Univariate gradient statistic for a marginal cure rate model with high-dimensional covariates

by

Jennifer Delzeit

B.S., Kansas State University, 2017

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Wei-Wen Hsu

Copyright

Abstract

Cure rate models, also known as two-component mixture models, have been well established and widely used in the literature for analyzing the lifetime data of long-term survivors. Owing to the advancement of genomic technology, it is now of interest to identify the significant genes or microarrays that are highly associated with the survival outcome under the cure rate model framework. The identification procedure using these genomic data will involve the technique of variable selection for high-dimensional covariates. However, the cure rate model requires the modeling of the cure fraction and the survival function of the uncured individuals, which inevitably leads to a more complicated variable selection process. In this paper, we propose a gradient-statistic-based variable selection method under a marginal representation of the cure rate model. This marginal model can produce interpretable covariate effects on the overall survival response by relating the marginal mean hazard rate to high-dimensional covariates directly while regarding the cure fraction as a nuisance parameter. A univariate gradient score is then used iteratively to determine significant covariates. Coupled with the use of a False Discovery Rate approach, the top-ranked list of covariates can be easily obtained by controlling the family-wise error rate. The proposed method is evaluated by extensive simulations and illustrated with an application of the TCGA breast cancer dataset which contains more than 400,000 microarrays.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cure rate models are being studied in the literature more extensively in the past decade due to the rise of data with long-term survivors (Chaves and Rodrigues, 2011; Baghestani et al., 2015; Bernhardt, 2016; Kim, 2017). Nevertheless, the cure rate model that is being studied comes with many challenges. A major challenge comes when fitting this model with high-dimensional covariates. Particularly, the modeling of the cure fraction becomes a problem as it doubles the number of variables in the model. Thus, modeling the cure fraction substantially adds to the computing time and further complicates any type of variable selection procedure. An additional challenge with the cure rate model is that there is no meaningful and straightforward interpretation that relates the effects of the covariates to the overall survival response. From a practical standpoint, this interpretation would be extremely useful and is often of interest to the researcher. In order to resolve the two preceding challenges, we propose a marginal cure rate model for high-dimensional covariates. Currently, there is no study in the literature that focuses on the marginal cure rate model because both the marginal survival function and hazard function are nonstandard due to the long-term survivors. Therefore, we consider a marginal mean hazard model developed by Chen as the marginal cure rate model in this report (Chen, 2019). Technically, this model is built upon the use of the average overall hazard rate and a Weibull baseline hazard function.

Not only are there no current studies over the marginal cure rate model, but in present-day literature there are only a few studies that focus on variable selection for high-dimensional variables under the cure rate model (Li, 2014; Fan et al., 2017; Masud et al., 2018). However, there are several variable selection methods that have been derived for the regular Cox Proportional Hazards Model, even though the Cox model does not have the capability to handle long-term survivors like the cure rate model does. Some of these methods under the Cox model are complex requiring a bunch of iterations or an inverse of the design matrix, such as Tibshirani's Least Absolute Shrinkage and Selection Operator termed Lasso (Tibshirani, 1997) or a stability selection method (Yin and Zhang, 2017). Others that simply want to identify covariates that are associated with survival, take a univariate approach. The classical approach to find these covariates involves testing the hypothesis that the covariate is not associated with survival against the hypothesis that the covariate is associated with survival. However, with high-dimensional data, there are too many hypotheses tested and multiple testing becomes a huge issue. One method that has been well-studied for high-dimensional data under the survival setting is the univariate Cox score (Witten and Tibshirani, 2010). The univariate Cox score is a straightforward method that involves calculating the score, the information matrix, the restricted maximum likelihood, and the regular maximum likelihood.

For the purpose of this report, we adopt a gradient score which is seldom discussed in the literature, but was originally derived from the regular score statistic (Terrell, 2002). Out of the aforementioned covariate selection procedures, the gradient score is the most easy to implement as it only requires the first partial derivatives of the model parameters and both the unrestricted and restricted maximum likelihood estimates. It does not require the use of the information matrix. For the reason of the simplicity of this gradient score, we derive this method under the proposed marginal cure rate model. We then pair our method with a false discovery rate algorithm to control the familywise error rate.

Upon the completion of deriving both the model and the gradient score, we proceed to evaluate the proposed method with a series of extensive simulations under different model

settings. Afterwards, we illustrate this method by applying it to the National Cancer Institute's The Cancer Genome Atlas (TGCA) breast cancer dataset that contains almost 400,000 microarrays and over 600 individuals. We then discuss our results and both the benefits and limitations of using the proposed method.

The remainder of this report is organized as follows: in Chapter 2 the derivation as well as the assumptions of the proposed marginal cure rate model will be discussed. In Chapter 3, we extend the Gradient test to our proposed model for high-dimensional variable selection. In Chapter 4, we evaluate our proposed method by extensive simulation studies and illustrate our method by an application to the TGCA breast cancer dataset. In Chapter 5, we discuss the usage of our proposed method as well as state some of the benefits and limitations of using it.

Chapter 2

The Marginal Cure Rate Model

In this chapter, we will very briefly discuss the origin of our chosen model, the marginal cure rate model. Please note this work was done by Jianfeng Chen as part of his Doctoral Thesis at Kansas State University (Chen, 2019).

2.1 Motivation of the Marginal Cure Rate Model

A cure rate model is different than other models as at the end of the study, some patients or some individuals are cured and are considered long-term survivors. This type of model is very important in survival analysis and has become more prominent in the literature in the past decade as more data have become available with long-term survivors. The standard cure rate model assumes there are two groups: an uncured group and a cured group, assuming U is the latent indicator variable that distinguishes these two groups. The uncured group can be thought of as the individuals who will experience the event of interest, $U_i = 1$ say, and the cured group are the individuals that will never experience this event, $U_i = 0$ where $i = 1, 2, ..., n$. The probability that an individual from the uncured group experiences the event of interest is π_i, $P(U_i = 1) = \pi_i$. This probability is often called the uncured fraction in the literature. Let t_i denote the censoring time or the time it takes for the individual to experience the event. Also, let $S_u(t_i)$ be the survival function for that of the uncured group.

Now, the marginal survival function for the overall population becomes:

$$S_M(t_i) = (1 - \pi_i) + \pi_i S_u(t_i), \qquad 0 \leq \pi_i \leq 1 \tag{2.1}$$

In this equation, the overall survival function $S_M(t_i)$ is confounded by the subpopulation of the uncured group's survival rate $S_u(t_i)$ and uncure fraction π_i. By assuming π_i and $S_u(t_i)$ share the same set of covariates, the additional modeling of the uncure fraction in the conditional cure rate model doubles the number of modeled parameters. These relationships and confounding variables make the interpretation between the effects of the covariates and the overall hazard rate extremely difficult.

Therefore, in order to better explain the effect that the covariates have on the hazard rate of the population and to eradicate using a uncure fraction, we propose a marginal mean hazard rate model. We first derive the marginal hazard rate model under the cure rate model (i.e., under the conditional model).

The marginal hazard rate can be described by the following:

$$h_M(t_i) = \frac{\pi_i f_u(t_i)}{1 - \pi_i + \pi_i S_u(t_i)}, \quad \text{where } f_u(t_i) = h_u(t_i) S_u(t_i) \text{and } h_u(t_i) = h_{u_0}(t_i) e^{\boldsymbol{\eta} \boldsymbol{w_i}} \tag{2.2}$$

In Equation 2.2, $h_u(t_i)$ is the hazard function of the uncured individuals and $\boldsymbol{\eta}$ is the vector of parameter coefficients for the covariates w_i for the i^{th} individual. The most popular baseline hazard function is the Weibull distribution as the Weibull distribution tends to be more flexible than other commonly used distributions like the exponential and lognormal due to having an additional scale parameter (Farewell, 1982). When the Weibull distribution is assumed, the baseline hazard becomes:

$$h_{u_0}(t_i) = \alpha \lambda t_i^{\alpha - 1} \qquad \text{with } \alpha > 0 \text{ and } \lambda > 0$$

2.2 Derivation of the Marginal Cure Rate Model

We derive the marginal cure rate model by considering the risk of an event on average, the expectation of the marginal hazard function with respect to time t. We then re-parameterize the model so that the marginal mean hazard rate and the effects of the covariates have the same support. The hazard rate needs to be a non-negative value, so we let the mean hazard rate be:

$$E[h_M(t_i)] = e^{\boldsymbol{\beta}' x_i} \tag{2.3}$$

In this case, $\boldsymbol{\beta}$ is the vector of parameter coefficients for the covariates and x_i are the observed covariates for the i^{th} individual. We can then derive the marginal mean hazard rate $E[h_M(t_i)]$ from the idea of a latent variable U_i by using:

$$h_M(t_i) = \begin{cases} h_u(t_i) & \text{when} \quad U_i = 1 \\ 0 & \text{when} \quad U_i = 0 \end{cases} \tag{2.4}$$

where U_i is the unobserved cure indicator. $U_i = 1$ indicates the individual is uncured with probability π, which is now a constant and does not depend on the individual (i.e., $\pi_i = \pi$ for all i). Conversely, $U_i = 0$ indicated the individual is cured with probability $1 - \pi$. Using the above, we can get the marginal mean hazard rate:

$$\begin{aligned} E[h_M(t_i)] &= E[E(h_M(t_i)|U_i)] \\ &= E[\pi h_u(t_i)] \\ &= \pi E[h_u(t_i)] \end{aligned} \tag{2.5}$$

