LUNDS UNIVERSITET

Bachelor's Program in Mathematics

Variable selection for generalized linear mixed model by L1 penalization for predicting clinical parameters of ovarian cancer

by

Lan Hoa Diep

MASK11
Bachlor's Thesis (15 credits ECTS)
June, 2021
Co-Supervisor: Magnus Wiktorsson
Co-Supervisor: Anna Sandström Gerdtsson
Examiner: Anna Lindgren

Abstract

The quantity of biomarkers, which are proteins in this case, in ovarian cancer (OC) tumor and immune tissue regions of interest (ROIs) were measured with the new technology Digital Spatial Profiler (DSP). These measurements were used to construct regression models on the biomarkers to predict for two clinical parameters; tumor type ("Type 1" vs "Type 2") and the immune infiltration type ("Cavities" vs "Dispersed"). The dataset was divided into tumor and immune ROIs to analyze separately. A total of three models were constructed: immune ROI with immune infiltration type, immune ROI with tumor type, and tumor ROI with tumor type. Since there were repeated measurements on the same patient but on different ROIs, logistic linear mixed model with random intercept was used to account for the dependency of ROIs and allow for the intercept to vary between patients. Since there were too many biomarkers to regress on, Lasso was used in combination with mixed model (GLMMLasso) for automatic variable selection. The tuning parameter λ in Lasso was chosen using BIC with some supervision. The model of immune ROI with immune infiltration level included four variables with coefficients that make biological sense and has good fit with both the training and test data. The model of immune ROI with tumor type had three variables that also makes biological sense and fitted the training data well, but not too well for test data. The model of tumor ROI with tumor type had a total of 12 variables but some of the variable coefficients do not make sense biologically. It could probably be optimized by including fewer variables in the model. For any certain conclusion to be made about the predictability of the models, bigger sample size would be needed for refitting as well as testing the models.

Acknowledgements

Firstly, I would like to express my deepest appreciation to my co-supervisor Anna Sandström Gerdtsson for providing invaluable advises as well as the data, biological backgrounds and interpretations from beginning to end. Without whom the coherence of this paper would suffer tremendously. I also wish to thank my co-supervisor Magnus Wiktorsson for his insightful suggestions and guidance which points me at the right direction for the theis. Furthermore, I'd like to gratefully acknowledge the help that I received from Lavanya Lokhande, when I was stuck with terrible technical problems. Additionally, I want to thank my friend Adam Lindström for the discussions that helps me understand the mathematical theory behind this thesis better. Last but not least, I want to thank all my friends and family for always supporting and being patient with me throughout my study.

Popular science summary

Predicting aspects of ovarian cancer with proteins and machine learning

Ovarian cancer, like any other type of cancer, is a dangerous disease that we are still learning how to fight off. For now, the 5-year survival rate of ovarian cancer patient is only 38%, so more research still has to be done to improve this number. Fortunately, the body is smart enough to defense against cancer by sending immune cells into tumor areas. Therefore the immune infiltration level, indicating if the immune cells are spread out or clustered together, is an important metric to learn about. Moreover, the type of ovarian cancer, with type 2 being more aggressive than type 1, is also a factor that is important for patient survival.

Recent technology allows for the measurements of proteins in chosen areas to be known. With this, scientists can choose specific areas on the autopsy samples of cancer patient and know the amount of certain proteins, especially the ones that characterize immune and tumor cells. There are many such proteins (over 40 of them were measured), and it would be interesting to narrow down to the most important proteins that can be predictive of the immune infiltration level and cancer type. Regression is a popular supervised machine learning method that can do exactly that. Like other machine learning methods, regression needs to be trained in order to make accurate predictions. So in this paper, I report on the process of training a specific type of regression to do these specific tasks.

The regression method that was chosen is special in two ways. Firstly, it takes into account that some autopsy samples come from the same patient and are therefore somewhat related, and this violates the independence assumptions that normal regression has. Secondly, it can try out different combinations of proteins and choose the ones that are most likely to be predictive, while trying to keep the number of proteins included relatively small. There are a total of 3 predictive models produced at the end and 2 of them work well with up to 70% accuracy, while the last one still needs some more tuning and training to perform better. But it needs to be noted that the number of samples in this study is quite small (around 70), so before anything can be claimed with certainty, more samples would have to be studied in the future.

Contents

Chapter 1

Introduction

1.1 Research problem

Tumor mincroenvironment (TME) is a system of tumor cells, immune cells surrounding them and their interactions. Profiling the TME is important as it can give insights into the type of tumor as well as potential effective treatments, particularly given the recent emerge of immunotherapeutic drugs, which are directing the immune cells of the TME to battle the tumor. Ovarian cancer (OC) is the 4th most common cancer-associated cause of death in women, with few treatment options beyond surgery and chemotherapy, which in most cases is non-curative. The heterogeneous immune response in OC requires characterization of the TME to define biomarker signatures that can subgroup patients and identify those that could benefit from immunotherapy.

Recent technological breakthroughs have enabled parallel measurement of large numbers of tumor and immune biomarkers in tumor tissue biopsies, enabling characterization of the TME in much more detail than what was previously possible. One such new technology is the Nanostring GeoMx Digital Spatial Profiler (DSP), in which tissue biopsies are stained with large antibody panels enabling quantitation of biomarkers in tumor and immune regions of interest (ROIs) [1]. A visualization of tumor and immune ROIs can be seen in figure 1.1, where the pink parts are the tumor cells and green parts are the immune cells.

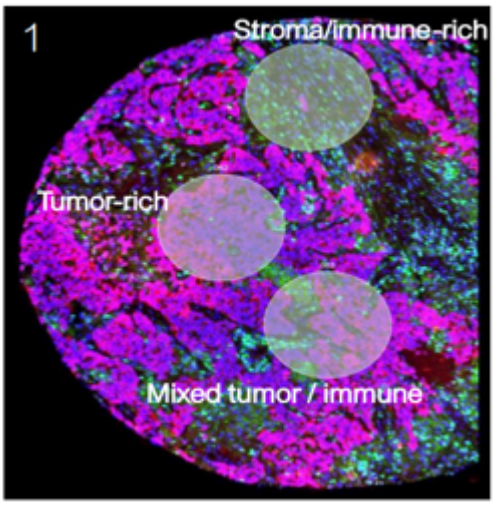

Figure 1.1: Tumor, immune and mixed regions of interest.

The purpose of this study was to construct a regression model on DSP data to identify predictive biomarkers. The dataset consisted of 44 biomarkers and was divided into immune and tumor cell ROIs, which were analyzed separately as they were too different in regard to e.g. number of cells and biomarker expression to be grouped together. As the majority of proteins were immune biomarkers, the immune data was deemed more clinically relevant to analyze. However, as not all tumor tissues had high enough immune infiltration to define immune ROIs, the immune dataset had a much smaller number of samples compared to the tumor dataset. Thus, regression was conducted in both datasets separately.

Various sample annotations could be interesting parameters to base the model on, such as disease outcome (overall survival time), OC histological subgroups, tumor type, and immune infiltration type. Preliminary statistical analysis showed low correlation to overall survival time. In addition, we opted for constructing the regression model on a categorical variable first. The four histological subtypes were ruled out as the sample groups unfortunately were too small to be divided into training and test sets. Hence, we constructed regression models based on tumor type (type 1 vs type 2) and immune infiltration type ('cavities' vs 'dispersed'). The immune infiltration type can be seen from figure 1.2.

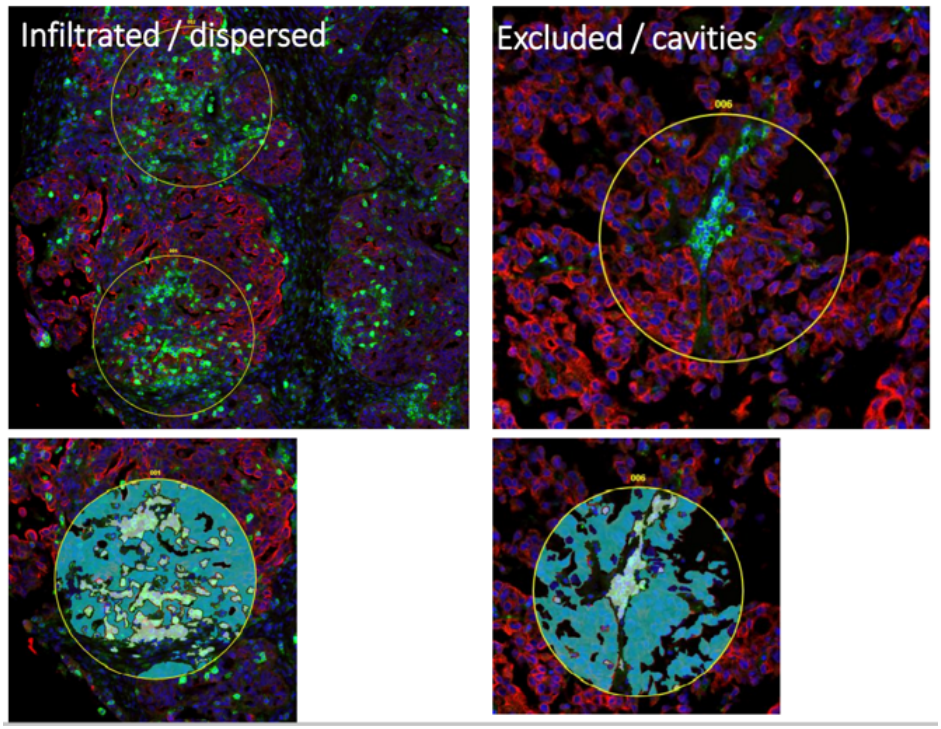

Figure 1.2: Immune infiltration type with dispersed and cavities.

