Statistical Learning by Quasi-linear Predictor

OMAE KATSUHIRO

Doctor of Philosophy

Department of Statistical Science
School of Multidisciplinary Sciences
SOKENDAI (The Graduate University for Advanced Studies)

Statistical Learning by Quasi-linear Predictor

Katsuhiro Omae

Doctor of Philosophy

Department of Statistical Science

School of Multidisciplinary Sciences

SOKENDAI (The Graduate University for Advanced Studies)

2017

Abstract: In a biomedical study, a linear predictor is widely used as the regression and prediction functions because of the learnability and understandability. Many statistical models are formulated by the linear predictor. Fisher's linear discriminant analysis is a representative example of such formulations for a binary classification. It has Bayes risk consistency in the classification problem of two normal samples with the same variance. The linear form is thus useful in fitting for data with homogenous structure. However, especially in the field of biomedical science, it was revealed that there are several diseases once considered to be homotypic but later elucidated as heterotypic. Nevertheless, only the linear model is applied and the estimated model is used for the discussion in most biomedical studies. This fact may yield misleading for therapy efficacy by failing to reveal the potentially complex mixture of substantial benefits or harm. Such heterogeneous diseases do not allow only a single set of the biomarkers to be predictive for all patients. Moreover, the heterogeneity may obstruct us to detect the predictive biomarkers and to learn the predictive model. We therefore need to consider the predictor in consideration with the heterogeneous structure. Taking such a background into account, we derive the quasi-linear predictor defined as log-sum-exp form. It is a special case of the generalized average known as Kolmogorov-Nagumo average. The quasi-linear predictor is made up of the combination of some linear predictors with the different intercepts and coefficients. The shape of the quasi-linear predictor is determined not only by the parameters of these linear predictors but also by the tuning parameter for adjusting the overall nonlinearity. The quasi-linear predictor converges to minimum, maximum and linear predictor in the limiting sense for this tuning parameter. For analytical purpose, the tuning parameter is determined by Bayes information criteria from the learning dataset. It results in that the suitable non-linearity of the regression or prediction function is estimated by the data. In the thesis, we extend two ordinary models, linear logistic model and Cox's proportional hazard model to the quasi-linear logistic model and the quasi-linear relative risk model, respectively. The optimality of the quasi-linear logistic model is assured when we consider the binary classification problem of mixture normal and normal samples with equal variance for each component. In order to get more parsimonious model expression, we derive the restricted quasi-linear logistic model, which is defined

as the logistic model with the quasi-linear predictor, where each predictor is composed by known disjoint clusters of covariates. The restricted model gives easier interpretation for the estimated model compared to the regular quasi-linear model because each covariate is incorporated in only one linear predictor but not in the other linear predictors. Moreover, in case such clusters are unknown, we derive the quasi-linear logistic model with the cross-L_1 regularization. In this regularization method, we add a penalty for product of absolute values of coefficients for same covariate in different clusters to the likelihood function. Sufficiently higher cross-L_1 regularizations therefore results in the restricted quasi-linear model and the resultant disjoint sets of covariates are automatically given through the model estimation procedure. The second extension, the quasi-linear relative risk model, is regarded as the extension of a mixture hazard model. In fact, the mixture hazard model with same baseline hazard function corresponds to the quasi-linear relative risk model with a specific tuning parameter. As is the case with the quasi-linear logistic model, the quasi-linear relative risk model with the cross-L_1 regularization is derived. Because the extensions of these two models are performed simply by replace the linear predictor to the quasi-linear predictor, the other extensions for the linear model are easily combined and implemented. We derive the L_1 and L_2 penalized versions for both of the quasi-linear logistic and relative risk models. In the simulation study for the binary classification problem, we checked the consistency of the parameter estimation and compared the predictive performance between the linear logistic model and the restricted quasi-linear logistic model. In the application studies for the binary classification problem, we compared the performance among the restricted quasi-linear logistic model, linear logistic model and ordinary classification methods including decision tree, random forest, support vector machine, naive Bayes, group lasso, neural network, L_1 and L_2 penalized linear logistic models. These simulation and application studies show that the restricted quasi-linear logistic model has better performance in some simulated examples and real datasets than the ordinary methods. For the regression problem on the survival time data, we checked the true model selection probability by the Bayes information criteria and the parameter consistency for some situations in the simulation studies and compared the quasi-linear relative risk model with cross L_1 penalty and Cox's proportional hazard

model in the application studies. The simulation studies show that the parameter estimation empirically has the consistency and the selection of the tuning parameter by Bayes information criteria works very well. The application studies for the regression problem on the survival time data show that the quasi-linear relative risk model has better performance in some real datasets compared to the Cox's proportional hazard model. Finally, we discuss about the role of the quasi-linear predictor in traditional clustering method, and the relationship among the quasi-linear model, mixture of experts model and neural network model for more discussions and future works.