Assuming the baseline hazard rate for the uncured group is derived from the Weibull survival function, $h_u(t_i) = \alpha\lambda t_i^{\alpha-1} e^{\mu_i}$ where $\mu_i = \boldsymbol{\eta w_i}$ from Equation 2.5, then Equation 2.3 can be

rewritten as:

$$\begin{aligned} e^{\boldsymbol{\beta' x_i}} &= \pi \int \alpha\lambda t_i^{\alpha-1} e^{\mu_i} f_u(t_i) dt \\ &= \pi\alpha\lambda e^{\mu_i} \int t_i^{\alpha-1} \lambda\alpha e^{\mu_i} t_i^{\alpha-1} \exp\{ -\lambda t_i^{\alpha} e^{\mu_i}\} dt \\ &= \pi\alpha\lambda e^{\mu_i} E[t_i^{\alpha-1}] \end{aligned} \tag{2.6}$$

We know that $E[t^k] = (\lambda e^{\mu})^{-\frac{k}{\alpha}}\Gamma(1+\frac{k}{\alpha})$ is the k^{th} moment of a Weibull random variable. Thus, we rewrite Equation 2.6 as:

$$e^{\boldsymbol{\beta' x_i}} = \pi\alpha[\lambda e^{\mu_i}]^{\frac{1}{\alpha}}\Gamma\left(2-\frac{1}{\alpha}\right) \tag{2.7}$$

Finally, we solve for λe^{μ_i}:

$$\lambda e^{\mu_i} = \left[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha} \tag{2.8}$$

Next, we will consider the likelihood function for the mixture data from both the risk and non-risk group. Suppose we observe, $\{t_i, \delta_i, \mathbf{x}_i\}$ for the i^{th} observation, then the likelihood function for the marginal cure rate model becomes:

$$\begin{aligned} \ell(\alpha, \boldsymbol{\beta}, \pi | t_1, ..., t_n) &= \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} [\pi f_u(t_i)]^{\delta_i} [1-\pi+\pi S_u(t_i)]^{1-\delta_i} \\ &= \prod_{\imath=1}^{n} \left\{ \pi\alpha \left[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha} t_i^{\alpha-1} \exp\left\{ -t_i^{\alpha}\left[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha}\right\}\right\}^{\delta_i} \\ &\qquad \left\{1-\pi+\pi\exp\left\{ -t_i^{\alpha}\left[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha}\right\}\right\}^{1-\delta_i} \end{aligned} \tag{2.9}$$

where π is the uncured rate, t_i is the survival time for the i^{th} individual, and δ_i is the censor indicator in which $\delta_i = 1$ if the individual is not censored and 0 otherwise.

Chapter 3

Variable Selection for High-Dimensional Covariates

3.1 The Gradient Score

In Chapter 2, the underlying model, the marginal cure rate model, was introduced. In this chapter, the proposed test for testing high-dimensional data under the marginal cure rate model will be proposed. In current literature, there are many variable selection methods for high-dimensional data under the Cox Proportional Hazards Model that have not been translated to the cure rate model. The Cox Proportional Model and the cure rate model can both be used for the analysis of survival data. Nevertheless, the Cox Proportional Model does not have the capability of handling long-term survivors. For the Cox model with high-dimensional data, one of the most straightforward methods for variable selection is the univariate Cox score, a score statistic (Witten and Tibshirani, 2010). There are other methods that have been proposed as well. Some of these methods include using a Wald score instead of the Cox score; however, the disadvantage with the Wald score is that in the high-dimensional setting, it requires iteratively fitting a Cox model for each covariate. Another method for variable selection includes the lassoed principal components method.

This method is primarily used for genomic data when researchers have a belief that a given gene or microarray is associated with the survival outcome if it is correlated with a set of other genes that already appear to be associated with the survival outcome (Witten and Tibshirani, 2008, 2010). The downside to this approach is that it requires the use of eigenvectors of the data matrix and there is an adaptively chosen tuning parameter. For the cure rate model, there are very few methods for high-dimensional variable selection discussed in the literature. The method proposed by Y. Li in their doctoral thesis involves a very complex expectation-maximization algorithm as well as other calculations such as the second partial derivatives (Li, 2014). Yet another proposed method uses an uncommon assumption of proportional relationships between the covariates and uncure fraction (Fan et al., 2017).

Due to the complexity of the aforementioned methods, we propose a variable selection method under the proposed marginal cure rate model that mimics the Cox Score. Technically, we propose a gradient score first derived by Terrel in 2002 (Terrell, 2002). The novelty of the gradient score is that it does not use the information matrix as does the Cox score. Therefore, when partial second derivatives of the model parameters become too complex as is the case with our marginalized cure rate model, the gradient score is a good, easy alternative. As defined by Terrell in 2002 and explained by Lemonte in 2012, the gradient score tests the null hypothesis, H_0:$\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$ against H_1:$\boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_{20}$ where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, $\boldsymbol{\theta}_1 = (\theta_1, ..., \theta_q)'$ represents the parameters that are not of interest or the nuisance parameters, $\boldsymbol{\theta}_2 = (\theta_{q+1}, ..., \theta_p)'$ represents the parameters that are of interest in testing, and $\boldsymbol{\theta}_{20}$ represents a fixed vector with $p - q$ dimensions (Terrell, 2002; Lemonte and Ferrari, 2012). The gradient score, then, for testing the null hypothesis, H_0:$\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$, is equated by:

$$F = \mathbf{U}(\tilde{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \tag{3.1}$$

In this equation, the restricted maximum likelihood estimate under the null hypothesis is

represented by $\hat{\boldsymbol{\theta}}$ and is partitioned into a vector containing the nuisance and of interest parameters, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2^T)'$. Similarly, the unrestricted or normal maximum likelihood estimate is represented by $\tilde{\boldsymbol{\theta}}$ and is again partitioned, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1', \tilde{\boldsymbol{\theta}}_2')'$. Finally, let the $\mathbf{U}(\boldsymbol{\theta})$ be the partitioned score function, $\mathbf{U}(\boldsymbol{\theta}) = \partial\ell/\partial\boldsymbol{\theta} = (\mathbf{U_1}(\boldsymbol{\theta})', \mathbf{U_2}(\boldsymbol{\theta})')'$. It is now worthy to note that Equation 3.1 can be even more simplified. Due to the constant nature of the nuisance parameters, the score function evaluated at the unrestricted maximum likelihood of those parameters, ie $\mathbf{U_1}(\tilde{\boldsymbol{\theta}}) = 0$. Thus, the gradient statistic, as written out by Lemonte in 2012, now becomes:

$$F = \mathbf{U_2}(\tilde{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{20}) \tag{3.2}$$

The gradient score has a central chi-square distribution with $p - q$ degrees of freedom.

3.2 The Gradient Score Under the Marginal Cure Rate Model

We can now use Equation 3.2 to derive the gradient test statistic for the marginalized cure rate model. Recall the marginal cure rate model, as shown in Equation 2.9 is:

$$\begin{aligned}\ell(\alpha, \boldsymbol{\beta}, \pi | t_1, ..., t_n) = \prod_{i=1}^{n} \Bigg\{ \pi\alpha \left[\frac{e^{\boldsymbol{\beta}' x_i}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha} t_i^{\alpha-1} \exp\left\{ -t_i^{\alpha}\left[\frac{e^{\boldsymbol{\beta}' x_i}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha} \right\} \Bigg\}^{\delta_i} \\ \left\{ 1 - \pi + \pi \exp\left\{ -t_i^{\alpha}\left[\frac{e^{\boldsymbol{\beta}' x_i}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha} \right\} \right\}^{1-\delta_i}\end{aligned} \tag{3.3}$$