Type 1 OC corresponds to so called low grade tumors which typically grow slower and metastasize late, while type 2 OC are so called high grade and more aggressive. Identifying differences in the TME of type 1 and type 2 OC would be interesting from a tumor biology perspective to understand the role of immune infiltration on disease progression in the different types. Also, as Type 1 and Type 2 tumors are relatively easily discriminated based on morphology (tumor cell appearance under the microscope), it would be clinically relevant to find differences in immune infiltration (and thus treatment options) that are reflected in standard histological grading.

During the sampling process, it had become evident that the type of immune infiltration looked different, and thus the ROIs had been annotated by visual inspection of the stained tissue into cavities (immune cells located in 'cavities' between tumor cells), dispersed (immune cells spread in between tumor cells), and insignificant (zero to low immune infiltration). With traditional methods such as imaging one or a few markers in parallel, or multiplex bulk tissue analysis where the tissue has been disrupted and the spatial information lost, this type of separation of different types of immune segments has not been possible. Thus, identifying differences in immune cells that are dispersed among the tumor cells versus those residing in cavities in between the tumor structures, would be highly interesting for advancing the understanding of TME heterogeneity in OC, and for defining biomarkers for treatment selection based on imaging.

The dataset included multiple (1-3) samples per patient, thus independency could not be assumed from samples derived from the same patient. When there are more than one sample, an average is usually taken, given that the variance between samples is not too large. However, this was not a valid strategy in this case since different types of ROIs, i.e. reflecting different TME structures, had been intentionally sam-

pled from the same patient. For example, ROIs of both the 'cavities' and 'dispersed' immune infiltration type could have been sampled from the same patient. Hence, the variance within a patient was sometimes big, and using averages would result in loss of information. The solution suggested by Nanostring is to use mixed model for this type of data, which introduces an extra factor to account for the clustering of samples coming from the same patient. That way we can regress with all available samples without having to take an average.

When trying mixed model regression on the immune dataset, there was immediately a problem with the number of variables being bigger than the number of samples for the immune section. The number of biomarkers hence had to be reduced somehow. One such method of feature reduction is Lasso regression. Thus, we decided to use a general linear mixed model (GLMM) combined with Lasso regression (GLMMLasso) to predict immune infiltration and tumor type.

1.2 Aim and Scope

The aim of this study was to evaluate the use of GLMMLasso regression to generate prediction models in DSP data, which frequently includes multiple samples per patient. The models may serve to identify predictive combinations of TME biomarkers related to various sample parameters, such as the location of immune cells in the tumor, and tumor subtypes. A combination of Lasso and generalized linear mixed model was applied with patient ID as a grouping variable, on measured protein biomarkers to construct three models:

- Predicting immune infiltration level ("Cavities" and Dispersed") for immune ROIs
- Predicting tumor type ("Type 1" and "Type 2") for immune ROIs
- Predicting tumor type ("Type 1" and "Type 2") for tumor ROIs.

Chapter 2

Theory

2.1 Least absolute shrinkage and selection operator (Lasso)

Lasso is a shrinkage method through $L1$ penalization. It maximizes the log-likelihood $l(\boldsymbol{\beta})$ while having constraint on the $L1$ norm of parameter vector $\boldsymbol{\beta}$. The Lasso estimate is defined as:

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{argmax}\ l(\boldsymbol{\beta}), \tag{2.1}$$

subject to $||\boldsymbol{\beta}||_1 \leq s$. With $s \geq 0$ and $||.||_1$ being the $L1$ norm. The Lasso estimate can also be derived from solving the optimization problem

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{argmax}[l(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_1]. \tag{2.2}$$

Both s and λ are tuning parameters and need to be determined by optimizing with either cross-validation or information criteria.

2.2 Generalized linear mixed model (GLMM)

The details and notations of generalized linear mixed models are mainly from the paper by Andreas Groll [2], with some minor differences.

Let y_{it} be observation t in cluster i of the response variable, where $i = 1, ..., m$ and $t = 1, ..., T_i$. The y_{it}'s are elements in the response vector $\mathbf{y}_i^T = (y_{i1}, ..., y_{iT_i})$. Note that T on $\mathbf{y}_i^T$ is for the transposition of the vector. Denote $\mathbf{x}_{it}^T = (1, x_{it1}, ..., x_{itp})$ as the covariate vector of fixed effects, in this case they are the measurement of different biomarkers of observation t in cluster i. Denote $\mathbf{z}_{it}^T = (1, z_{it1}, ..., z_{itq})$ as the covariate vector of random effect. Moreover, y_{it}'s are assumed to be conditionally independent with mean $\mu_{it} = E(y_{it}|\mathbf{b}_i, \mathbf{x}_{it}, \mathbf{z}_{it})$ and variance $var(y_{it}|\mathbf{b}_i) = \phi v(\mu_{it})$, with ϕ being dispersion parameter.

When the response variable is assumed to have any distribution other than Gaussian, specific link functions are used to allow for non-linear relationship between μ_{it} and predictors. Let g be a link function, then g is a monotonic, continuous function that

relates μ to the predictors as following:

$$g(\mu_{it}) = \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{z}_{it}^T\mathbf{b}_i = \eta_{it}^{par} + \eta_{it}^{rand}, \tag{2.3}$$

where $\eta_{it}^{par} = \mathbf{x}_{it}^T\boldsymbol{\beta}$ is a linear term with respect to the parameters in $\boldsymbol{\beta}^T = (\beta_0, ..., \beta_p)$, including the intercepts. While $\eta_{it}^{rand} = \mathbf{z}_{it}^T\mathbf{b}_i$ contains the cluster-specific effect $\mathbf{b}_i \sim N(0, \mathbf{Q})$, with a $q \times q$ covariance matrix $\mathbf{Q}$.
By collecting samples within cluster, equation (2.3) also has the form

$$g(\boldsymbol{\mu_i}) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \tag{2.4}$$

where $\mathbf{X}_i^T = (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT_i})$ and the design matrix $\mathbf{Z}_i^T = (\mathbf{z}_{i1}, ..., \mathbf{z}_{iT_i})$. With all the samples, one gets:

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \tag{2.5}$$

with $\mathbf{X}^T = [\mathbf{X}_1^T, ..., \mathbf{X}_m^T]$ and a block diagonal matrix $\mathbf{Z} = diag(\mathbf{Z}_1, ..., \mathbf{Z}_m)$. The random effect vector $\mathbf{b}^T = (\mathbf{b}_1^T, ..., \mathbf{b}_m^T)$ is assumed to have normal distribution with block diagonal covariance matrix $\mathbf{Q_b} = diag(Q, ..., Q)$.

A method for optimizing GLMM is penalized quasi-likelihood (PQL) that is suggested by Lin and Breslow [3]. It is assumed that the conditional density of y_{it}, given $\boldsymbol{\beta}$ and $\mathbf{b}_i$, belongs to a simple exponential family, which has the form

$$f(y_{it}|\mathbf{x}_{it}, \mathbf{b}_i) = exp\left\{\frac{y_{it}\theta_{it} - \kappa(\theta_{it})}{\phi} + c(y_{it}, \phi)\right\}, \tag{2.6}$$

where
$\theta_{it} = \theta(\mu_{it})$ is the natural parameter,
ϕ is dispersion parameter,
$\kappa(.)$ and $c(.)$ are specific functions corresponding to the type of exponential family [4].