Contents

Chapter 1

Motivations

Statistical learning is usually divided into two popular types of learning: supervised and unsupervised learning. These strategies play important roles in biomedical research, not only in traditional setting but also in more complicated and higher dimensional setting (Foster *and others*, 2014) that many researchers are working on today. For example, Casanova *and others* (2011) applied a L_1 logistic regression model to high dimensional structural magnetic resonance imaging (MRI) data of brain image of Alzheimer patients for early detection of their Alzheimer's diseases. Because the structural MRI data has several hundreds of thousands voxels[1], machine learning methods for high dimensional data play essential roles in the analysis (Suzuki *and others*, 2017). Likewise, shrinkage methods by L_1 and L_2 regularizations have been frequently used in the context of prediction (Brimacombe, 2014). van't Veer *and others* (2002) used the cross validation method to optimize the combination of genes from a number of candidate genes to develop the predictor for breast cancer metastatic events. Amato *and others* (2013) discussed the effectiveness of using artificial neural network model in diagnosis and they stated that it would help physicians perform diagnosis of various diseases. The unsupervised learning methods also reveal some novel findings in biomedical science. For example, in the work of Sørlie *and others* (2001), hierarchical clustering algorithm was used to visualize the differentially expressed genes of breast carcinomas and find new subtypes without any class labels. Other than their work, clustering methods

[1]The voxel is a unit in which pixels are extended to three dimensions.

has been used in the context of interpretation (Oghabian *and others*, 2014).

In supervised learning, the primary objective is to learn a regression or prediction function from a training dataset whose true labels or outcomes are known. We seek a better regression function which accurately estimates an expectation of outcome measurement, or better prediction function which correctly predicts true labels consisting of binary or more multiple label indicators. In unsupervised learning, the goal is to get knowledge discovery including discovering clusters, discovering latent factors and discovering graph structure (Murphy, 2012). In these strategies, a wide variety of learning method can be regarded as a problem of either or both of how to construct a statistical model and what loss is considered for the learning. In the thesis, we extend linear logistic model and Cox's proportional hazard model, and give a unified discussion about K-means clustering and maximum entropy clustering.

The linear logistic model is one of the most frequently used models in binary classification problems. The general formulation of the logistic model was given by Cox (1958) although it had been already applied in some areas (Berkson, 1953, 1955). In this model, we assume that the log odds of an event are approximated by a linear predictor. Then, a binary response which indicates whether or not an event was occurred is regarded as following binomial distribution with the mean parameter of logistic transformation of the linear predictor. In this sense, the logistic model is a member of generalized linear model (*GLM*, Nelder and Wedderburn (1972)). Such a formulation gives an easy interpretation of the fitted linear predictor because the predictor means the log odds of the event (McCulloch *and others*, 2008). The parameter estimation on the linear logistic model is often performed by the Newton-Raphson method. It is easy to fit the model to the data because the iteratively reweighted least squares (*IRLS*), the special version of Newton-Raphson method, is easy to implement and working very fast (Murphy, 2012). The linear logistic model for binary classification is often compared with Fisher's linear discriminant analysis. It is well known that Fisher's linear discriminant rule has Bayes risk consistency if both of the covariate vectors conditioned on each binary response follow normal distribution with the equal variance, say normal-normal assumption. Accordingly, categorical variables should not be included in the

model even though these are commonly included in the real world data. Even if all variables are continuous, the performance of Fisher's linear discriminant rule drops sharply when the normal-normal assumption is violated. Consequently, the linear logistic model is preferred because it relaxes the normal-normal assumption of Fisher's linear discriminant rule. Moreover, even when the normal-normal assumption is correct, the logistic model still has not so much inferior performance to Fisher's linear discriminant analysis. Anyway, normal-normal assumption seems a little strict for specific situations. Especially in the field of biomedical science, it was revealed that there are several diseases once considered to be homotypic but later elucidated as heterotypic (Wallstrom *and others*, 2013). This fact may yield misleading for therapy efficacy by failing to reveal the potentially complex mixture of substantial benefits or harm (Kravitz *and others*, 2004). Many researchers have been focusing on such disease heterogeneity and discussed in biomarker researches (Di Camillo *and others*, 2012; Komori *and others*, 2013; Ein-Dor *and others*, 2005; Omae *and others*, 2016). Heterogeneous diseases do not allow only a single set of biomarkers to be predictive for all patients. Moreover, the heterogeneity may obstruct us to detect predictive biomarkers and to learn the predictive model. We therefore need to consider the predictor in consideration with the heterogeneous structure. Taking such a background into account, it seems that more natural assumption might be that the covariates of disease samples follow mixture distributions. In contrast with the normal-normal assumption, one of the simplest heterogeneous assumptions is described by the case that these are distributed with normal mixture distributions. In this setting, Bayes optimal classifier is no longer derived as a linear fashions, but rather we need to replace it to the log likelihood ratio of normal mixture distribution and normal distribution. Then the likelihood ratio is written by the probability-weighted mean of the exponential transformation of linear predictors. Accordingly, the log likelihood ratio results in the form of log-sum-exp. Thus we derived the logistic model with the log-sum-exp predictor that we call the quasi-linear predictor defined in Chapter 2. The quasi-linear logistic model is basically a logistic regression model, rather than the predictor is replaced by the log-sum-exp form of some linear predictors and so it is a non-linear model. Such simple extension does not lose learnability and understandability unlike any other non-linear

methods. In Omae *and others* (2017), which proposed the quasi-linear logistic model first, disjoint clusters of markers were specified by a hard clustering method in advance. The linear predictors by these clusters were combined as the quasi-linear predictor. In this manner, the quasi-linear logistic model could combine the unsupervised and supervised statistical learning in a natural way. The details are described in Chapter 4.