Next, we take the natural logarithm of the marginal cure rate model (Equation 3.3). This allows us to be able to begin taking partial first derivatives with respect to our parameters and to calculate both the unrestricted and restricted maximum likelihood underneath the

null hypothesis. Taking the natural logarithm of both sides of our model yields:

$$\begin{aligned}\mathcal{L}(\alpha,\boldsymbol{\beta},\pi|t_1,...,t_n) &= \log[\ell(\alpha,\boldsymbol{\beta},\pi|t_1,...,t_n)] \\ &= \sum_{i=1}^{n}\delta_i\Bigg\{\log(\pi)+\log(\alpha)+\alpha\boldsymbol{\beta}'x_i-\alpha\log(\alpha)-\alpha\log(\pi) \\ &\quad -\alpha\log\left(\Gamma\Big(2-\frac{1}{\alpha}\Big)\right)+\alpha\log(t_i)-\log(t_i)-\left(\frac{t_i e^{\boldsymbol{\beta}'x_i}}{\alpha\pi\Gamma\big(2-\frac{1}{\alpha}\big)}\right)^{\alpha}\Bigg\} \\ &\quad +\sum_{i=1}^{n}(1-\delta_i)\Bigg\{\log\left(1-\pi+\pi\exp\Big\{-t_i^{\alpha}\Big[\frac{e^{\boldsymbol{\beta}'x_i}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\Big]^{\alpha}\Big\}\right)\Bigg\}\end{aligned} \tag{3.4}$$

Finally, we are able to calculate the partial first derivatives with respect to the three parameters, α, π, and $\boldsymbol{\beta}=(\beta_0,\beta_j)'$.

$$\begin{aligned}\frac{\partial\mathcal{L}(\alpha,\boldsymbol{\beta},\pi|t_1,...,t_n)}{\partial\alpha} &= \sum_{i=1}^{n}\delta_i\Bigg\{\frac{1}{\alpha}+\boldsymbol{\beta}'x_i+\log\left(\frac{t_i}{\alpha\pi\Gamma\big(2-\frac{1}{\alpha}\big)}\right)-G-N^{\alpha}\log(N)-G-2\Bigg\} \\ &\quad +\sum_{i=1}^{n}(1-\delta_i)\Bigg\{\frac{1}{D}\Big(\pi\exp\big\{-N^{\alpha}\big\}\big[-N^{\alpha}\big(\log(N)-1-G\big)\big]\Big)\Bigg\} \\ \frac{\partial\mathcal{L}(\alpha,\boldsymbol{\beta},\pi|t_1,...,t_n)}{\partial\pi} &= \sum_{i=1}^{n}\delta_i\Bigg\{\frac{1-\alpha}{\pi}+N^{\alpha}\Big(\alpha\pi^{-(\alpha+1)}\Big)\Bigg\} \\ &\quad +\sum_{i=1}^{n}(1-\delta_i)\Bigg\{\frac{1}{D}\Big[-1+\exp\big\{-N^{\alpha}\big\}+\pi\exp\big\{-N^{\alpha}\big\}\big(\pi N\big)^{\alpha}\alpha\pi^{-(\alpha+1)}\Big]\Bigg\} \\ \frac{\partial\mathcal{L}(\alpha,\boldsymbol{\beta},\pi|t_1,...,t_n)}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n}\delta_i\Big\{\alpha\mathrm{x}_i-\alpha x_i N^{\alpha}\Big\}+\sum_{i=1}^{n}(1-\delta_i)\Big\{\frac{-\alpha x_i N^{\alpha}}{D}\Big\}\end{aligned} \tag{3.5}$$

where

$$N = \frac{t_i e^{\boldsymbol{\beta}' x_i}}{\alpha\pi\Gamma\left(2-\frac{1}{\alpha}\right)}$$
$$G = \frac{1}{\alpha\Gamma(2-\frac{1}{\alpha})}\left(\frac{\partial}{\partial\alpha}\Gamma\left(2-\frac{1}{\alpha}\right)\right)$$
$$D = 1-\pi+\pi\exp\left\{-t_i^{\alpha}\left[\frac{e^{\boldsymbol{\beta}'\boldsymbol{x_i}}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha}\right\}$$

Using the partial derivatives in Equation 3.5, we can now begin calculating the score statistic under the marginal cure rate model. To finish calculating the gradient score, the maximum likelihood estimates and the restricted maximum likelihood estimates under the null hypothesis need to be calculated and put into Equation 3.2. From Equation 3.2 and using the hypotheses: $H_0 : \beta_j = 0$ and $H_1 : \beta_j \neq 0$ with $j = 1, ..., p$, the closed form of the gradient statistic is:

$$F_j = U(\tilde{\boldsymbol{\theta}}, \tilde{\beta}_j) \times \hat{\beta}_j, \qquad \text{where } U(\tilde{\boldsymbol{\theta}}, \tilde{\beta}_j) = \left.\frac{\partial\mathcal{L}(\alpha, \boldsymbol{\beta}, \pi|t_1, ..., t_n)}{\partial\beta_j}\right|_{\boldsymbol{\theta_1}=\tilde{\boldsymbol{\theta}}_{\mathbf{1}}, \beta_j=\tilde{\beta}_j=0} \tag{3.6}$$

Here $\boldsymbol{\theta_1} = (\alpha, \pi, \beta_0)'$, $\tilde{\boldsymbol{\theta}}_{\mathbf{1}} = (\tilde{\alpha}, \tilde{\pi}, \tilde{\beta}_0)'$ are the restricted maximum likelihood estimates for α, π, and β_0 evaluated at $\beta_j = 0$, and $\hat{\beta}_j$ is the maximum likelihood estimate of β_j.

The gradient score is treated much in the standard way a score statistic is treated; however, the gradient score is non-negative and therefore the interpretation only becomes if the feature is significant at predicting survival (Witten and Tibshirani, 2010).

3.3 Variable Selection with the Gradient Score

Once computed, the gradient test needs some way to tell which of the features (or covariates) are significantly associated with survival. From a classical perspective, the significance for each feature would be tested using a hypothesis test. The null hypothesis in this case would

be that there is no association of that feature with survival and our test hypothesis would be that there is an association of that feature with survival. Owing to high-dimensional features, we would then adjust the p-values using some form of multiple comparison adjustment such as Holm's method (Holm, 1979), Tukey's method (Tukey, 1991), or any other method. If these p-values were small enough, the feature would then be considered significantly associated with survival. Depending on the data, it may be appropriate to make this adjustment in different ways. Given high-dimensional genomic data, it can be found that adjustment of the p-values leads to meaningless and oftentimes little to no results (Witten and Tibshirani, 2010). This largely comes from the fact that there are so many features and not enough change within the decimal places of the p-values to find anything that is truly significant. In this scenario, it may be beneficial for the researcher to accept some features knowing that some of these features will be false positives as in the feature will show up significant, but will not actually be significant. In this case, a false discovery rate (FDR) is appropriate and can be very useful. The FDR was originally published by Benjamini in 1995 (Benjamini and Hochberg, 1995) and later in 2003, Storey applied the FDR to genomic studies (Storey and Tibshirani, 2003). Storey in 2003 found that the original version of the FDR was not powerful enough to handle ultra high-dimensional genomic studies, but found that a permutation version of the False Discovery Rate could deal with such a limitation. The algorithm for the permutated FDR used by Storey was written out in Witten's work in 2010 (Storey and Tibshirani, 2003; Witten and Tibshirani, 2010). We adapted the algorithm used by both Storey and Witten to the gradient statistic under the marginal cure rate model and is written out below.

1. Compute the gradient statistic for each feature, $j = 1, 2, ..., p$, denoted by F_j.

2. For $i \in \{1, 2, ..., M\}$ where M = 1000 or a large enough number

   (a) Permute the individual's survival times t_i, i.e., randomly assign the survival times to the feature measurements, x_j

(b) Compute the gradient statistic for the permuted data; F_j^{*i}.

3. Estimate the FDR at a pre-determined threshold value, c by:

$$\widehat{FDR(c)} = \frac{\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{p}\mathbb{1}(F_j^{*i} \geq c)}{\sum_{j=1}^{p}\mathbb{1}(F_j \geq c)} \tag{3.7}$$

In Equation 3.7, the $\mathbb{1}(\cdot)$ is an indicator variable and the numerator is the average or expected number of features that exceed the threshold value under the null hypothesis where the denominator is the actual or observed number of features that exceed that threshold value. Please note that Equation 3.7 was originally written as an algorithm for calculating the FDR given the Cox Score Test was being calculated (Witten and Tibshirani, 2010). For this report, this algorithm was slightly changed to incorporate estimating the false discovery rate using the gradient statistic.

Chapter 4

Results

In this chapter we will discuss the results from both our rigorous simulation analyses and from an application to the TGCA breast cancer dataset. First, we will discuss the simulation analyses and then we will discuss the real data analysis.

4.1 Simulation Results

First, we simulated data from our model under two different scenarios. One scenario, we included 200 individuals ($n = 200$) with 100 features ($p = 100$). The second scenario, we included 200 individuals, but 300 features ($p = 300$), to mimic a more typical high-dimensional data structure. With these two scenarios, we included a different number of truly significant features $p^* \in \{5, 10, 15\}$. In order to obtain data simulated from our model, we had to define some parameters for the model. We defined Σ, the covariance matrix as $\Sigma_{i,j} = \rho^{|i-j|}$ where $i = 1, 2, ..., p$; $j = 1, 2, ..., p$; and ρ is the correlation between the features. For this study, we looked at three different values of the correlation, $\rho \in \{0, 0.2, 0.5\}$. We set the value of α in the marginal cure rate model to 1.1 and the uncure rate π to 0.85. The measurement for each feature was found from a multivariate normal distribution with a mean of 0 and the variance of Σ as discussed above.