Moreover, the covariance matrix $\mathbf{Q}$ of the random effect $\mathbf{b}_i$ is dependent on an unknown parameter vector $\boldsymbol{\sigma}$. The penalized-based likelihood-function is specified by $\boldsymbol{\gamma}^T = (\phi, \boldsymbol{\sigma}^T)$ and $\boldsymbol{\delta}^T = (\boldsymbol{\beta}, \mathbf{b}_i)$. So the log-likelihood is:

$$l(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{m} log\left(\int f(y_i|\boldsymbol{\delta}, \boldsymbol{\gamma})p(\mathbf{b}_i, \boldsymbol{\gamma})d\boldsymbol{b}_i\right), \tag{2.7}$$

where $p(\mathbf{b}_i, \boldsymbol{\gamma})$ denotes the density of random effect.

The approximate log-likelihood is derived by Clayton and Breslow [5] as:

$$l^{app}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{m} logf(y_i|\boldsymbol{\delta}, \boldsymbol{\gamma}) - \frac{1}{2}\mathbf{b}^T\mathbf{Q}^{-1}(\boldsymbol{\delta})\mathbf{b}. \tag{2.8}$$

PQL distinguishes between the estimation of $\boldsymbol{\delta}$, given the plugged-in estimate $\hat{\boldsymbol{\gamma}}$, and the estimation of $\boldsymbol{\gamma}$.

2.3 GLMMLasso

To combine GLMM and Lasso, the $L1$ penalty term is added to the likelihood function (2.8) to yield the penalized log-likelihood

$$l^{pen}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = l^{app}(\boldsymbol{\delta}, \boldsymbol{\gamma}) - \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i|. \tag{2.9}$$

Given $\hat{\boldsymbol{\gamma}}$, the optimization problem becomes

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{argmax}\, l^{pen}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\delta}}{argmax} \left[l^{app}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}}) - \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i| \right]. \tag{2.10}$$

2.4 Bayesian information criterion (BIC)

The tuning number λ controls how much shrinkage one wants for the parameters. Generally, the bigger the λ, the more shrinkage it is, thus the fewer parameters are going to be included in the model. The number λ is not selected automatically, the user therefore need to test and choose the λ that works best for the data. One way to do this somewhat automatically is by using the Bayesian information criterion (BIC), which allows for the comparison between models with different number of parameters [4]. The BIC for GLMMLasso, depending on the selected λ, has the form:

$$BIC_\lambda = -2\, l_\lambda^{pen}(\delta, \gamma) + p_\lambda\, log(n), \tag{2.11}$$

where p_λ is the number of parameters included in the model and n is the number of observations.

As seen from equation (2.11), BIC is the negative of log-likelihood with the additional penalty term for models with too many parameters. By choosing the model with the lowest BIC, it is equivalent to choosing the model with highest likelihood but also penalizes models that are too big, because likelihood always increases with the addition of parameters.

2.5 Model details

Since the response variable y_{it} is binary, the most obvious assumption for the distribution is $y_{it} \in Bin(1, \mu_{it})$. The natural link function for this is the log of odds, also called logit:

$$g(\mu_{it}) = ln\left(\frac{\mu_{it}}{1-\mu_{it}}\right) = ln\left(\frac{P(y_{it}=1)}{P(y_{it}=0)}\right). \tag{2.12}$$

Logit is used specifically for binary outcomes because it squeezes the probability of the event between 0 and 1. It can be seen easily from (2.12) that:

$$P(y_{it}=1) = \frac{1}{1+e^{-(\eta_{it}^{par}+\eta_{it}^{rand})}}. \tag{2.13}$$

Furthermore, the dependency of observations is accounted for by adding an intercept for the random effect. This is equivalent to assuming that the intercept is different for different patient. Since the slope of random effect won't be taken into account, the covariance matrix $\mathbf{Q}$ for the mix model is just a 1×1 matrix containg σ. This means that the specific effect b_i is just a random number with distribution $N(0, \sigma)$. While Z is a sparse matrix where each row is for an observation, and each column is for a patient with 0 and 1 indicating which observation belongs to which patient.

Chapter 3

Data and Methods

3.1 Data

All 44 biomarkers, which were proteins in this case, can be seen in Appendix A. To predict for immune infiltration ("Cavities" and "Dispersed"), only the immune ROIs were used. For this model, there was a total of 52 samples, 23 of which were "Cavities" and the other 29 were "Dispersed". To predict for tumor type ("Type 1" and "Type 2"), the immune and tumor ROIs were used separately. There were 55 samples/ROIs in the immune data, of which 26 were "Type 1" and 29 were "Type 2". There were 91 samples/ROIs in the tumor data, with 43 of them being "Type 1" and 48 "Type 2". For all of these three models, 80 % of the samples were used for training and the remaining 20 % were used for testing.

Prior to the work described in this study, data had been normalized in two steps:
1. Scaling based on control probe spike-ins to adjust for technical variation in the read-out/quantitation process.
2. Scaling based on geometric mean value of two house-keeper proteins (S6 and GAPDH) to adjust for sample-based variation, including region size/number of cells measured and background signal. Tumor and immune segments differed in regards to number of cells captured (in general fewer immune cells than tumor cells) and biomarker level of immune- and tumor-specific markers, respectively. Hence, data was separated into tumor and immune sets prior to the second normalization step, to avoid having to use disproportionately large scaling factors that may transform the data so that biological variation is masked. After normalization, data was log2 transformed.

3.2 Model selection

All data analysis was done in R version 4.0.4 (2021-02-15) and Rstudio version 1.4.1106. Function glmmLasso() from R Package "glmmLasso" [6] was used for mixed model Lasso regression. This function requires the user to input a λ value. The best λ is chosen by fitting the model with a range of λ values and choose the model with the lowest BIC. The computational cost of this process is quite high so a suitable range was determined first by screening a big range with poor "resolution", for example from 0 to 100 with step size of 1, to get a decent λ. Then a smaller range around this λ with better "resolution" was used to get to a λ with more significant

figures, for example from 0 to 10 with step size of 0.01.

Selecting a model completely based on BIC, sometimes resulted in "overfitting", which is evident by the inclusion of excessive parameters to the point that the standard errors for coefficient estimates is absurdly large and the p values of all estimate are basically 1. Examples of this can be seen in Appendix B. To select a model that does not suffer from this issue but still have one of the lowest BIC, the range of λ was restricted to exclude λs that are too small. Since low λ corresponds to bigger numbers of parameters, choosing the range of λ to be above certain number means limiting the number of parameters being included in the model. To choose the lower limit of the range of λ, it is helpful to look at the graph of BIC against λ.

3.3 Classifying threshold

Since this is a logistic model, the fitted value is the probability that a sample belongs to a category, or "success". Therefore it is necessary to decide on a cutoff value, which is a threshold for classifying a sample as "success". Probability falls between 0 to 1, so threshold is a number between 0 and 1. If the fitted value for a sample is bigger than the threshold, the sample is classified as "success" and if not, it is classified as "failure". Intuitively, it makes sense for the cutoff value to be 0.5. This is the default case if there is no additional information given. The number of false positive and false negative depends on the cutoff value. And in some cases, one may want to trade off false positive for false negative or vice versa, especially when making one type of error results in significantly worse consequences than the other. For example in the case of detecting tumor, it is detrimental to have a false negative, but a false positive is corrected easily through more examining.

The ROC curve, which plots true positive rate against false positive rate, was used to determine the optimal cutoff value to get the desired false positive and negative rate. Using the function ROCInfo() found on Github [7], which calculates penalty for errors depending on the costs of false positive and false negative that are input by user, the best cutoff value was selected to minimize the penalty. In this case, the costs of both are the same and are set to 100.

3.4 Model validation

The optimal cutoff value was chosen using the ROC curve of training data as mentioned above. The accuracy of the model with that specific cutoff value can be presented in what is called a confusion matrix. Both the confusion matrix of the training data as well as the test data were included for examining the accuracy of the model.

Chapter 4

Result

4.1 Model of immune segment with immune infiltration level as response variable

The range of λ was first chosen to go from 0 to 100 with step size of 1 for initial screening. The BIC of glmmLasso models with the given λ's is plotted against λ's in figure 4.1. It can be seen from the figure that once λ is bigger than around 10, the BIC is constant as they are models without any parameter. The λ range was therefore chosen again to go from 0 to 10 with step size of 0.01. With this choice of λ range, the best model chosen with BIC resulted in an overfitted model which can be seen in table B.2 in Appendix B. Looking at the plot of BIC against λ in figure 4.2, the λ range was chosen again from 5 to 10 with step size of 0.01 since there seemed to be some instability right before 5. The final model chosen within this range of λ can be seen in table 4.1. Note that in this model, "Dispersed" is 1 and "Cavities" is 0. The ROC curve and the optimal cutoff value of the train data is shown in figure 4.3. The confusion matrix of the train and test data using the optimal cutoff value are shown respectively in table 4.2 and table 4.3.