The motivation for the quasi-linear logistic model, the disease heterogeneity, leads us to the extension of the regression model. In the field of biomedical science, one of the most important models is the regression model for survival time data. Cox's proportional hazard model (Cox, 1972) is the most commonly used model in the regression problem of survival time analysis. The model is a member of relative risk models (Aalen *and others*, 2008) and it is a semi-parametric model which combines baseline hazard and regression functions as the non-parametric and parametric part, respectively. Cox (1972) proposed the maximum partial likelihood method for inference of the regression coefficients. A rigorous theory of behavior of the estimator in large sample setting was investigated by Tsiatis (1981). Anderson and Gill (1982) gave more natural derivations for mathematical theory of Cox's proportional hazard model by counting process and martingale theory (Aalen *and others*, 2008). Moreover, the derivation from the viewpoint of counting process produced martingale residuals as a natural evaluation in regression models of survival time (Barlow and Prentice, 1988) and the residuals enable us to evaluate the goodness-of-fit for fitted model. Such a model evaluation is very important because Cox's proportional hazard model needs a noncasual assumption. In fact, the model assumes that the log hazard is decomposed into the time-dependent term and time-independent linear predictor. Nevertheless, understandability of the model interpretation and the good performance for several applications has lead to widely use of Cox's proportional hazard model especially in the field of biomedical science. However, for the same reason with the quasi-liner logistic model discussed above, the sample heterogeneity should be included in building a hazard model in the specific situation. We note that a frailty model was developed by Vaupel and Yashin (1985) to describe unobserved heterogeneity, but even the observable heterogeneity might not be caught up by a linear predictor. In fact, the different markers for different population, which predict

the prognosis of each group well, were detected in the analysis of the gene expressions data (Wang *and others*, 2005). They used the label of different population to detect them while such meaningful labels are unknown in general. We thus may need the relative risk model for capturing heterogeneous structure. Although the mixture of hazard model were previously proposed and applied by Louzada-Neto *and others* (2002); Elmahdy and Aboutahoun (2013); Zhang *and others* (2014) in parametric setting and Rosen and Tanner (1999) in semi-parametric setting, they are in the restrictive setting and give no general formulations. Hence we develop the quasi-liner relative risk model to describe the observable heterogeneity and give the general formulation based on the relative risk model. In the special case, this model is equivalent to the relative risk model of Rosen and Tanner (1999) making assumption that the hazard rate forms mixture distribution. The details of derivation for the quasi-linear relative risk model are described in Chapter 5.

For these extensions, the key is in the form of log-sum-exp function. This is a member of Kolmogorov-Nagumo average (Eguchi and Komori, 2015). The log-sum-exp function is occasionally used in the field of machine-learning theory and its applications. First, it has been used as the computation technique to prevent the numerical calculations from overflows. The technique is called a log-sum-exp trick. Second, the softplus function (Belisle *and others*, 2002) is used as an activation function in the neural network model which is defined by log-sum-exp function. The limiting case of the softplus is rectified linear unit function proposed by Glorot *and others* (2011). Because the log-sum-exp form is considered to be an activation of the linear predictors which have large values, the quasi-linear model is very close to the neural network model. This fact is discussed in Chapter 6. Third, the fuzzy clustering proposed by Rose *and others* (1990) uses an energy function of log-sum-exp form. This class includes K-means clustering in a limiting sense. We discuss about more broad family by the generalized average form in Chapter 6.

The rest part of the thesis is organized as follows. The characteristics of the quasi-linear predictor and the relation of it to the linear predictor is discussed in Chapter 2. In this chapter, we also introduce the generalized linear predictor to discuss the role of it in the traditional settings in Chapter 6. We summarize notations of the GLM and relative risk

model in Chapter 3 in order to prepare for extensions of them by the quasi-linear predictor in the next two sections. In Chapter 4, we introduce the quasi-linear logistic model, which is extended logistic model by the quasi-linear predictor. To avoid the loss of the parameter identifiability and instability of the parameter estimation, we propose the restricted quasi-linear logistic model and cross-L_1 regularized quasi-linear logistic model. The performance of the restricted quasi-linear logistic model is evaluated by simulation studies and application studies to the real datasets. In Chapter 5, we introduce the quasi-linear relative risk model, which is extended Cox's proportional hazard model by the quasi-linear predictor. We discuss that it is very natural model because the quasi-linear relative risk model is also regarded as the extension of the mixture hazard model. The performance of the cross-L_1 penalized quasi-linear relative risk model is evaluated by simulation studies and application studies to the real datasets. In Chapter 6, we give the discussion of the connection between the quasi-linear predictor and traditional method including K-means clustering, maximum entropy clustering, mixture of experts model (Jordan and Jacobs, 1993) and neural network model (Rosenblatt, 1958; Hinton *and others*, 1984). We also give the comprehensive discussion and ongoing of the method of the quasi-linear predictor.

Chapter 2

Quasi-linear predictor

2.1 Definition and properties of quasi-linear predictor

Let $\boldsymbol{x}$ be a covariate vector. Then we define the quasi-linear predictor as the log-sum-exp average of K linear predictors:

$$F_\tau(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{\tau}\log\left(\frac{1}{K}\sum_{k=1}^{K}\exp(\tau\alpha_k + \tau\boldsymbol{\beta}_k^\top\boldsymbol{x})\right), \tag{2.1.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K)^\top$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \cdots, \boldsymbol{\beta}_K^\top)^\top$. Here, τ is a tuning parameter of non-zero real number, α_k is an intercept and $\boldsymbol{\beta}_k$ is a coefficient vector for the k-th linear predictor. Below, we may unify or omit the arguments and write $F_\tau(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta})$ simply as $F_\tau(\boldsymbol{x};\boldsymbol{\theta})$, $F_\tau(\boldsymbol{x})$, $F_\tau(\boldsymbol{\theta})$ or F_τ, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$. When $K = 1$, the quasi-linear predictor reduces to a linear predictor defined by

$$F(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = \alpha + \boldsymbol{\beta}^\top\boldsymbol{x}. \tag{2.1.2}$$