The true measurements, $\boldsymbol{\beta}^*$, for each feature were then supplied. The measurements

depended on how many truly significant features there were $p^* \in \{5, 10, 15\}$.

When $p^* = 15$,

$$\boldsymbol{\beta^*} = (\underbrace{1.75, 1.5, -1.5, 1.75, -1.5, 1.5, 1.75, 1.5, -1.75, -1.5, 1.25, -1.75, 2, 1.5, -1.5, 1.75}_{\text{intercept + 15 significant features}}, \underbrace{0, 0, ..., 0, 0}_{\text{100-16 or 300-16}})'$$

When $p^* = 10$,

$$\boldsymbol{\beta^*} = (\underbrace{1.75, 1.5, -1.5, 1.75, -1.5, 1.5, 1.75, 1.5, -1.75, -1.5, 1.25}_{\text{intercept + 10 signficant features}}, \underbrace{0, 0, ..., 0, 0}_{\text{100-11 or 300-11}})'$$

When $p^* = 5$,

$$\boldsymbol{\beta^*} = (\underbrace{1.75, 1.5, -1.5, 1.75, -1.5, 1.5}_{\text{intercept + 5 significant features}}, \underbrace{0, 0, ..., 0, 0}_{\text{100-6 or 300-6}})'$$

In all three scenarios, the intercept β_0^* was set to 1.75 and was not included in the count of truly significant features. Next, the true marginal mean hazard rate was found by taking only the non-zero values of $\boldsymbol{\beta}$ and multiplying them by the measurement simulated by the multivariate normal distribution. The marginal mean hazard rate was then estimated by $\widehat{E[h_M(t_i)]} = e^{\hat{\boldsymbol{\beta}} x_i}$ where $\hat{\boldsymbol{\beta}}$ are the estimated parameter coefficients for the covariates x_i observed for the i^{th} individual.

The censoring rate was determined using a randomized exponential distribution with rate of $\lambda_c \in \{0.0002, 0.001, 0.002\}$. The censoring time was set to have a maximum of 300 days for each individual. Next, we got a true survival time t_i by taking a randomized binomial distribution with probability set to the uncure rate π. If the randomized binomial distribution gave a value of 1, the individual was assigned a true survival time t_i to be uncured by using:

$$\left[-\log(1-r) \left(\frac{\alpha \pi \Gamma\left(2 - \frac{1}{\alpha}\right)}{e^{\hat{\boldsymbol{\beta}}' x_i}} \right) \right]^{\frac{1}{\alpha}}$$

where r is a random uniform distribution taking values in $(0, 1)$, $\alpha = 1.1$, and $\pi = 0.85$.

Conversely, if the randomized binomial distribution gave a value of 0, the individual was assigned a true survival time t_i to be cured and this time was set to a large value, $t_i = 10000000$. Finally, we compared each individual's censoring time and true survival time. If the censoring time was equal to or exceeded the true survival time, a value of 1 was recorded for the censor, i.e., $\delta_i = 1$, otherwise a value of 0 was recorded $\delta_i = 0$.

With the data being generated from the proposed marginal cure rate model, we then calculated the gradient statistic from Equation 3.2 and used the regular False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). We took 1000 repetitions of these data and averaged the false positive rate (FPR), false negative rate (FNR), and accuracy. The results for nominal levels of 0.05 and 0.01 are shown in Tables 4.1 and 4.2, respectively.

From Tables 4.1 and 4.2, we show what happens as we not only increase the correlation ρ, but also what happens if the censoring coefficient λ_c, the number of truly significant features p^*, and the number of features p, in the dataset are changed. Also from Tables 4.1 and 4.2, we see that as we increase the censoring rate, there is little to no change in the FPR, the FNR, and the accuracy. This result is seen when we look at a unique combination of number of truly significant features p^*, number of features in the dataset p, correlation value ρ, and nominal level. For example, at $p = 100$, $p^* = 5$, $\rho = 0$, and nominal level of 0.01, we see that the FPR is 0.009, the FNR is near 0.125, and the accuracy is around 0.985 at all levels of censoring. These statistics may fluctuate a bit due to randomness and the random way in which these data are generated. Now, looking at a unique combination of p^*, λ_c, p, and nominal level, we see that majority of the time, as the correlation between features increases, the FPR decreases, the FNR increases, and the accuracy stays about the same. The intrigue behind this result may be due to the fact that the covariance structure that incorporates the correlation between the features is not taken into consideration in the gradient test. Thus, as the correlation increases, the algorithm is not finding enough features significant due to the correlation between the neighboring covariates. Thus, taking away from the false positive rate and adding to the false negative rate, but overall having the same accuracy

with this increase and decrease. Next, increasing the number of features from $p = 100$ to a high-dimensional case of $p = 300$ remaining at a constant p^*, ρ, λ_c, and nominal level; we see that the FPR decreases and both the FNR and accuracy increase. This result of a decrease in the FPR and an increase in both the FNR and the accuracy is also seen when looking at the effect of changing the nominal level from 0.05 to 0.01.

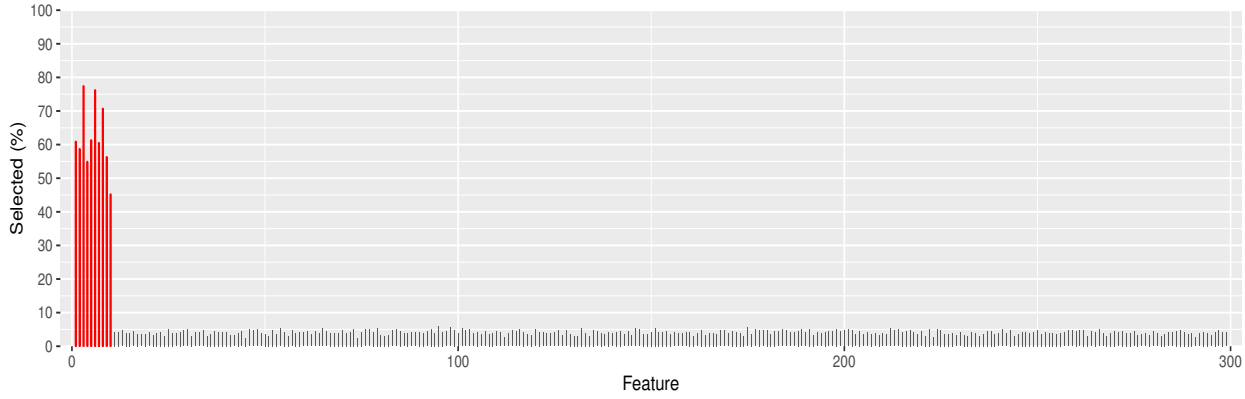
Table 4.1: Simulation Results at nominal level 0.05

		$n = 200, p = 100$			$n = 200, p = 300$		
	ρ	FPR	FNR	$Accuracy$	FPR	FNR	$Accuracy$
$p^* = 5$				Mild Censoring ($\lambda_c = 0.0002$)			
	0	0.027	0.071	0.971	0.017	0.097	0.982
	0.2	0.019	0.200	0.972	0.010	0.254	0.986
	0.5	0.009	0.511	0.966	0.004	0.581	0.987
				Moderate Censoring ($\lambda_c = 0.001$)			
	0	0.028	0.068	0.970	0.017	0.095	0.982
	0.2	0.020	0.197	0.972	0.010	0.257	0.986
	0.5	0.009	0.520	0.967	0.004	0.579	0.987
				Heavy Censoring ($\lambda_c = 0.002$)			
	0	0.029	0.066	0.969	0.018	0.089	0.981
	0.2	0.018	0.192	0.974	0.010	0.257	0.986
	0.5	0.009	0.530	0.965	0.003	0.585	0.987
$p^* = 10$				Mild Censoring ($\lambda_c = 0.0002$)			
	0	0.059	0.324	0.914	0.041	0.378	0.948
	0.2	0.052	0.400	0.913	0.034	0.457	0.952
	0.5	0.042	0.473	0.914	0.026	0.522	0.958
				Moderate Censoring ($\lambda_c = 0.001$)			
	0	0.064	0.308	0.912	0.043	0.370	0.946
	0.2	0.056	0.398	0.909	0.037	0.452	0.950
	0.5	0.042	0.472	0.915	0.027	0.511	0.957
				Heavy Censoring ($\lambda_c = 0.002$)			
	0	0.067	0.307	0.909	0.047	0.354	0.943
	0.2	0.058	0.392	0.909	0.041	0.435	0.946
	0.5	0.044	0.466	0.913	0.029	0.507	0.955
$p^* = 15$				Mild Censoring ($\lambda_c = 0.0002$)			
	0	0.084	0.503	0.853	0.065	0.561	0.911
	0.2	0.081	0.510	0.854	0.062	0.560	0.913
	0.5	0.081	0.484	0.858	0.054	0.531	0.923
				Moderate Censoring ($\lambda_c = 0.001$)			
	0	0.085	0.498	0.853	0.070	0.545	0.906
	0.2	0.086	0.504	0.851	0.064	0.546	0.912
	0.5	0.084	0.483	0.855	0.058	0.522	0.919
				Heavy Censoring ($\lambda_c = 0.002$)			
	0	0.090	0.491	0.849	0.076	0.528	0.901
	0.2	0.094	0.500	0.845	0.071	0.532	0.906
	0.5	0.084	0.475	0.857	0.062	0.517	0.915