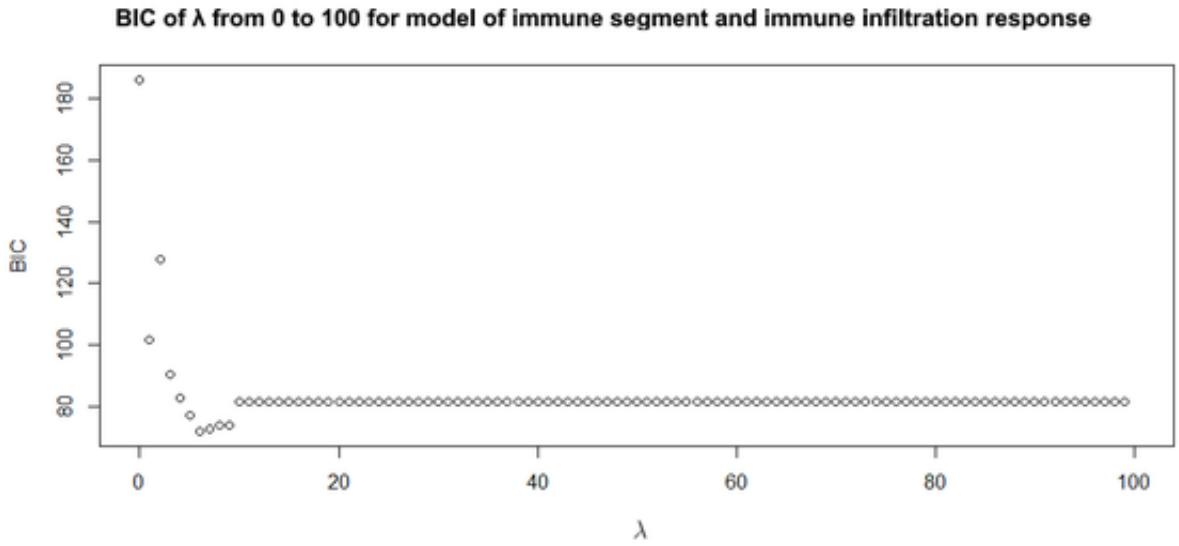

Figure 4.1: Plot of BIC against λ for model of immune segment with immune infiltration as response where λ ranges from 0 to 100.

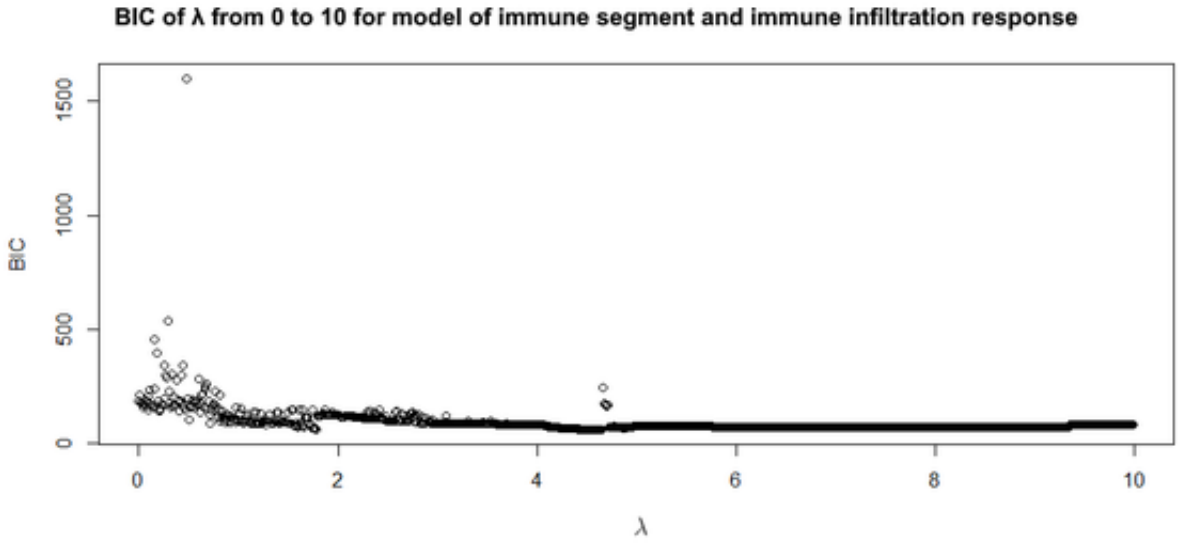

Figure 4.2: Plot of BIC against λ for model of immune segment with immune infiltration as response where λ ranges from 0 to 10.

	Estimate	StdErr	z.value	p.value
(Intercept)	-11.3436662	0.60925654	-18.6188665	2.2595E-77
GZMB	1.28842474	0.80553665	1.59946136	0.10971813
X4.1BB	0.76448876	0.82047025	0.93176902	0.35145591
GITR	0.56180648	0.63402964	0.88608868	0.37556974
IDO1	0.81720263	0.44633831	1.83090407	0.06711486
λ	6.69			
StdDev of Patients	1.790226			

Table 4.1: Final model for immune segment with immune infiltration as response variable.

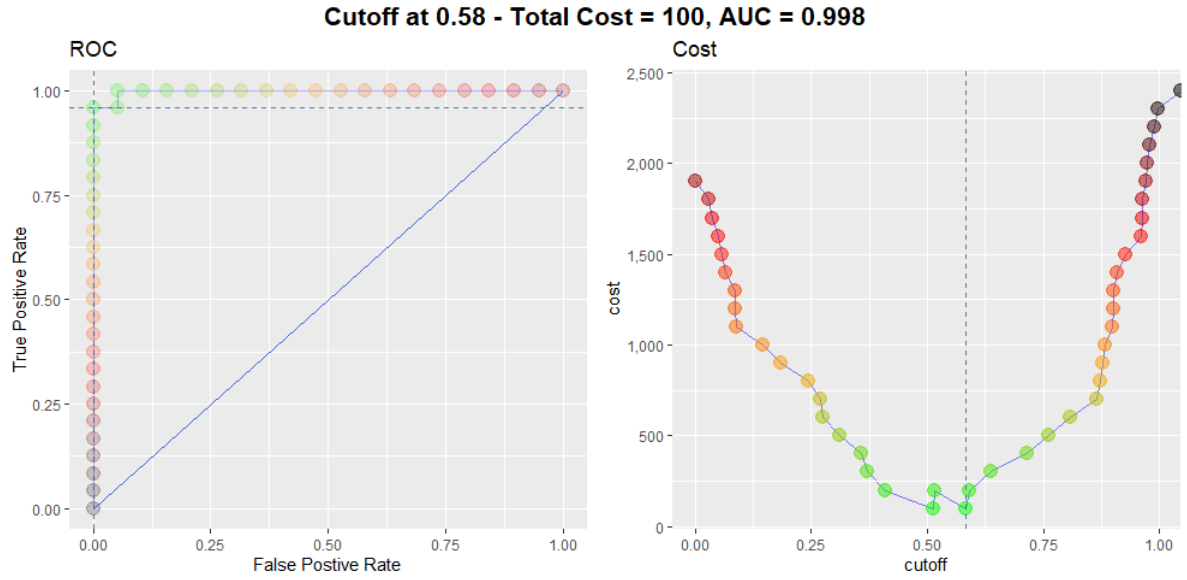

Figure 4.3: ROC curve and optimal cutoff value using train data for final model of immune segment with immune infiltration response.

Prediction	Reference	
	Cavities	Dispersed
Cavities	19	1
Dispersed	0	23
Sensitivity		1.0000
Specificity		0.9583
Accuracy		0.9767

Table 4.2: Confusion matrix using training data for model of immune segment with immune infiltration response.

Prediction	Reference	
	Cavities	Dispersed
Cavities	3	1
Dispersed	1	4
Sensitivity		0.7500
Specificity		0.800
Accuracy		0.7778

Table 4.3: Confusion matrix using test data for model of immune segment with immune infiltration response.

4.2 Model of immune segment with tumor type as response variable

Similarly, λ range at first was set from 0 to 100, the BIC plot of which can be seen in figure 4.4. The range of λ was then narrowed down to 0 to 20, the plot of which is seen in figure 4.5. This range also resulted in an overfitted model with λ being 6.99. This model can be seen in table B.4 in Appendix B. The λ range was therefore restricted from 7 to 20. This range resulted in the final model in table 4.4. Note that in this model, "Type 2" is 1 and "Type 1" is 0. Its ROC curve and optimal cutoff value for train data is seen in figure 4.9. The confusion matrices for the train data and test data are shown in table 4.5 and table 4.6 .