As shown in Proposition 1, a one-parameter family $\{F_\tau|\tau \in \mathbb{R}\backslash\{0\}\}$ connects between the minimum and maximum of linear predictors. We define them as $F_{-\infty}$ and F_∞:

$$F_{-\infty}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{1\le k\le K} F(\boldsymbol{x};\alpha_k,\boldsymbol{\beta}_k), \tag{2.1.3}$$

$$F_{\infty}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = \max_{1\le k\le K} F(\boldsymbol{x};\alpha_k,\boldsymbol{\beta}_k). \tag{2.1.4}$$

We show that the tuning parameter τ controls linearity of the quasi-linear predictor by the following proposition.

Proposition 1. *The following properties follow:*

1. $$F_{-\infty}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) \leq F_{\tau}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) \leq F_{\infty}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{2.1.5}$$

2. $$\lim_{\tau\to-\infty} F_{\tau}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = F_{-\infty}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{2.1.6}$$

3. $$\lim_{\tau\to 0} F_{\tau}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{K}\sum_{k=1}^{K} F(\boldsymbol{x};\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k), \tag{2.1.7}$$

4. $$\lim_{\tau\to\infty} F_{\tau}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = F_{\infty}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{2.1.8}$$

The proofs of Proposition 1 are given in Appendix A.1. These characteristics of the quasi-linear predictor are visualized in Figure 2.1. We note that the average of linear predictors (2.1.7) is regarded as the linear predictor because

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K} F(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) &= \frac{1}{K}\sum_{k=1}^{K}(\alpha_k + \boldsymbol{\beta}_k^{\top}\boldsymbol{x}) \\
&= \frac{1}{K}\sum_{k=1}^{K}\alpha_k + \left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\beta}_k^{\top}\right)\boldsymbol{x} \\
&= F\left(\boldsymbol{x}; \frac{1}{K}\sum_{k=1}^{K}\alpha_k, \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\beta}_k\right).
\end{aligned} \tag{2.1.9}$$

Hence the quasi-linear predictor approaches the linear predictor as τ goes to 0. In this sense, the parameter family of the quasi-linear predictors is a broader class than the sets of all linear predictors. The tuning parameter τ controls the non-linearity and the linear predictor corresponds to take simple averaging whereas the quasi-linear predictor reflects the predictor with a larger (if $\tau > 1$) or a smaller (if $\tau < -1$) value more. For example,

$$\frac{1}{10}\log\left(\frac{1}{2}\left(\exp(10\cdot 20) + \exp(10\cdot 5)\right)\right) = 19.93 \approx 20$$

and

$$\frac{1}{-10}\log\left(\frac{1}{2}(\exp(-10\cdot 20) + \exp(-10\cdot 5))\right) = 5.069 \approx 5.$$

Thus the parameter family $\{F_{\tau}; -\infty < \tau < \infty, \tau \neq 0\}$ includes the minimum, mean and

maximum functions of K values. The nonlinearity of the quasi-linear predictor can be adjusted by the tuning parameter. The case that τ is equal to 1 has a special meaning from the perspective of Bayes risk consistency in the quasi-linear logistic model as discussed in Chapter 3. It enables us to seek the proper non-linearity for a given dataset. The tuning of τ is discussed in Chapter 5.

We need the derivatives of the quasi-linear predictor by parameter $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$ to investigate the nature of the predictor and build algorithms for parameter estimation of the model developed later chapters. The first order derivative with respect to parameter $\boldsymbol{\theta}$ of the quasi-linear predictor is given as

$$\frac{\partial F_\tau(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (\eta_1, \cdots, \eta_K, \boldsymbol{x}^\top \eta_1, \cdots, \boldsymbol{x}^\top \eta_K)^\top, \tag{2.1.10}$$

where

$$\eta_k(\boldsymbol{x}; \tau\boldsymbol{\alpha}, \tau\boldsymbol{\beta}) = \frac{\exp(\tau\alpha_k + \tau\boldsymbol{\beta}_k^\top \boldsymbol{x})}{\sum_{\ell=1}^K \exp(\tau\alpha_\ell + \tau\boldsymbol{\beta}_\ell^\top \boldsymbol{x})} \tag{2.1.11}$$

is called a softmax function (Murphy, 2012). Consider the case when $K = 2, \tau = 1$ and $x \in \mathbb{R}$. Then, the first derivative is written as $(\eta_1, \eta_2, \eta_1 x, \eta_2 x)$. The softmax η_1 and η_2 are in the relationship of seesaw while maintaining $\eta_1 + \eta_2 = 1$. If the linear predictor of the first cluster has extremely higher value than the second cluster, the derivative is approximated by $(1, 0, x, 0)$. This means that the surface of the quasi-linear predictor is almost approximated by a linear surface in the local area. The example in the case of $\boldsymbol{x} = (x_1, x_2)^\top \in \mathbb{R}^2, \alpha_1 = \alpha_2 = 1, \boldsymbol{\beta}_1 = (1, 0)^\top$ and $\boldsymbol{\beta}_2 = (0, 1)^\top$ is shown in Figure 2.1, which shows that the local area $\{\boldsymbol{x} : |(\alpha_1 + \beta_1 x_1) - (\alpha_2 + \beta_2 x_2)| \gg 0\}$ is approximated by the linear surface.