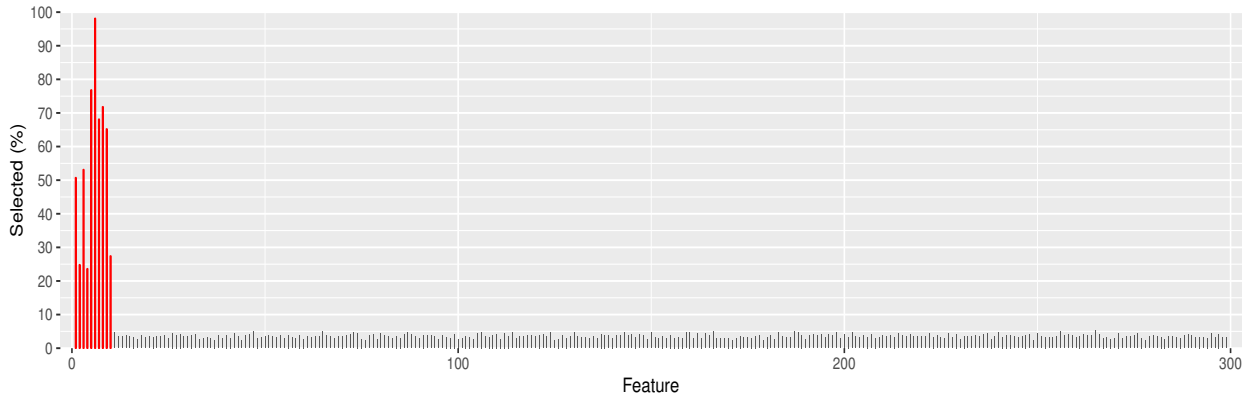
Table 4.2: Simulation Results at nominal level 0.01

		$n = 200, p = 100$			$n = 200, p = 300$		
	ρ	FPR	FNR	$Accuracy$	FPR	FNR	$Accuracy$
$p^* = 5$				Mild Censoring ($\lambda_c = 0.0002$)			
	0	0.009	0.127	0.985	0.005	0.173	0.993
	0.2	0.005	0.318	0.979	0.002	0.374	0.991
	0.5	0.002	0.612	0.667	0.001	0.670	0.988
				Moderate Censoring ($\lambda_c = 0.001$)			
	0	0.009	0.125	0.985	0.005	0.165	0.993
	0.2	0.005	0.312	0.979	0.002	0.403	0.991
	0.5	0.002	0.620	0.967	0.001	0.678	0.988
				Heavy Censoring ($\lambda_c = 0.002$)			
	0	0.009	0.123	0.986	0.005	0.158	0.993
	0.2	0.005	0.313	0.979	0.003	0.388	0.991
	0.5	0.002	0.637	0.966	0.001	0.677	0.988
$p^* = 10$				Mild Censoring ($\lambda_c = 0.0002$)			
	0	0.021	0.474	0.934	0.013	0.545	0.970
	0.2	0.018	0.533	0.930	0.011	0.587	0.970
	0.5	0.013	0.565	0.932	0.007	0.607	0.973
				Moderate Censoring ($\lambda_c = 0.001$)			
	0	0.022	0.469	0.933	0.015	0.527	0.968
	0.2	0.018	0.531	0.930	0.012	0.582	0.969
	0.5	0.013	0.571	0.930	0.008	0.607	0.972
				Heavy Censoring ($\lambda_c = 0.002$)			
	0	0.022	0.459	0.934	0.014	0.528	0.969
	0.2	0.020	0.516	0.930	0.012	0.567	0.970
	0.5	0.014	0.564	0.930	0.008	0.597	0.972
$p^* = 15$				Mild Censoring ($\lambda_c = 0.0002$)			
	0	0.031	0.660	0.874	0.022	0.713	0.943
	0.2	0.030	0.647	0.877	0.021	0.699	0.945
	0.5	0.031	0.591	0.884	0.018	0.635	0.951
				Moderate Censoring ($\lambda_c = 0.001$)			
	0	0.032	0.651	0.874	0.024	0.705	0.942
	0.2	0.032	0.641	0.876	0.022	0.686	0.945
	0.5	0.033	0.590	0.883	0.020	0.629	0.950
				Heavy Censoring ($\lambda_c = 0.002$)			
	0	0.035	0.647	0.872	0.026	0.684	0.941
	0.2	0.035	0.635	0.874	0.024	0.678	0.943
	0.5	0.033	0.579	0.885	0.021	0.620	0.949

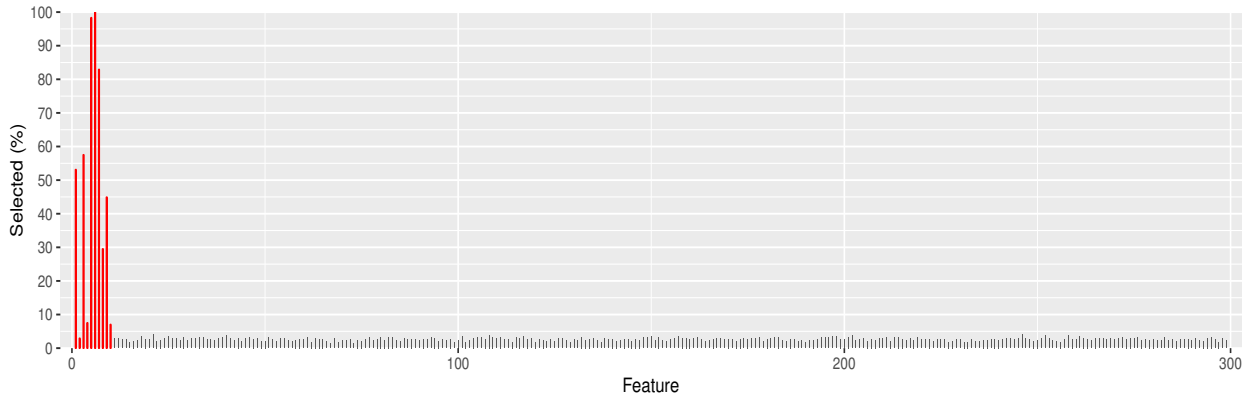
Afterwards, we looked at the percentage of times each feature was selected. Results are shown for $p^* = 10$, $n = 200$, and $p = 300$ in Figures 4.1 and 4.2. Figure 4.1 shows the results when the nominal level was 0.05. The increasing correlations, $\rho = 0$, $\rho = 0.2$, and $\rho = 0.5$ are shown in Figures 4.1a, 4.1b, and 4.1c, respectively. Conversely, Figure 4.2 shows the results with a nominal level of 0.01 and Figures 4.2a, 4.2b, and 4.2c show $\rho = 0$, $\rho = 0.2$, and $\rho = 0.5$, respectively. The red lines in each of these six figures represent the selected percentage of the truly significant features.