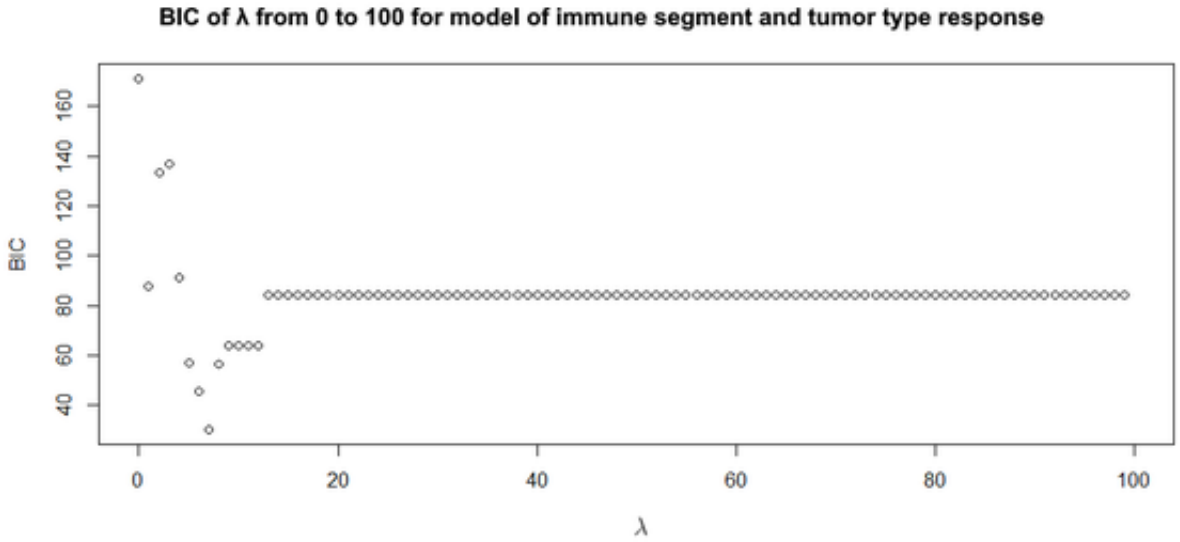

Figure 4.4: Plot of BIC against λ for model of immune segment with tumor type as response where λ ranges from 0 to 100.

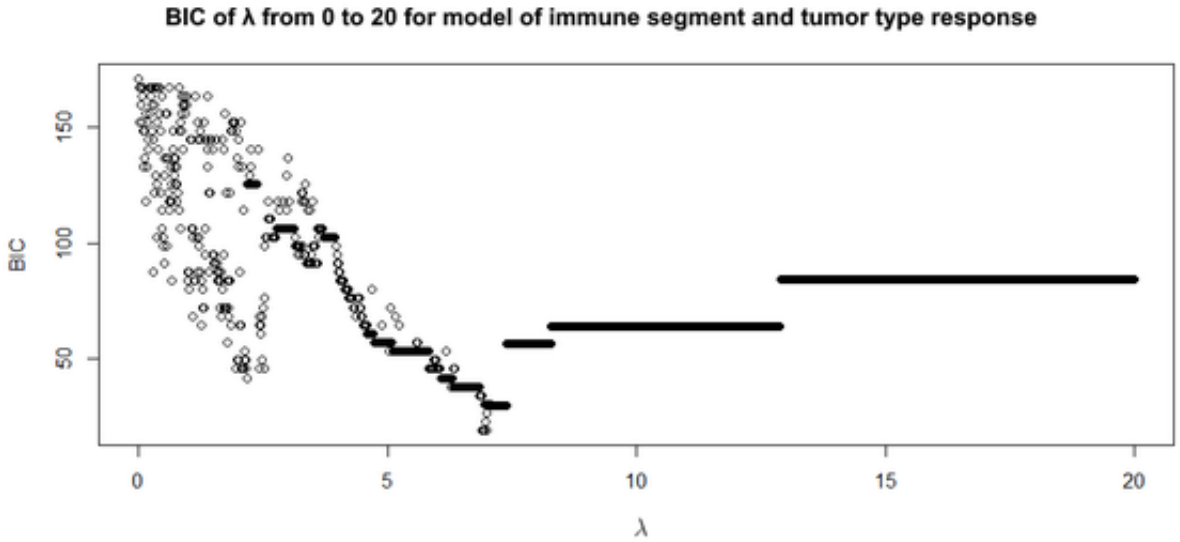

Figure 4.5: Plot of BIC against λ for model of immune segment with tumor type as response where λ ranges from 0 to 20.

	Estimate	StdErr	z.value	p.value
(Intercept)	-45.6676301	0.97419606	-46.877248	0
CD44	6.75053553	4.58727213	1.47157948	0.14113447
CD45RO	-7.68681207	5.72766732	-1.34204933	0.17958
B7.H3	5.19858718	3.16918547	1.64035435	0.10093151
λ		7.33		
StdDev of Patients		1.05507		

Table 4.4: Final model for immune segment with tumor type as response variable.

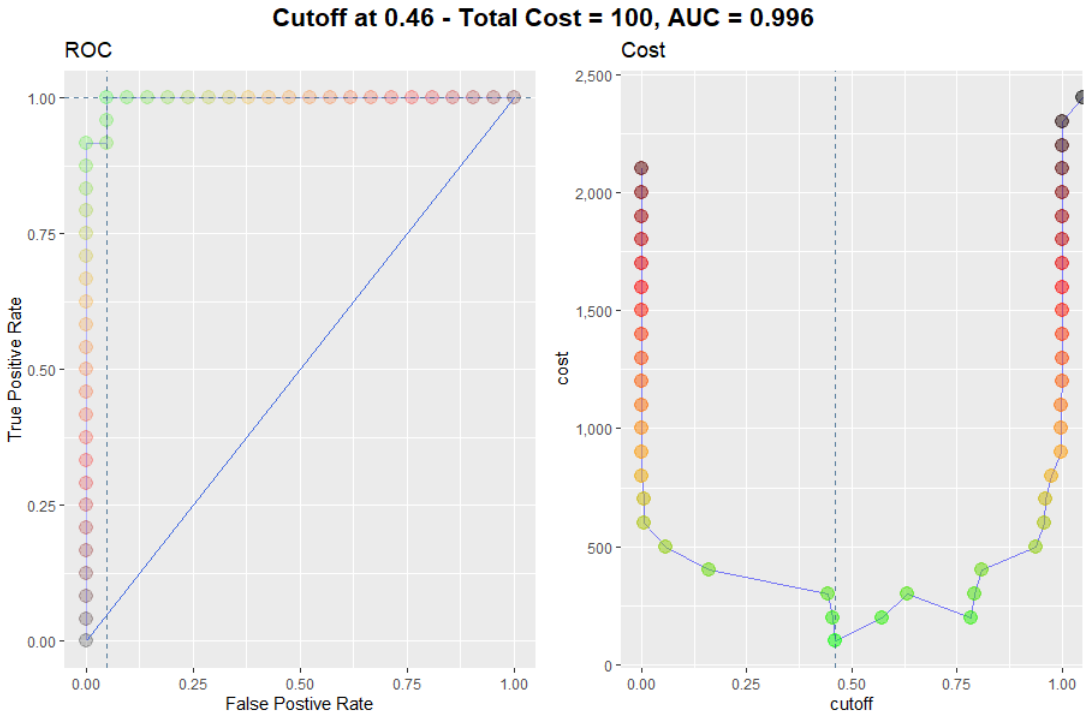

Figure 4.6: ROC curve and optimal cutoff value for final model of immune segment with tumor type response.

Prediction	Reference	
	Type 1	Type 2
Type 1	20	0
Type 2	1	24
Sensitivity		0.9524
Specificity		1.0000
Accuracy		0.9778

Table 4.5: Confusion matrix using training data for model of immune segment with tumor type response.

Prediction	Reference	
	Type 1	Type 2
Type 1	3	1
Type 2	2	4
Sensitivity		0.60
Specificity		0.80
Accuracy		0.70

Table 4.6: Confusion matrix using test data for model of immune segment with tumor type response.

4.3 Model of tumor segment with tumor type as response variable

Similarly to the previous models, a screening step with λ going from 0 to 100 with step size of 1 was used. From the plot in figure 4.7, the λ range was narrowed down to 0 to 20 with step size 0.1 to focus on the interesting range. Once again, this resulted in a strange model that can be seen in table B.3 in Appendix B with absurdly big coefficients, a small change in any parameter leads to an extremely big change in the prediction. A model like this is too unstable so the range of λ was chosen again for model selection. As seen in figure 4.8, the BIC seemed to be unstable right before λ of 10. So the λ range was narrowed down to 10 to 20 with step size of 0.1. The final model can be seen in table 4.7. Similar to the above model, "Type 2" is 1 and "Type 1" is 0 for this model. The ROC curve and optimal cutoff value of train data can be seen in figure 4.9. The confusion matrices of train and test data are respectively in table 4.8 and table 4.9.