The second order derivative with respect to parameter $\boldsymbol{\theta}$ of the quasi-linear predictor is

given as

$$\frac{\partial^2 F_\tau(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{pmatrix} 1 & \boldsymbol{x}^\top \\ \boldsymbol{x} & \tau \boldsymbol{x}\boldsymbol{x}^\top \end{pmatrix} \otimes \begin{pmatrix} \eta_1(1-\eta_1) & -\eta_1\eta_2 & \cdots & -\eta_1\eta_K \\ -\eta_2\eta_1 & \eta_2(1-\eta_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ -\eta_K\eta_1 & \cdots & \cdots & \eta_K(1-\eta_K) \end{pmatrix}, \quad (2.1.12)$$

where $\otimes$ denotes a Kronecker product. While the linear predictor has the zero-matrix, the quasi-linear predictor has non-zero matrix as the second derivative.

The quasi-linear predictor is defined by the form of log-sum-exp. The form has been used for the computation technique to prevent the numerical calculations from overflows or underflows. Consider an example of calculation on $\exp(x) + \exp(y)$. To prevent extremely large values of $\exp(x)$ and $\exp(y)$ from the overflows, we often take logarithms of such values: $\log(\exp(x) + \exp(y))$. However, when x or y itself has a large value, the returned value would be log(Inf) and the overflow occurs. It can be avoided by taking calculation as $\log(\exp(x)(1+\exp(y-x)) = x + \log(1+\exp(y-x))$ expecting that $y-x$ is an enough small value to calculate the exponential. The trick is also efficient against the numerical problem of the quasi-linear predictor for all methods derived in the thesis.

2.2 Generalized quasi-linear predictor

As shown in the previous section, the quasi-linear predictor includes the linear predictor. Moreover, we can consider broader class. In fact, the log-sum-exp average is a member of the generalized average, called Kolmogorov-Nagumo average (Kolmogorov, 1930; Nagumo, 1930), defined as

$$G(\boldsymbol{z}) = \phi^{-1}\left(\frac{1}{K}\sum_{k=1}^{K}\phi(z_k)\right), \quad (2.2.1)$$

where $\boldsymbol{z} = (z_1, z_2, \cdots, z_K)$ and ϕ is any invertible real-valued function. One of the most popular examples is power mean: $\phi(z) = z^p$. For $p \neq 0$ and $z_k > 0$ for $1, 2, \cdots, K$, it is

defined as

$$G_{\text{power}}(\boldsymbol{z}) = \left(\frac{1}{K} \sum_{k=1}^{K} z_k^p \right)^{1/p}. \tag{2.2.2}$$

This is very close to the quasi-linear predictor because the same properties shown in Proposition 1 follow for tuning parameter p. It also includes harmonic ($p = -1$) and geometric ($p \to 0$) mean. Although we cannot use the power mean for combining linear predictors directly because of the limited support, it is useful form in the extension of the clustering method based on the distance as discussed in Chapter 6.

Next, we introduce the quasi-linear predictor with a cumulative distribution function Φ as ϕ defined by

$$C(z) = \frac{1}{\tau} \Phi^{-1} \left(\frac{1}{K} \sum_{k=1}^{K} \Phi(\tau z_k) \right), \tag{2.2.3}$$

Because the cumulative distribution function gives an non-negative value, it plays a similar role to the power mean. For example, consider an exponential distribution function as Φ. In this case, we get that $\Phi(z) = 1 - \exp(-z)$ and $\Phi^{-1}(z) = -\log(1 - z)$. Then

$$\begin{aligned} C_{exp}(z) &= -\frac{1}{\tau} \log \left\{ 1 - \frac{1}{K} \sum_{k=1}^{K} (1 - \exp(-\tau z_k)) \right\} \\ &= -\frac{1}{\tau} \log \left\{ \frac{1}{K} \sum_{k=1}^{K} \exp(-\tau z_k) \right\}. \end{aligned} \tag{2.2.4}$$

We see that $C_{exp}(z) = -F_\tau(-z_k)$ and it is closely related to the maximum entropy clustering introduced in Section 6.

For the other example, consider a logistic distribution function as Φ. In this case, we get that $\Phi(z) = 1/(1 + \exp(-z))$ and $\Phi^{-1}(z) = \log(z/(1 - z))$. Then

$$\begin{aligned} C_{logit}(z) &= \frac{1}{\tau} \log \frac{\frac{1}{K} \sum_{k=1}^{K} \frac{1}{1+\exp(-\tau z_k)}}{1 - \frac{1}{K} \sum_{k=1}^{K} \frac{1}{1+\exp(-\tau z_k)}} \\ &= \frac{1}{\tau} \log \frac{\frac{1}{K} \sum_{k=1}^{K} \frac{1}{1+\exp(-\tau z_k)}}{\frac{1}{K} \sum_{k=1}^{K} \frac{\exp(-\tau z_k)}{1+\exp(-\tau z_k)}} \end{aligned} \tag{2.2.5}$$

If $\tau = 1$, C_{logit} is written by the difference between two log-sum-exp averages

$$\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{1+\exp(-z_k)}\right) - \log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\exp(-z_k)}{1+\exp(-z_k)}\right). \tag{2.2.6}$$

We see that the average (2.2.6) makes sense when we consider K logistic models. Each k-th component of both terms of (2.2.6) is regarded as the estimated odds π_k in the k-th logistic model. The average (2.2.6) combines these odds as $\log(\frac{1}{K}\sum_{k=1}^{K}\pi_k) - \log(\frac{1}{K}\sum_{k=1}^{K}(1-\pi_k))$. We note that $\frac{1}{K}\sum_{k=1}^{K}\pi_k + \frac{1}{K}\sum_{k=1}^{K}(1-\pi_k) = 1$ and therefore (2.2.6) gives summarized the odds ratio of multiple logistic models. Besides this, Normal distribution function would give us similar interpretation while it recalls the Probit regression model not the logistic model.