(a) Selected features: $\rho = 0$ and nominal level of 0.05

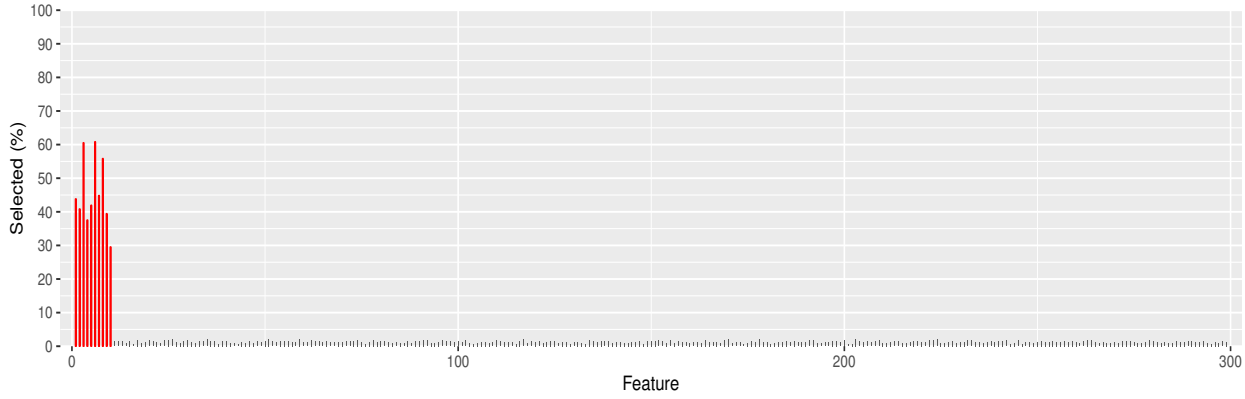


(b) Selected features: $\rho = 0.2$ and nominal level of 0.05

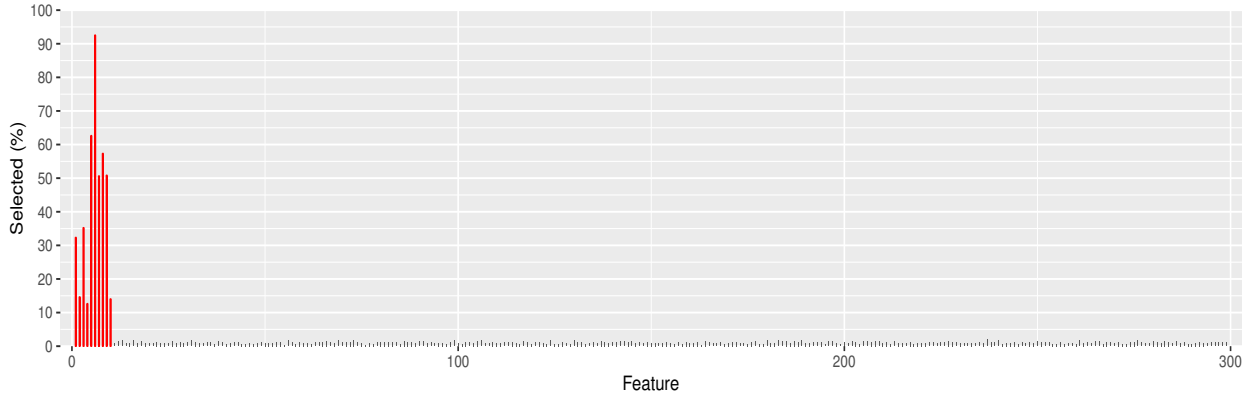


(c) Selected features: $\rho = 0.5$ and nominal level of 0.05

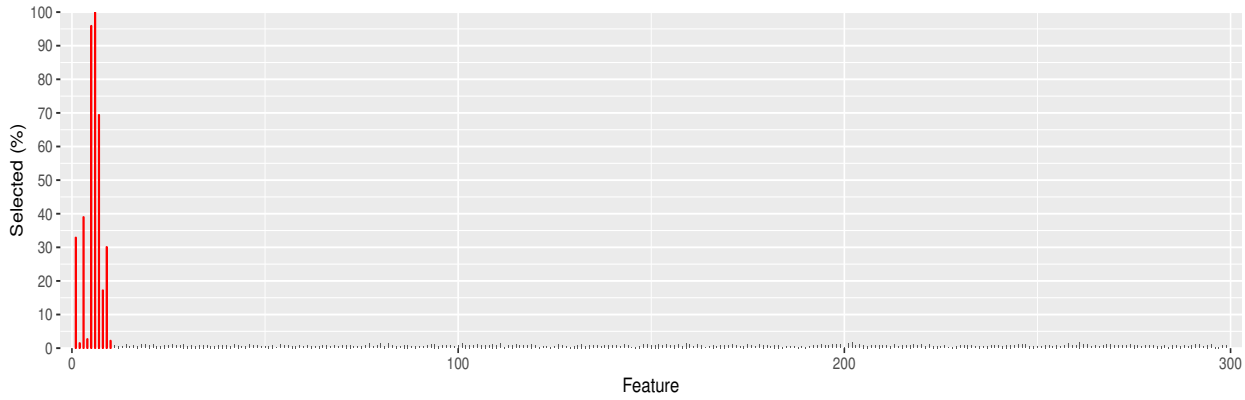
Figure 4.1: The preceding figures had 10 truly significant features $p^* = 10$ and censor rate $\lambda_c = 0.0002$ with a sample size $n = 200$ and $p = 300$ features



(a) Selected features: $\rho = 0$ and nominal level of 0.01



(b) Selected features: $\rho = 0.2$ and nominal level of 0.01



(c) Selected features: $\rho = 0.5$ and nominal level of 0.01

Figure 4.2: The preceding figures had 10 truly significant features $p^* = 10$ and censor rate $\lambda_c = 0.0002$ with a sample size $n = 200$ and $p = 300$ features

In Figure 4.1a, all 10 truly significant features, as shown by the red lines are all selected over 45% of the time; whereas the non-significant features are selected around 5% of the time. As the correlation increases to $\rho = 0.2$ in Figure 4.1b, the sixth truly significant feature is almost selected 100% of the time; whereas, three truly significant features drop to being selected less than 30% of the time. These three features continue to be selected even fewer times when the correlation is increased to 0.5 in Figure 4.1c. These three truly significant features drop to being selected under 10% of the time, almost mimicking that of the non-significant features. At the same time, there are three truly significant features that increase to being selected almost 100% of the time when the correlation is increased to 0.5. Similarly, these patterns are found in Figures 4.2a, 4.2b, and 4.2c. These figures all show that as the correlation increases, the significant features that are in the middle - the fourth, fifth, and sixth feature - all get selected nearly 100% of the time; whereas the features that are further away from the middle get selected less and less. This pattern is due to the value that the significant features were originally set to. Recall that when $p^* = 10$, the first 10 truly significant parameter values without the intercept in $\boldsymbol{\beta}^*$ were $\boldsymbol{\beta}^* = (1.5, -1.5, 1.75, -1.5, 1.5, 1.75, 1.5, -1.75, -1.5, 1.25)'$. In the figures, when the feature takes on a negative value, as in the second feature, the parameter is selected fewer times as the correlation levels increase between the features. When there are two negatively set features in a row, the second negative cancels out the first negative and is selected more times even if it has a negative value. This is seen in the ninth feature. Since there are two negatives in a row, the next feature, the tenth feature, although positive is selected less due to the preceding negative features (Fan and Lv, 2008; Goh and Dey, 2019).

4.2 Ultra High-Dimensional Simulation Results

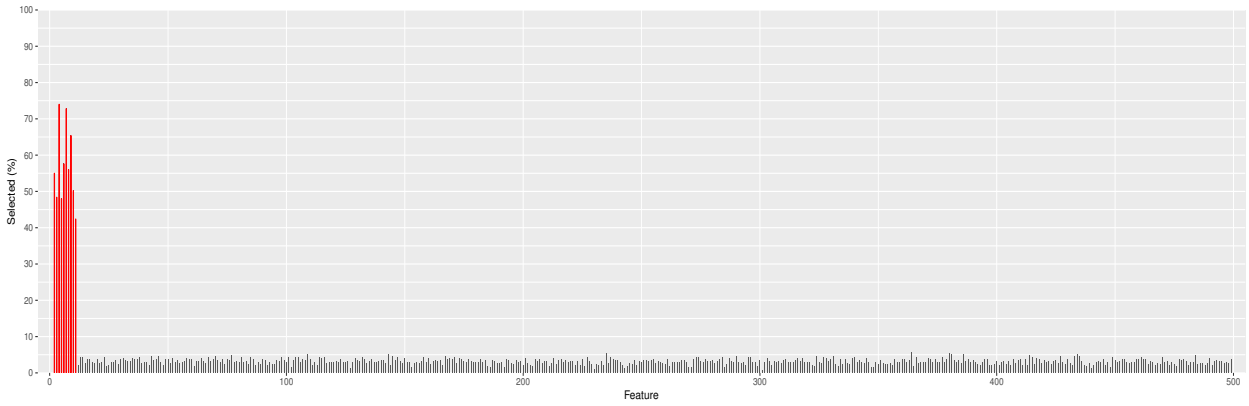
Using the methods described in Section 4.1, we now simulate data from an ultra high-dimensional setting using 200 individuals ($n = 200$) and 1000 covariates ($p = 1000$). The

results for the nominal level at 0.05 are seen in Table 4.3.