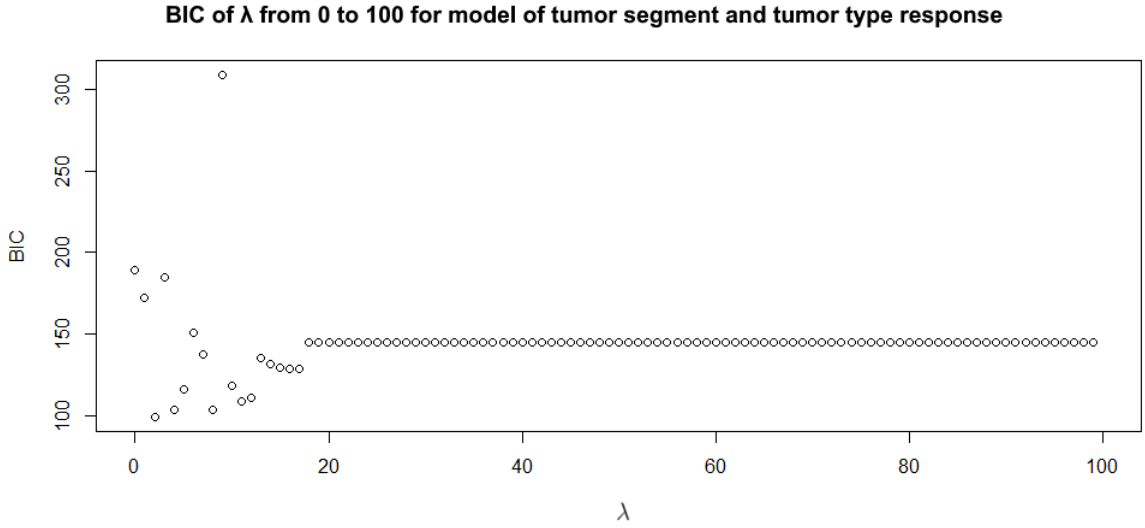

Figure 4.7: Plot of BIC against λ for model of tumor segment with tumor type as response where λ ranges from 0 to 100.

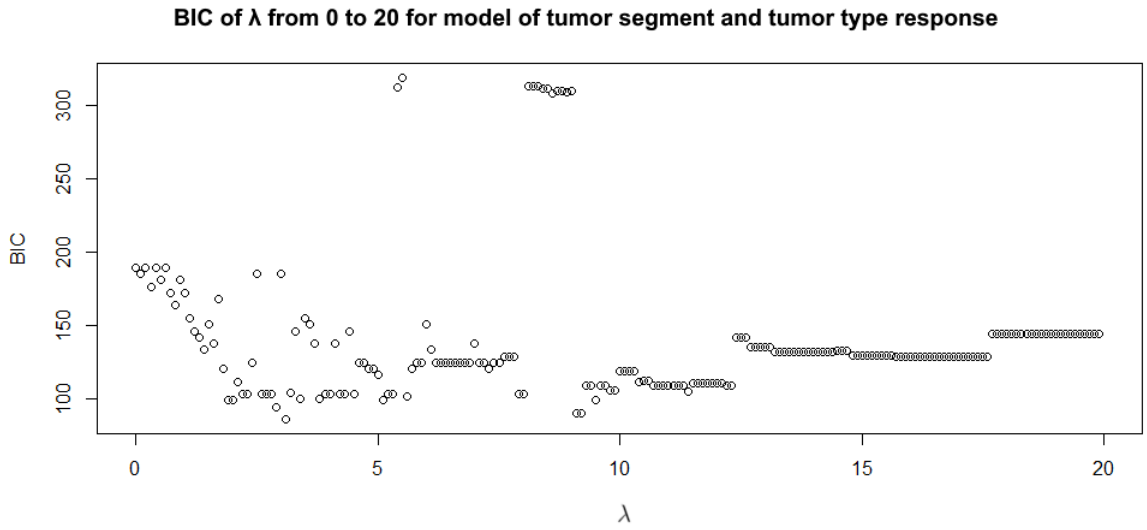

Figure 4.8: Plot of BIC against λ for model of tumor segment with tumor type as response where λ ranges from 0 to 20.

	Estimate	StdErr	z.value	p.value
(Intercept)	2.63358529	0.43899663	5.99910134	1.9841E-09
CD56	0.61931942	0.58267606	1.06288805	0.28783272
CD4	-0.20652548	1.08904289	-0.18963944	0.84959168
PanCk	-1.04171873	0.44926723	-2.31870621	0.02041097
Ki.67	1.07339272	0.44613494	2.40598221	0.01612905
Histone.H3	-1.396274	0.86318391	-1.61758576	0.1057519
CD3	0.33695207	0.67306295	0.50062489	0.61663514
CD11c	1.78662254	1.05953598	1.68623112	0.09175129
CD34	-2.00201903	0.79773991	-2.50961374	0.01208633
X4.1BB	0.91186451	0.5601158	1.62799284	0.1035264
ARG1	0.78913183	0.78118092	1.01017806	0.31240999
IDO1	-0.42418091	0.34385816	-1.23359268	0.21735472
B7.H3	1.06950008	0.65675991	1.6284491	0.10342969
λ		11.4		
StdDev of Patients		0.3532506		

Table 4.7: Final model for tumor segment with tumor type as response variable.

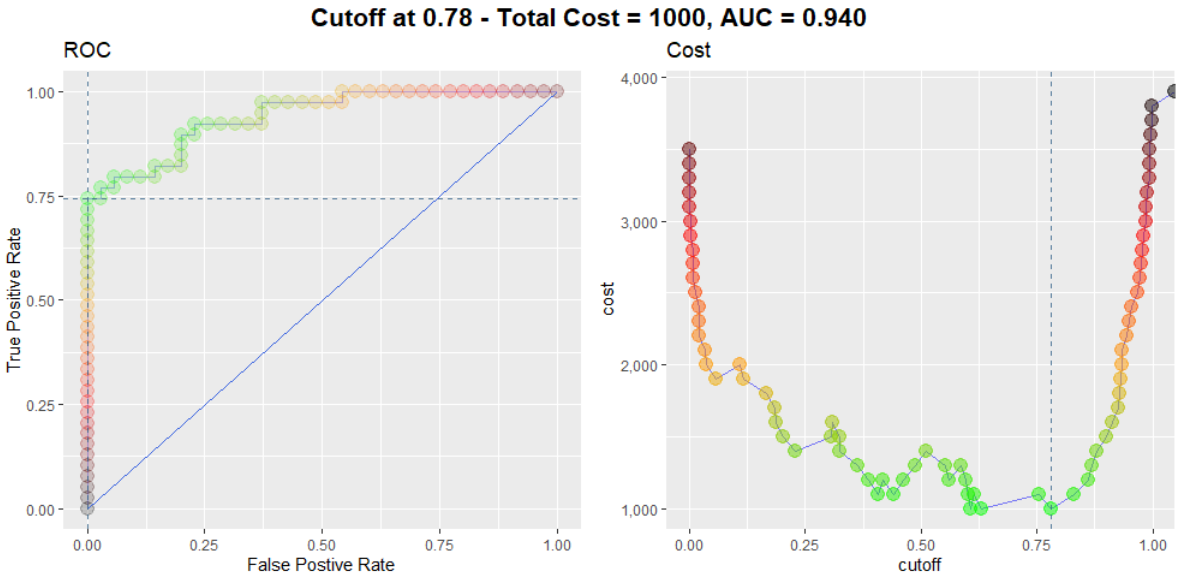

Figure 4.9: ROC curve and optimal cutoff value for final model of tumor segment with tumor type response.

Prediction	Reference	
	Type 1	Type 2
Type 1	35	10
Type 2	0	29
Sensitivity		1.0000
Specificity		0.7436
Accuracy		0.8649

Table 4.8: Confusion matrix using training data for model of tumor segment with tumor type response.

Prediction	Reference	
	Type 1	Type 2
Type 1	6	3
Type 2	2	6
Sensitivity		0.7500
Specificity		0.6667
Accuracy		0.7059

Table 4.9: Confusion matrix using test data for model of tumor segment with tumor type response.