By introducing the generalized quasi-linear predictor, the roles of the quasi-linear form in the traditional clustering method as the basis of the energy functions are easily extended, including K-means and maximum entropy clustering as discussed in Chapter 6. Other than it, the method extended by the quasi-linear predictor in the thesis can be further extended broader class by the generalized quasi-linear predictor. The details are discussed in Chapter 6.

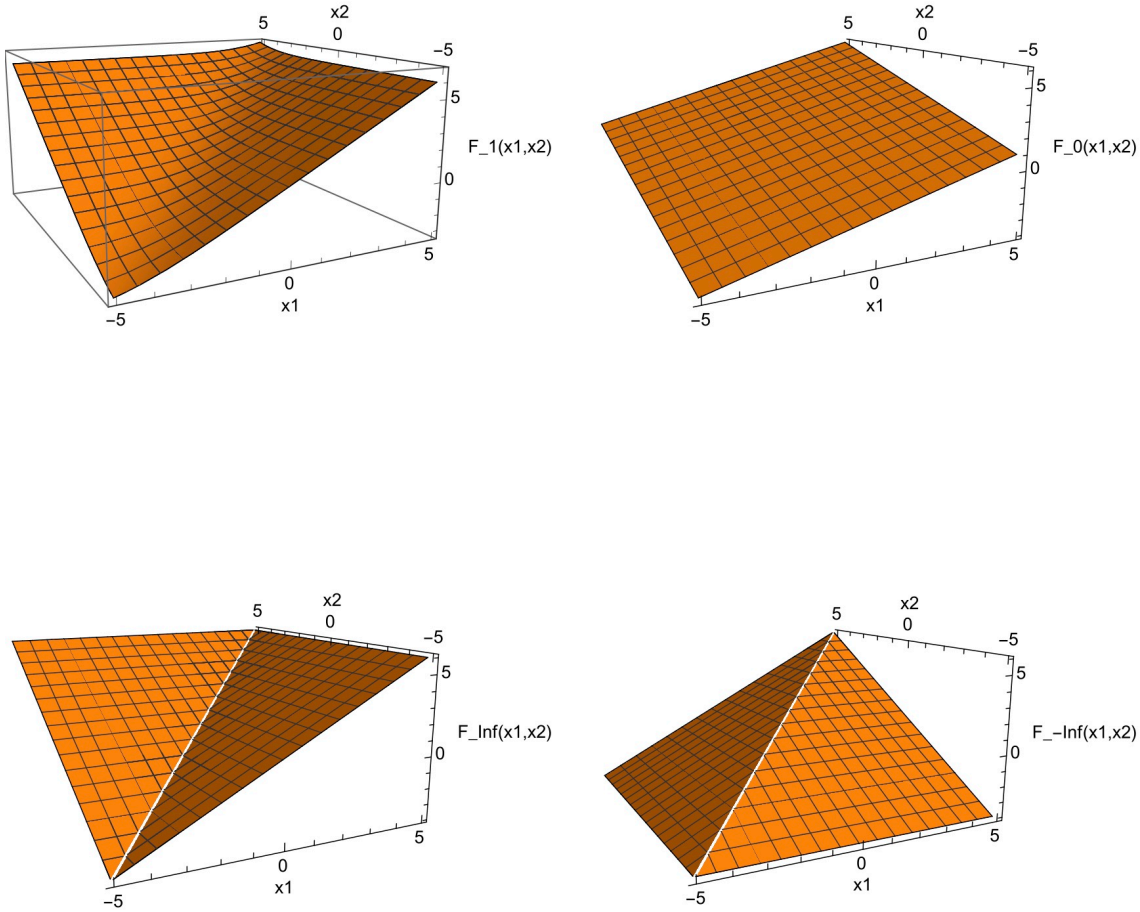

Figure 2.1: The contour of the quasi-linear predictor for two dimensional settings. The parameters are set as follows: $\boldsymbol{\alpha} = (1, 1)^\top$, $\boldsymbol{\beta} = (1, 0, 0, 1)^\top$. The top left and right figures show the surfaces of the quasi-linear predictor when $\tau = 1$ and the linear predictor, respectively. The bottom left and right figures show the surfaces of the minimum and maximum functions, respectively.

Chapter 3

Existing methods

In this chapter, we summarize the traditional two models, which are extended by the quasi-linear predictor in the thesis. In Section 3.1, we give the framework of the generalized linear model. In Section 3.2, we give the framework of the relative risk model, which includes Cox's proportional hazard model.

3.1 Generalized linear model

In Chapter 4, we extend the logistic model by the quasi-linear predictor. It is expected sufficiently to deal with the data which has heterogeneous structure as shown by Bayes risk consistency of the normal mixture-normal assumption. We therefore give the basic modeling notation of the logistic model via the framework of the generalized linear model. We note that while we focus only on the logistic model in the thesis, the generalized linear model is similarly extended as the logistic model.

