated failure time models with nonlinear covariates effects. *Aust. N. Z. J. Stat*, **49(2)**, 155-172.

- [56] Li, H. and Gui, J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, **20(1)**, 208-215.
- [57] Li, L. and Li, H. (2004) Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20(18)**, 3406-3412.
- [58] Li, H. and Luan, Y. (2003) Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium of Biocomputing*, **8**, 65-76.
- [59] Li, H. and Luan, Y. (2005) Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, **21(10)**, 2403-2409.
- [60] Li, C. and Li, H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24(9)**, 1175-1182.
- [61] Lin, D. Y. and Wei, L. J. (1989) The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, **84(408)**, 1074-1078.
- [62] Liu, Z. *et al.* (2010) Kernel based methods for accelerated failure time model with ultra-high dimensional data. *BMC Bioinformatics*, **11**, 606.
- [63] Lu, W. and Li, L. (2008) Boosting method for nonlinear transformation models with censored survival data. *Biostatistics*, **9**, 658-667.

- [64] Martinussen and Scheike (2006) Dynamic Regression Models for Survival Data. *Statistics for Biology and Health*.
- [65] Mehmood, T. *et al.* (2012) A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, **118**, 62-69.
- [66] Mendez, E. *et al.* (2009) A genetic expression profile associated with oral cancer identifies a group of patients at high-risk of poor survival. *Clinical Cancer Research*, **15**, 1353-1361.
- [67] Nagelkerke, N. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78(3)**, 691-692.
- [68] Nguyen, D. V. and Rocke, D. M. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625-1632.
- [69] Nguyen, D. V. (2005) Partial least squares dimension reduction for microarray gene expression data with censored response. *Mathematical Biosciences*, **193**, 119-137.
- [70] Novak, P. (2010) Checking goodness-of-fit of the accelerated failure time model for survival data. *WDS'10 Proceedings of Contributed Papers*, **I**, 189-194.
- [71] Nygard, S. *et al.* (2008) Partial least squares Cox regression for genome-wide data. *Lifetime Data Anal*, **14**, 179-195.

- [72] O’Quigley, J. *et al.* (2005) Explained randomness in proportional hazards models. *Statistics in Medicine*, **24(3)**, 479-489.
- [73] Pang, H. *et al.* (2012) Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9(5)**, 1422-1431.
- [74] Park, P. *et al.* (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, 120-127.
- [75] Peri, S. *et al.* (2013) Meta-Analysis Identifies NF-B as a Therapeutic Target in Renal Cancer. *PLoS ONE*, **8(10)**, e76746.
- [76] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- [77] Rosenwald, A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *New England Journal of Medicine*, **346**, 1937-1947.
- [78] Rouam, S. *et al.* (2010) Identifying common prognostic factors in genomic cancer studies: A novel index for censored outcomes. *BMC Bioinformatics*, **11(150)**.
- [79] Rouam, S. *et al.* (2011) A pseudo-R² measure for selecting genomic markers with crossing hazards functions. *BMC Medical Research Methodology*, **11**, 28.

- [80] Saintigny, P. *et al.* (2011) Gene expression profiling predicts the development of oral cancer. *Cancer Prev Res (Phila)*, **4**, 218-229.
- [81] Schemper, M. *et al.* (2009) The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine*, **28**, 2473-2489.
- [82] Song, X. and Zhou, X. (2008) A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistical Sinica*, 947-965.
- [83] Soofi, E. *et al.* (1995) Information Distinguishability with Application to Analysis of Failure Data. *Journal of the American Statistical Association*, **90(430)**, 657-668.
- [84] Spitzner, D. (2004) Construction Concepts for Continuum Regression. *Technical Report, Department of Statistics, Virginia Tech.*
- [85] Stute, W. (1993) Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, **45(1)**, 89-103.
- [86] Sundberg, R. (2002) Continuum Regression. *Encyclopedia of Statistical Science*, **2nd Ed.**
- [87] Storey, J. D. (2002) A direct approach to false discovery rates. *Royal Statistical Society*, **64(3)**, 479-498.
- [88] Tabatabai, M. A. *et al.* (2012) Clinical and multiple gene expression variables in survival analysis of breast cancer: Analysis with the hypertabastic survival model. *BMC Medical Genomics*, **6(63)**.
- [89] TCGA Research Network: <http://cancergenome.nih.gov/>

- [90] Tibshirani, R. (1997) The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.
- [91] Tothill, R.W. *et al.* (2008) Novel molecular subtypes of serous and endometroid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, **14**, 5198-5208.
- [92] Tusher, V. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, **98(9)**, 5116-5121.
- [93] van der Net, J. B. *et al.* (2008) Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. *European Journal of Human Genetics*, **16**, 1111-1116.
- [94] Verweij, P. and Houwelingen, H. (1994) Penalized likelihood in Cox regression. *Statistics in Medicine*, **13**, 2427-2436.
- [95] Wang, S. *et al.* (2008a) Doubly penalized Buckley- James method for survival data with high-dimensional covariates. *Biometrics*, **64**, 132-140.
- [96] Wang, Z. and Wang, C. Y. *et al.* (2010) Buckley-James boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetic Molecular Biology*, **9(24)**.
- [97] Wei, Z. and Li, H. (2007) Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*, **8**, 265-284.

- [98] van Wieringen, W. N. *et al.* (2009) Survival prediction using gene expression data: A review and comparison. *Computational Statistics and Data Analysis*, **53**, 1590-1603.
- [99] Witten, D. and Tibshirani, R. (2008) Testing significance of features by lassoed principal components. *The Annals of Applied Statistics*, **2(3)**, 986-1012.
- [100] Wold, S. *et al.* (2002) Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. *Encyclopedia of Computational Chemistry*.
- [101] Xu, C. *et al.* (2010) Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. *Molecular Cancer*, **9**, 143.
- [102] Xu, J. *et al.* (2005) Survival analysis of microarray expression data by transformation models. *Computational Biology and Chemistry*, **29**, 91-94.
- [103] Yang, S. and Prentice, R. (2005) Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika*, **92(1)**, 1-17.
- [104] Youden, W. J. (1950) Index for rating diagnostic tests. *Cancer*, **3:1**, 32-35.
- [105] Zhang, H. and Lu, W. *et al.* (2007) Adaptive-Lasso for Cox's Proportional Hazards Model. *Biometrika*, **94**, 691-703.

- [106] Zhang, W. *et al.* (2013) Network-based Survival Analysis Reveals Sub-network Signatures for Predicting Outcomes of Ovarian Cancer Treatment. *PLoS Comput Biol*, **9(3)**, e1002975.
- [107] Zou, H. (2008) A Note on Path-Based Variable Selection in the Penalized Proportional Hazards Model. *Biometrika*, **95**, 241-247.