Chapter 5

Discussion

The final model for immune segment data with immune infiltration (dispersed vs cavities) as a response variable included 4 biomarkers, all with higher levels in "dispersed" (Table 4.1). The biomarkers were all functional molecules that indicated a more active immune response in the dispersed type. GZMB (Granzyme B) is released by immune cells and induces apoptosis (cell death) of tumor cells. 4-1BB and GITR (glucocorticoid-induced TNFR-related protein) are costimulatory immune checkpoint markers expressed by activated T-cells. IDO (Indoleamine-pyrrole 2,3-dioxygenase) has previously been demonstrated overexpressed in OC, and shown to correlate with immune infiltration [8]. All the biomarkers included were proteins expected to have higher expression in dispersed type of immune infiltration compared to tumor excluded immune cells located in cavities, which supports the validity of the model. The model also performed well in predicting the limited number (9) of samples in the test data (Table 4.3) but will have to be validated in larger datasets.

The final model for immune segment data with OC type as a response variable, included 3 biomarkers (Table 4.4). CD44 is among other things a marker that distinguishes active from naïve T-cells [9]. CD45RO is also expressed on several type of (activated) immune cells but is primarily a marker of memory T-cells. The opposing coefficients of CD44 (higher in Type 2) and CD45RO (higher in Type 1) in the model could imply that Type 2 ovarian tumors have a higher ratio of effector T-cells, while type 1 tumors have more memory T-cells. The third marker, B7-H3 and is known to suppress immune response to tumors and has previously been correlated to high grade (Type 2) OC [10], so the coefficient in the model for B7-H3 also makes sense. The model, however, had relatively poor accuracy in predicting the limited number (n=10) of test data samples, and will have to be further optimized in larger cohorts.

The final model for tumor segment data with OC type as a response variable, included 12 biomarkers (Table 4.7). A higher expression of proliferation marker Ki67 in Type 2 was expected, as high grade tumors are more aggressive with a faster proliferation rate. The higher expression of epithelial marker PanCK in Type 1 was also expected as low grade tumors are more differentiated and generally have a higher levels of epithelial markers. Hence, the inclusion of Ki67 and panCK and their relatively low individual p-values adds validity to the model. In contrast, the overall expression level of Histone H3 was not expected to differ between Type 1 and Type 2. In addition, its model coefficient indicated lower expression in Type 2, while

the univariate comparison showed that the signal was higher in Type 2, both in the complete data (including all ROIs) and for mean value per patient (Figure 5.1). The remaining variables in the model were mainly biomarkers of immune phenotype or function, and even though some of them (e.g. B7-H3 and IDO1) can be expressed by tumor cells as well, their inclusion in the model may be a result from bleed-over signal from the immune segments. Some markers, in particular CD4 and CD3, also show high individual p-values, suggesting that the model could be overtrained. This is also indicated by the relatively poor accuracy in both training (Table 4.8) and test (Table 4.9) data. The prediction could be further optimized, likely using less biomarkers, which can be done by once again increasing the lower bound of lambda range.

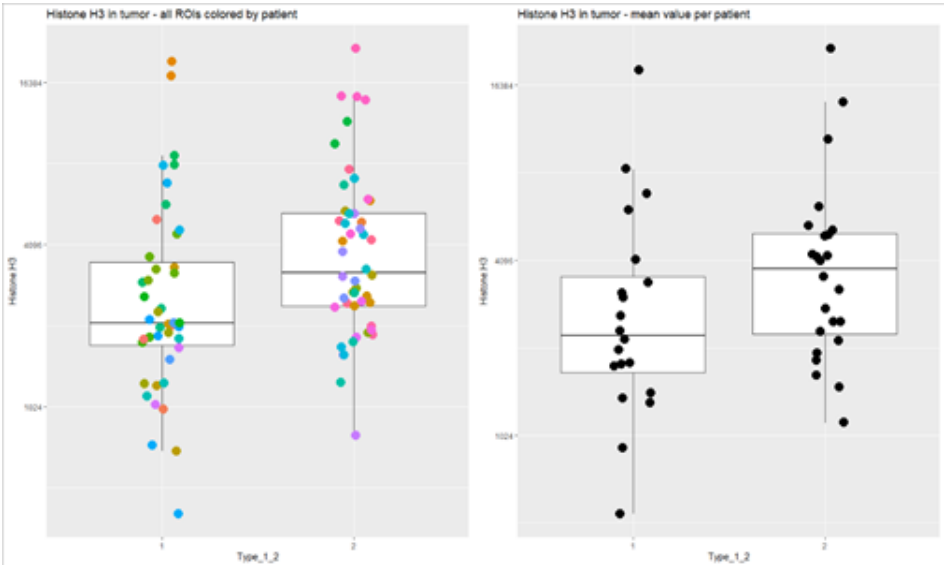

Figure 5.1: Measurements of Histone H3 in tumor segments. Left panel includes data from all ROIs. Right panel includes average values per patient

In the three final models above, it can be seen that the p-values of the coefficients are are not significant. Some of the p-values are even as high as 0.8 such as the one for $CD4$ in table 4.7. It is unclear how reliable the p-values are for the models since there is an additional penalty term. A possible way to obtain better p-values may be refitting the model with a mixed model method but using only the leftover parameters obtained from GLMMLasso. This is equivalent of using GLMMLasso only as a way to select out the most important parameters. Alternative method for more significant parameter can be forward inclusion of parameter with significant p-value until the addition of parameter does not improve the p-values. This method can replace Lasso as a variable selection method.

When selecting λ and corresponding model automatically using BIC, the resulting model tends to overfit the training data. This may be due to the fact that the number of samples is not big enough for BIC to punish big models properly. After all, the probability of BIC working consistently approaches 1 as $n \to \infty$, so the more samples there are, the more reliable the BIC would be. Moreover, looking at figure 4.8, it can be seen that just changing λ slightly can result in big fluctuation of BIC. This might just be due to the small number of samples as mentioned, or it might

indicate that the method GLMMLasso is not very stable for the dataset.

By allowing the intercept to vary between patients, this mixed model takes into account the fact that different patients have different probability of belonging in a certain group even if all the measurements are the same. Which is an appropriate assumption that also accounts for the dependency of observations on the patients. If one wants more variation, one can also allow for the parameter coefficients to vary. This would mean that the effect of changing biomarkers is different for different patients. This is not done here but can potentially be added in future models if deemed suitable.

Moreover, with the availability of more data, an evaluation set can be used to find a better cutoff value from its ROC curve. This would potentially provide a better cut-off value for the test data since both of these datasets are "unseen" data. Therefore, a more unbiased cutoff can be found using the evaluation data compared to the training data.

In conclusion, the GLMMLasso method has some unstability that might or might not be caused by the small sample size. Furthermore, the automatic way of selecting λ and model using BIC without any supervision is proved to be unreliable for this dataset. It is likely that this method works better with bigger sample size but supervision is still needed to make sure the models do not grossly overfit. Regardless, two of the models produced by GLMMLasso with supervised λ selection make biological sense and fit the training data well. For better assessment on test data, larger sample size would be needed as well as an evaluation dataset for finding cutoff value. For the third model, some biomarkers that are included do not have the expected coefficient signs. This is probably due to the fact that too many biomarkers were included. So far, the analysis done in this study is quite preliminary. As this type of data is still very new, more investigation needs to be done in order to conclude the effects of the biomarkers with more certainty.

Bibliography

[1] James Monkman, Touraj Taheri, Majid Ebrahimi Warkiani, Connor O'Leary, Rahul Ladwa, Derek Richard, Ken O'Byrne, and Arutha Kulasinghe. High-plex and high-throughput digital spatial profiling of non-small-cell lung cancer (nsclc). *Cancers (Basel)*, 12(12), 2020.

[2] Andreas Groll and Gerhard Tutz. Variable selection for generalized linear mixed models by l1-penalized estimation. *Stat Comput*, 24(2), 2014.

[3] Xihong Lin and Norman E. Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435), 1996.

[4] Ludwig Fahrmeir and Gerhard Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Model*. Springer, München, Germany, 2001.

[5] D. G. Clayton and Norman E. Breslow. Journal of the american statistical association. *Journal of the American Statistical Association*, 88(421), 1993.

[6] Andreas Groll. *glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*, 2017. R package version 1.5.1.

[7] Ethen Liu. Unbalanced. `https://github.com/ethen8181/machine-learning/blob/master/unbalanced/unbalanced_code/unbalanced_functions.R`, 2016.

[8] Quan Zhou, Fan-Hua Cao, Hui Liu, and Man-Zhen Zuo. Comprehensive analysis of the prognostic value and immune function of the ido1 gene in gynecological cancers. *American journal of translational research*, 13(4), 2021.

[9] Julia Schumann, Katarina Stanko, Ulrike Schliesser, MChristine Appelt, and Birgit Sawitzki. Differences in cd44 surface expression levels and function discriminates il-17 and ifn-γ producing helper t cells. *PloS one*, 10(7), 2015.