In the framework of generalized linear model, the linear predictor, defined as the function of covariates vector $\boldsymbol{X}$, is connected to the expectation of outcome Y by a link function g as $\mathrm{E}[Y|\boldsymbol{X}] = g^{-1}(F(\boldsymbol{X};\boldsymbol{\alpha},\boldsymbol{\beta}))$. Here, it is assumed that Y has a distribution in the exponential family (McCullagh and Nelder, 1989). The parameter estimation is often performed by minimizing negative log-likelihood function for parameters. The distribution function of

exponential family is taking the form

$$f(Y;\theta,\phi) = \exp\left(\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)\right), \tag{3.1.1}$$

where $a(\phi)$, $b(\theta)$ and $c(Y,\phi)$ is a known function. The parameter θ is called a canonical parameter and ϕ is called a dispersion parameter. If the data $\{Y_1, Y_2, \cdots, Y_N\}$ is given, the log-likelihood function is written as

$$l(\theta) = \sum_{i=1}^{N}\left(\frac{Y_i\theta - b(\theta)}{a(\phi)} + c(Y_i,\phi)\right). \tag{3.1.2}$$

The moments and derivatives are given by simple notations for the exponential family. For example, $\mathrm{E}[Y] = \dot{b}(\theta)$, $\mathrm{var}[Y] = \ddot{b}(\theta)a(\phi)$. This property leads to calling $b(\boldsymbol{\theta})$ cumulant function. Furthermore,

$$\frac{\partial l}{\partial \theta} = \frac{Y - \dot{b}(\theta)}{a(\phi)}, \tag{3.1.3}$$

$$\frac{\partial^2 l}{\partial \theta} = -\frac{\ddot{b}(\theta)}{a(\phi)}, \tag{3.1.4}$$

where a dot and two dots denote the first and second derivative with respect to θ, respectively. If link function g is defined as $g = \dot{b}^{-1}$, then $g(\mathrm{E}[Y|\boldsymbol{X}]) = \theta = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X}$. Thus the canonical parameter and the linear predictor are directly connected, and so such link function is called a canonical link function. Some examples of the generalized linear models are given below.

- Normal

  If $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$ and $c(Y,\theta) = -(Y^2/\phi + \log(2\pi\phi))/2$, then

  $$f(Y;\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(Y-\theta)^2}{2\sigma^2}\right\}. \tag{3.1.5}$$

  The generalized linear model for (3.1.5) is known as the linear model (McCulloch *and others*, 2008).

- Binomial

If $\phi = 1$, $a(\phi) = \phi$, $b(\theta) = \log(1 + \exp(\theta))$ and $c(Y, \theta) = 0$, then

$$f(Y; \theta) = \frac{\exp(Y\theta)}{1 + \exp(\theta)}. \tag{3.1.6}$$

The generalized linear model for (3.1.6) is known as the linear logistic regression model (McCulloch *and others*, 2008).

- Poisson

  If $\phi = 1$, $a(\phi) = \phi$, $b(\theta) = \exp(\theta)$ and $c(Y, \phi) = -\log Y!$, then

$$f(y; \theta) = \frac{(\exp(\theta))^Y}{Y!} \exp(-\exp(\theta)). \tag{3.1.7}$$

  The generalized linear model for (3.1.7) is known as the poisson regression model (McCulloch *and others*, 2008). The model is linked to Cox's proportional hazard model (Cox, 1972) as discussed in Section 3.2.

The parameter estimation of the generalized linear model is performed by the gradient method or Newton-Raphson method. The case of the logistic model is noted in Section 4.

3.2 Relative risk model

In Chapter 5, we extend Cox's proportional hazard model by the quasi-linear predictor with expectation that the extended model could capture heterogeneous structure similar to the quasi-linear logistic model. The counting process and martingale theory is very useful to evaluate the characteristics of Cox's proportional hazard model. We give only the simple notation in this section and put more rigorous discussion in Appendix A.2.

In Cox's proportional hazard model, log hazard and covariates are connected via a linear predictor. Let $N_i(t)$ be number count of occurrences of the event of interest for individual $i \in \{1, \cdots, N\}$. For individual i and some time point t, $N_i(t) = 1$ if by time t the event has been observed to occur, otherwise $N_i(t) = 0$. Let $\boldsymbol{X}_i$ be p-dimensional covariates vector.

Assume that the intensity process of N_i is written as

$$A_i(t) = Y_i(t)h(t|\boldsymbol{X}_i). \tag{3.2.1}$$

Here $h(t|\boldsymbol{X}_i)$ is called the hazard rate, and $Y_i(t)$ is an indicator taking the value 1 if individual i is at risk for the event of interest just before time t and the value 0 otherwise. If t_i denotes the time to event for individual i, $Y_i(t) = \mathbf{1}(t \leq t_i)$, where $\mathbf{1}(\cdot)$ is an indicator function.

The relative risk model is written as

$$h(t|\boldsymbol{X}_i, \boldsymbol{\theta}) = h_0(t)r(\boldsymbol{X}_i, \boldsymbol{\theta}), \tag{3.2.2}$$

where $r(\boldsymbol{X}_i, \boldsymbol{\theta})$ is a relative risk function with parameter $\boldsymbol{\theta}$, and $h_0(t)$ is called the baseline hazard function. We often normalize the $r(\boldsymbol{X}_i, \boldsymbol{\theta})$ as

$$r(\mathbf{0}, \boldsymbol{\theta}) = 1. \tag{3.2.3}$$

We give examples of the relative risk model proposed to date.