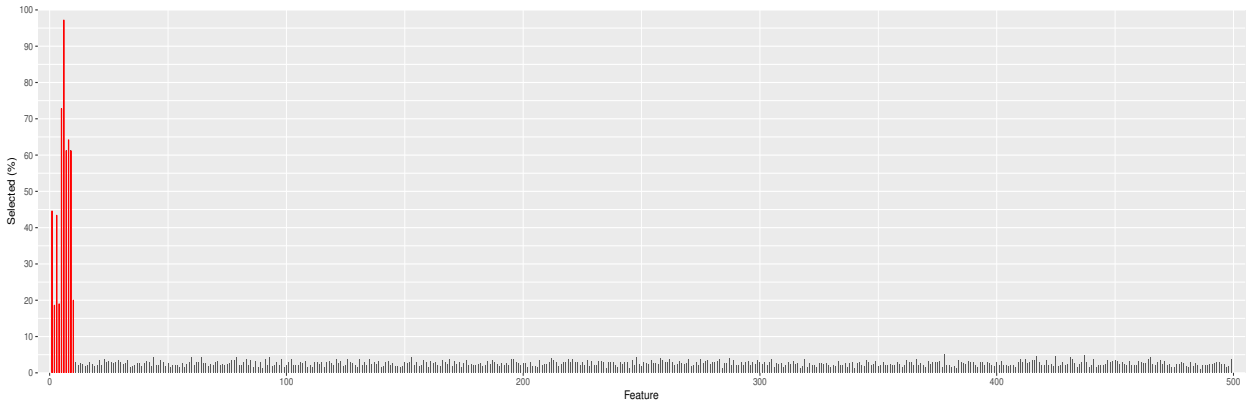
Table 4.3: Ultra high-dimensional results at nominal level 0.05

		$n = 200, p = 1000$		
	ρ	FPR	FNR	$Accuracy$
$p^* = 5$		Mild Censoring ($\lambda_c = 0.0002$)		
	0	0.010	0.122	0.989
	0.2	0.005	0.308	0.993
	0.5	0.001	0.650	0.996
		Moderate Censoring ($\lambda_c = 0.001$)		
	0	0.011	0.124	0.989
	0.2	0.005	0.465	0.997
	0.5	0.002	0.631	0.995
		Heavy Censoring ($\lambda_c = 0.002$)		
	0	0.012	0.114	0.988
	0.2	0.005	0.328	0.993
	0.5	0.001	0.650	0.995
$p^* = 10$		Mild Censoring ($\lambda_c = 0.0002$)		
	0	0.031	0.430	0.965
	0.2	0.026	0.498	0.970
	0.5	0.020	0.548	0.975
		Moderate Censoring ($\lambda_c = 0.001$)		
	0	0.035	0.410	0.962
	0.2	0.030	0.486	0.966
	0.5	0.021	0.544	0.974
		Heavy Censoring ($\lambda_c = 0.002$)		
	0	0.035	0.410	0.961
	0.2	0.030	0.469	0.965
	0.5	0.020	0.535	0.974
$p^* = 15$		Mild Censoring ($\lambda_c = 0.0002$)		
	0	0.058	0.574	0.935
	0.2	0.049	0.588	0.943
	0.5	0.044	0.548	0.948
		Moderate Censoring ($\lambda_c = 0.001$)		
	0	0.061	0.569	0.931
	0.2	0.057	0.579	0.935
	0.5	0.033	0.545	0.947
		Heavy Censoring ($\lambda_c = 0.002$)		
	0	0.067	0.550	0.926
	0.2	0.061	0.555	0.931
	0.5	0.049	0.533	0.944

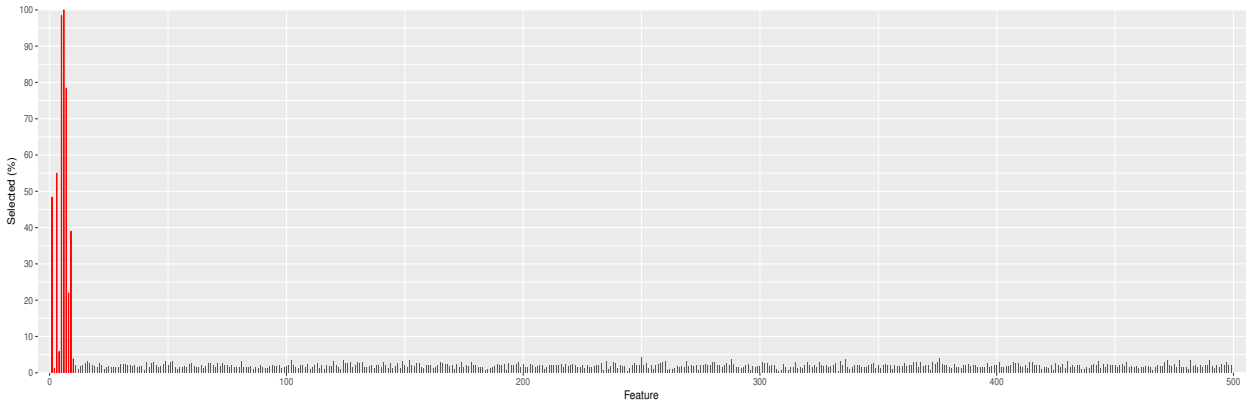
Next, we looked at the percentage of times each variable was selected. The results for the nominal level of 0.05 are shown in Figure 4.3. Please note, for this figure, only the first 500 variables are shown.



(a) Selected features: $\rho = 0$ and nominal level of 0.05



(b) Selected features: $\rho = 0.2$ and nominal level of 0.05



(c) Selected features: $\rho = 0.5$ and nominal level of 0.05

Figure 4.3: The preceding figures had 10 truly significant features $p^* = 10$ and censor rate $\lambda_c = 0.0002$ with a sample size $n = 200$ and $p = 1000$ features

Again, we see similar results as discussed in Section 4.1.

4.3 TCGA Breast Cancer Dataset Results

The Cancer Genome Atlas (TCGA) Breast Cancer Dataset has nearly 400,000 microarrays and 622 individuals in the study at the present time. Looking at the survival curve in Figure 4.4 from these data, we can see that approximately 60% of these individuals are long-term survivors. Thus, a cure rate model is needed. However, with the number of microarrays collected on each of these individuals, the cure rate model would be too complex and too difficult to interpret for a meaningful conclusion. Therefore, the marginal cure rate model is an appropriate option.

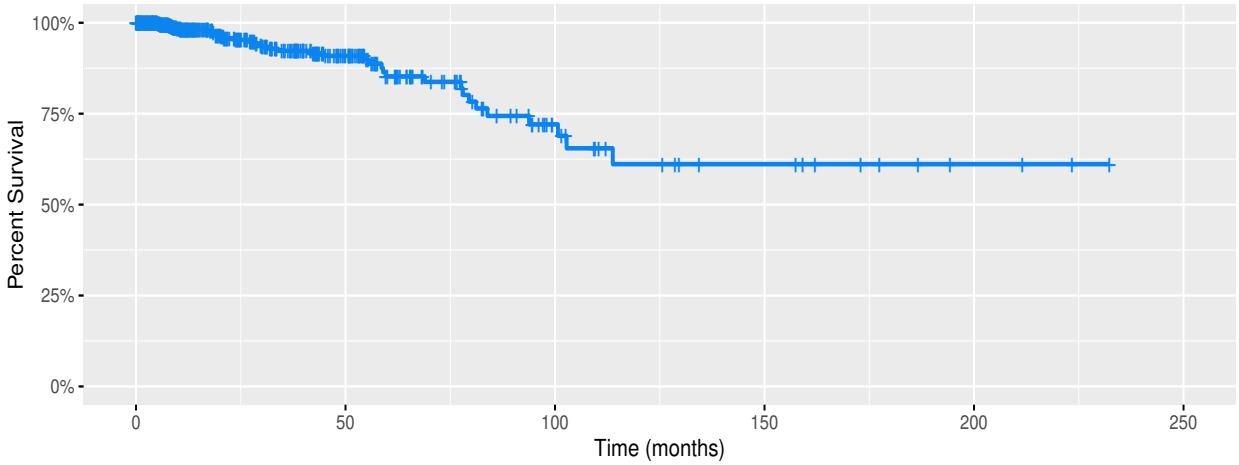


Figure 4.4: Survival curve of the TGCA breast cancer dataset

For the purpose of this report, only 4,000 microarrays were randomly chosen to analyze. Due to this random selection, it is possible that there are significant genes that are truly associated with breast cancer survival not in this selection. Of these 4,000 microarrays, we were able to find some significant microarrays by the proposed gradient test. From these 4,000 microarrays, we calculated the pairwise spearman correlation coefficients and found that a majority of the microarray pairs had very small correlations being around 0. These calculated pairwise correlations can be found in Figure 4.5.