[10] D. Cai, J. Li, and D. et al Liu. Tumor-expressed b7-h3 mediates the inhibition of antitumor t-cell functions in ovarian cancer insensitive to pd-1 blockade therapy. *Cell Mol Immunol*, 17, 2020.

Appendix A

Biomarkers

GZMB	CTLA4	CD56	CD45	HLA.DR	CD68
PanCk	PD.1	Fibronectin	CD20	SMA	Ki.67
CD8	Beta.2.microglobulin	CD11c	CD44	CD40	CD80
CD25	ICOS	CD27	CD163	FAP.alpha	FOXP3
CD34	CD45RO	X4.1BB	LAG3	ARG1	VISTA
STING	IDO1	B7.H3	Tim.3	CD4	PD.L1
Histone.H3	CD3	CD127	PD.L2	CD66b	CD14
OX40L	GITR				

Table A.1: All proteins used as parameters for regression.

Appendix B

Models with overfitting problems

The tables displayed in this Appendix are examples of bad models to show the signs of overfitting. They are in no way meant to represent useful models.

	Estimate	StdErr	z.value	p.value
(Intercept)	-2437.14859	9132359.67	-0.00026687	0.99978707
GZMB	-262.037679	104148569	-2.516E-06	0.99999799
CTLA4	85.0757482	30733820.5	2.7681E-06	0.99999779
CD56	-75.1177358	28672946.9	-2.6198E-06	0.99999791
CD45	-173.645797	48436545	-3.585E-06	0.99999714
HLA.DR	-54.0087012	39313286.5	-1.3738E-06	0.9999989
CD68	2.88769521	38328445.3	7.5341E-08	0.99999994
PD.L1	27.9247636	20977582.6	1.3312E-06	0.99999894
PanCk	-32.30393	21408999.5	-1.5089E-06	0.9999988
PD.1	67.3412654	34547308.8	1.9492E-06	0.99999844
Fibronectin	177.230166	21122060.7	8.3908E-06	0.99999331
CD20	53.6574807	34879961.1	1.5383E-06	0.99999877
SMA	-99.5682113	19463507.7	-5.1156E-06	0.99999592
Ki.67	-38.548418	23899836.4	-1.6129E-06	0.99999871
Histone.H3	344.974821	72183256.5	4.7792E-06	0.99999619
CD3	222.796261	43315296.8	5.1436E-06	0.9999959
CD8	-68.9641152	34399008.1	-2.0048E-06	0.9999984
Beta.2.microglobulin	-64.2026116	34248926	-1.8746E-06	0.9999985
CD11c	108.271944	66825001.1	1.6202E-06	0.99999871
CD44	-225.918002	31451613.9	-7.183E-06	0.99999427
CD40	15.9559193	31443105.7	5.0745E-07	0.9999996
CD80	20.9698071	25779308.8	8.1344E-07	0.99999935
CD127	148.576475	59554525.5	2.4948E-06	0.99999801
PD.L2	106.181472	29437082.1	3.6071E-06	0.99999712
CD25	95.575084	60355639.7	1.5835E-06	0.99999874
ICOS	-5.78815966	42979553.4	-1.3467E-07	0.99999989
CD27	-58.9451021	21144543.4	-2.7877E-06	0.99999778
CD163	-26.931924	21095114.9	-1.2767E-06	0.99999898
FAP.alpha	73.9965425	29533521.3	2.5055E-06	0.999998
FOXP3	-49.9685677	21823338.2	-2.2897E-06	0.99999817
CD66b	-67.0765295	17292600.7	-3.8789E-06	0.99999691
CD14	138.276757	64766572.1	2.135E-06	0.9999983
CD34	-114.146407	26453062.9	-4.3151E-06	0.99999656
CD45RO	-19.333995	48009789.6	-4.0271E-07	0.99999968
X4.1BB	-143.494154	24973021.2	-5.746E-06	0.99999542
LAG3	188.590872	36386915	5.1829E-06	0.99999586
ARG1	-26.9526558	29003501.4	-9.2929E-07	0.99999926
VISTA	65.7515697	47882378.6	1.3732E-06	0.9999989
OX40L	-73.3212253	29282966.8	-2.5039E-06	0.999998
GITR	-69.5311116	26627890.2	-2.6112E-06	0.99999792
STING	-45.5671405	29699424.5	-1.5343E-06	0.99999878
IDO1	85.7183511	14963054.5	5.7287E-06	0.99999543
B7.H3	-78.0330857	26311461.8	-2.9657E-06	0.99999763
Tim.3	122.622093	29149904.4	4.2066E-06	0.99999664
Lambda		2.14		
StdDev of Patients		0.5875498		

Table B.1: Optimal glmmLasso model based on BIC for tumor segment with immune inflitration as response variable.

Coefficients	Estimate	StdErr	z.value	p.value
(Intercept)	1.8740e+02	8.7660e+06	0	1
GZMB	2.5135e+02	1.5053e+07	0	1
CTLA4	8.9799e+01	1.0149e+07	0	1
PD.L1	1.7373e+02	1.4380e+07	0	1
Fibronectin	-3.1461e-01	1.4583e+07	0	1
CD20	-1.9060e+02	1.0268e+07	0	1
SMA	-1.0631e+02	1.4843e+07	0	1
CD3	3.1779e+01	1.7027e+07	0	1
CD8	2.4197e+01	1.2581e+07	0	1
CD11c	-1.2762e+02	1.4597e+07	0	1
CD44	-6.4837e+01	8.1435e+06	0	1
CD163	-2.7763e+01	1.0345e+07	0	1
CD45RO	1.4072e+01	2.2178e+07	0	1
X4.1BB	-1.6482e+01	1.6240e+07	0	1
VISTA	5.9707e+01	1.0228e+07	0	1
GITR	7.0934e+01	1.6078e+07	0	1
IDO1	9.2613e+01	8.5347e+06	0	1
Lambda		4.62		
StdDev of Patients		0.5705323		

Table B.2: Optimal glmmLasso model based on BIC for immune segment with immune inflitration as response variable.

	Estimate	StdErr	z.value	p.value
(Intercept)	9.7915E+15	4.5965E+14	21.302228	0
GZMB	7.5565E+14	63835415.5	11837479.6	0
CTLA4	-2.6387E+14	31186405.1	-8460902.74	0
CD56	-2.4221E+13	47132488.8	-513900.854	0
CD45	-1.6512E+15	68758311.9	-24014096.4	0
CD4	1.7437E+15	70005823.6	24907708.3	0
PD.L1	1.5006E+14	20867224	7191013.93	0
PanCk	-1.1085E+15	32985219.2	-33606632	0
PD.1	-6.0578E+14	27276014.3	-22209218.7	0
Fibronectin	-7.4287E+13	23775973	-3124447.17	0
CD20	-1.2352E+15	64274442.8	-19217789.7	0
SMA	4.265E+14	28154080.6	15148743.9	0
Ki.67	4.1428E+14	37506522	11045529.5	0
Histone.H3	-7.6543E+14	63802746.9	-11996888.6	0
CD3	1.4998E+15	44269781.9	33878065.7	0
CD11c	-1.7114E+14	64987458.1	-2633375.5	0
CD44	-1.2157E+14	27905195.5	-4356588.27	0
CD163	2.798E+14	25684805.2	10893767.2	0
CD34	1.6164E+14	43062704.7	3753504.85	0
X4.1BB	3.7306E+14	25663664.5	14536604.1	0
ARG1	6.5233E+14	31200247.5	20907837.5	0
IDO1	-3.5013E+13	25650303.5	-1365029.26	0
B7.H3	-3.1735E+14	47303844.5	-6708703.9	0

Lambda	9.2
StdDev of Patients	2.73146e+15

Table B.3: Optimal glmmLasso model based on BIC for tumor segment with tumor type as response variable.

	Estimate	StdErr	z.value	p.value
(Intercept)	-2249.41776	10003998.8	-0.00022485	0.99982059
HLA.DR	-53.5539263	12049842.1	-4.4444E-06	0.99999645
Beta.2.microglobulin	40.9280318	18800535.9	2.177E-06	0.99999826
CD44	284.452382	8885944.94	3.2011E-05	0.99997446
CD45RO	-363.351474	14575303	-2.4929E-05	0.99998011
IDO1	75.9050896	6856951.83	1.107E-05	0.99999117
B7.H3	289.432234	12909211	2.2421E-05	0.99998211

Lambda	6.99
StdDev of Patients	0.8032166

Table B.4: Optimal glmmLasso model based on BIC for immune segment with tumor type as response variable.