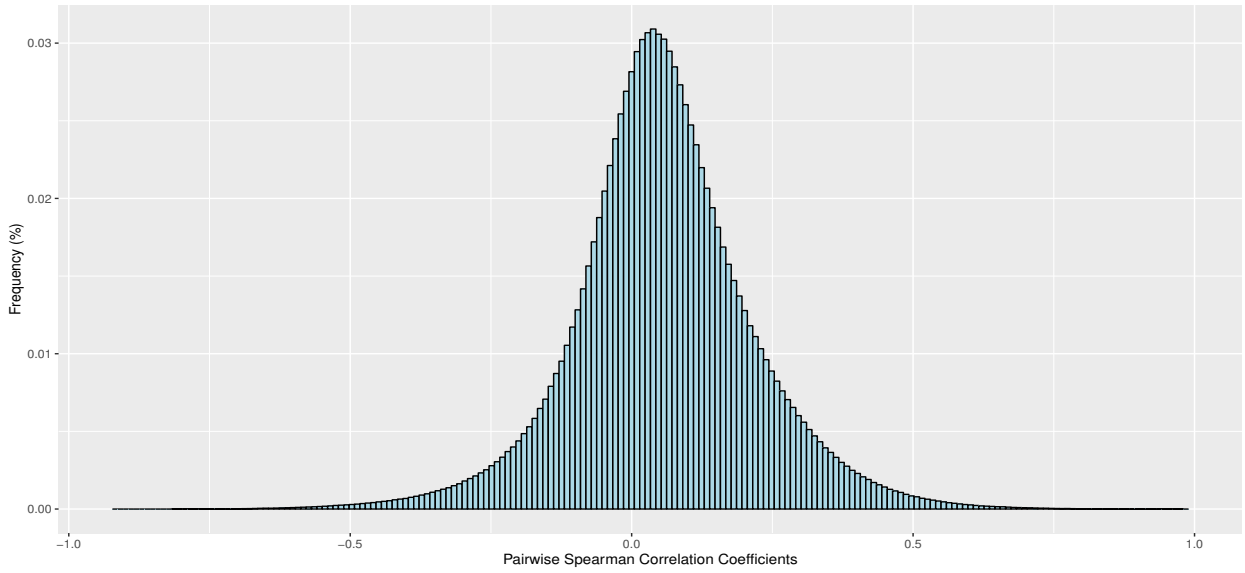


Figure 4.5: Pairwise Spearman correlation coefficients for the 4,000 microarrays in the TGCA dataset

The false discovery rate, as discussed in Section 3.3 was then calculated by using different threshold values, $c \in \{7.0, 7.5, 8.0, 8.5\}$, from Equation 3.7. 1,000 permutations were used in calculating these false discovery rates. Table 4.4 shows the selected microarrays along with the threshold value and calculated false discovery rate.

In Table 4.4, the selected microarrays when the threshold value is 8.5, are seen in all other results by different threshold values. We can also see the genes that each of these microarrays corresponds to. Interestingly, some of the selected microarrays and their genes are already known to be involved in important functions of the human cancer cell. *GNA12* is involved early on in the cancer cell's ability to spread throughout the body and to also invade healthy tissue (Rasheed et al., 2015). *GNG7* is involved in the way that the cancer cell can avoid natural cell death and therefore the cancer cell can live for a much longer time than a normal cell can (Liu et al., 2016). The estimated false discovery rates among these four results range from 0.076 to 0.097 which is in the acceptable range of false discovery rates for gene expression data 0.1 to 0.2 (Witten and Tibshirani, 2010). Nevertheless, during these calculations, some gradient scores that were not non-negative were noted. These scores were usually quite small and very close to zero, ranging from approximately -0.1 to -1.0.

Threshold Value c	7.0	7.5	8.0	8.5	Gene
Selected Microarrays					
	cg03961401	*cg03961401*	*cg03961401*	*cg03961401*	*GNA12*
	cg07739686	*cg07739686*	*cg07739686*	*cg07739686*	*TBC1D17*
	cg08933276	*cg08933276*	*cg08933276*	*cg08933276*	*GNA12*
	cg13701180	*cg13701180*	*cg13701180*	*cg13701180*	*GNG7*
	cg17290636	*cg17290636*	*cg17290636*	*cg17290636*	*TBCD*
	cg17579154	*cg17579154*	*cg17579154*	*cg17579154*	*TBCD*
	cg01205019	*cg01205019*	*cg01205019*		
	cg03303325	*cg03303325*	*cg05394010*		
	cg05394010	*cg05394010*	*cg13105522*		
	cg06544239	*cg06544239*			
	cg11021321	*cg13105522*			
	cg13105522				
	cg13573115				
	cg18639956				
	cg23866381				
	cg24502901				
	cg27023595				
$\widehat{FDR}$	0.094	0.097	0.089	0.076	

Table 4.4: The selected microarrays and estimated False Discovery Rate at each predetermined threshold value c for a subset of the TGCA breast cancer dataset

This is a problem as all gradient scores should be non-negative as was stated in Chapter 3.2. This problem is probably due to the fact that when calculating the maximum likelihood for α which has a support from $[0.5, \infty]$ was always near the 0.5 boundary. Since it was always near the end of its support boundary, the maximum likelihood algorithm could probably not get the true maximum likelihood estimate for α as there was that constraint put on the algorithm. Thus, some microarrays had small negative gradient scores that were close to zero and were considered to be zero.

Chapter 5

Discussion

In conclusion, we were able to derive the marginal cure rate model and a method of variable selection under our proposed model and receive meaningful results. The gradient statistic offers a simple and straightforward method for variable selection under the marginal cure rate model. This variable selection step is important as it gives a list of variables that is much smaller than the original set. In this study, we took a random portion of a dataset, 4,000 variables and were able to select 6-16 potentially important genes from it. The completion of this simple and easy to calculate variable selection step, simplifies the dataset and opens up many more opportunities for present-day data analyses.

Nevertheless, there are some limitations to both the marginal cure rate model and the gradient statistic. First, the marginal cure rate model assumes a Weibull distribution as the baseline hazard model. With the use of the Weibull distribution, we were able to have closed-form solutions for our log-likelihood which is more easily implemented and explained. For a more general model fitting, a nonparametric approach should be used for the baseline hazard model. Second, the univariat gradient statistic in this paper worked well for variable selection when there was little to no correlation between the features. However, there could be the case that there is substantial correlation between the features. In this scenario, the proposed variable selection method needs to incorporate the correlation structure into

its algorithm. Thus, future research should aim to add a correlation component into the gradient scores to better control for features that are potentially highly correlated.

Bibliography

A. Baghestani, M. Akbari F. Zayeri, L. Shojaee, N. Khadembashi, and P. Shahmirzalou. Fitting cure rate model to breast cancer data of cancer research center. *APJCP*, 16: 7923–7927, 2015.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57:289–300, 1995.

P. W. Bernhardt. A flexible cure rate model with dependent censoring and a known cure threshold. *Stat. Med.*, 35:4607–4623, 2016.

J. Chaves and J. Rodrigues. Standard exponential cure rate model with noninformative of informative uniform-exponential censoring. *Commun. Stat. Simul. Comput.*, 40:364–382, 2011.

J. Chen. *Marginal cure rate models for long-term survivors.* PhD thesis, Kansas State University, 2019.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B*, 70:849–911, 2008.

X. Fan, M. Liu, K. Fang, Y. Huang, and S. Ma. Promoting structural effects of covariates in the cure rate model with penalization. *Stat Methods Med Res*, 26:2078–2092, 2017.

V. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046, 1982.

G. Goh and D. Dey. Asymptotic properties of marginal least-square estimator for ultrahigh-dimensional linear regression models with correlated errors. *Am Stat.*, 73:4–9, 2019.

S. Holm. A simple sequentially rejective multiple test procedure. *Scand J Statist*, 6:65–70, 1979.

Y. Kim. Cure rate model with bivariate interval censored data. *Commun. Stat. Simul. Comput.*, 46:7116–7124, 2017.

A. Lemonte and S. Ferrari. The local power of the gradient test. *Ann. Inst. Stat. Math*, 64: 373–381, 2012.

Yanming Li. *High-dimensional variable selection for multivariate and survival data with applications to brain imaging and genetic association studies*. PhD thesis, 2014.

J. Liu, X. Ji, Z. Li, X. Yang, W. Wang, and X. Zhang. G protein gamma subunit 7 induces autophagy and inhibits cell division. *Oncotarget*, 7:24832–24847, 2016.

A. Masud, W. Tu, and Z. Yu. Variable selection for mixture and promotion time cure rate models. *Stat Methods Med Res*, 27:2185–2199, 2018.

S. Rasheed, C. Rong Teo, E. Beillard, P. Voorhoeve, W. Zhou, S. Ghosh, and P. Casey. Microrna-31 controls g protein alpha-13 (gna13) expression and cell invasion in breast cancer cells. *Molecular Cancer*, 14, 2015.

J. Storey and R. Tibshirani. Statistical significance for genomewide studies. *PNAS*, 100: 9440–9445, 2003.

G. R. Terrell. The gradient statistic. *Computing Science and Statistics*, 34:206–215, 2002.

R. Tibshirani. The lasso method for variable selection in the cox model. *Stats. Med.*, 16: 385–395, 1997.

J. Tukey. The philosophy of multiple comparisons. *Statitical Science*, 6:100–116, 1991.

D. Witten and R. Tibshirani. Testing significance of features by lassoed principal components. *Ann Appl Stat*, 2:986–1012, 2008.

D. Witten and R. Tibshirani. Survival analysis with high-dimensional covariates. *Stat. Methods Med. Res.*, 19:29–51, 2010.

Qing-Yan Yin and Chun-Xia Zhang. Ensembling variable selectors by stability selection for the cox model. *CIN*, 2017:2747431, 2017.