- Exponential relative risk (Cox's proportional hazard) model

$$r(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left(F(\boldsymbol{x}; \boldsymbol{\theta})\right) \tag{3.2.4}$$

- Linear relative risk model

$$r(\boldsymbol{x}, \boldsymbol{\theta}) = F(\boldsymbol{x}; \alpha = 1, \boldsymbol{\beta}) \tag{3.2.5}$$

- Excess relative risk model

$$r(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{j=1}^{p} F(x_j; \alpha = 1, \beta_j) \tag{3.2.6}$$

Combining (3.2.1) and (3.2.2), We get

$$A_i(t) = Y_i(t)h_0(t)r(\boldsymbol{X}_i, \boldsymbol{\theta}). \tag{3.2.7}$$

The likelihood function of the survival data is defined by (3.2.7). To show this we define the conditional probability

$$p_i = \frac{A_i(t)}{\sum_{\ell=1}^{n} A_\ell(t)}. \tag{3.2.8}$$

This is the probability that an individual experiences an event at time t, given that one of the individuals in the risk set experiences an event at this time. The likelihood is equivalent to the product of these probability over all individual who experiences an event during the observation as equation (3.2.10). We derive it from the perspective of the partial likelihood function.

Let T be a continuous random variable which indicates a survival time with a distribution function $G(t)$. We denote the density function, survival function and hazard function of T as $g(t)$, $S(t)$ and $h(t)$. Immediate consequence of these definitions result in their relationships as $g(t) = \partial G(t)/\partial t$, $S(t) = 1 - G(t)$, $h(t) = g(t)/S(t)$.

Consider the data $(\boldsymbol{X}_i, t_i, \delta_i)$ $(i = 1, 2, \cdots, N)$, where t_i is observed survival time to some event and δ_i is an event indicator which takes value 1 if the individual experiences an event by $t = t_i$ and value 0 otherwise. We assume that t_i and δ_i are independent for all

individuals. Then, the full likelihood of $\boldsymbol{\theta}$ is written as

$$\begin{aligned}
L^{full} &= \prod_{i=1}^{N} g(t_i|\boldsymbol{X}_i,\boldsymbol{\theta})^{\delta_i} S(t_i|\boldsymbol{X}_i,\boldsymbol{\theta})^{1-\delta_i} \\
&= \prod_{i=1}^{N} h(t_i|\boldsymbol{X}_i,\boldsymbol{\theta})^{\delta_i} S(t_i|\boldsymbol{X}_i,\boldsymbol{\theta}) \\
&= \prod_{i=1}^{N} \{h_0(t_i) r(\boldsymbol{X}_i,\boldsymbol{\theta})\}^{\delta_i} \exp\left\{-\int_0^{t_i} A_i(u)\right\} \\
&= \prod_{i=1}^{N} \{h_0(t_i) r(\boldsymbol{X}_i,\boldsymbol{\theta})\}^{\delta_i} \prod_{i=1}^{N} \exp\left\{-\int_0^{\infty} Y_i(u) h_0(u) r(\boldsymbol{X}_i,\boldsymbol{\theta}) du\right\} \\
&= \prod_{i=1}^{N} \left\{\frac{r(\boldsymbol{X}_i,\boldsymbol{\theta})}{\sum_{\ell=1}^{N} Y_\ell(t_i) r(\boldsymbol{X}_\ell,\boldsymbol{\theta})}\right\}^{\delta_i} \prod_{i=1}^{N} \left\{h_0(t_i)\left(\sum_{\ell=1}^{N} Y_\ell(t_i) r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\right)\right\}^{\delta_i} \\
&\quad \exp\left\{-\int_0^{\infty} \sum_{\ell=1}^{N} Y_\ell(u) r(\boldsymbol{X}_\ell,\boldsymbol{\theta}) h_0(u) du\right\}. \qquad (3.2.9)
\end{aligned}$$

The inference for the parameter θ is often based on a partial likelihood defined by

$$L = \prod_{i=1}^{N} \left\{\frac{r(\boldsymbol{X}_i,\boldsymbol{\theta})}{\sum_{\ell=1}^{N} Y_\ell(t_i) r(\boldsymbol{X}_\ell,\boldsymbol{\theta})}\right\}^{\delta_i}, \qquad (3.2.10)$$

rather than the full likelihood (3.2.9) (Klein and Moeschberger, 2003). The maximum partial likelihood estimator has large sample properties similar to the maximum likelihood estimator (Aalen *and others*, 2008).

Practically, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is got by maximizing the log partial likelihood defined as

$$l = \sum_{i=1}^{N} \delta_i \left\{\log\left(r(\boldsymbol{X}_i,\boldsymbol{\theta})\right) - \log\left(\sum_{\ell=1}^{N} Y_\ell(t_i) r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\right)\right\}. \qquad (3.2.11)$$

The estimator is the solution of the likelihood equation: $U(\boldsymbol{\theta}) = 0$, where $U(\boldsymbol{\theta})$ is defined as the first derivative of log likelihood function:

$$U(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i \left\{\frac{\frac{\partial}{\partial\boldsymbol{\theta}} r(\boldsymbol{X}_i,\boldsymbol{\theta})}{r(\boldsymbol{X}_i,\boldsymbol{\theta})} - \frac{\sum_{\ell=1}^{N} Y_\ell(t_i) \frac{\partial}{\partial\boldsymbol{\theta}} r(\boldsymbol{X}_\ell,\boldsymbol{\theta})}{\sum_{\ell=1}^{N} Y_\ell(t_i) r(\boldsymbol{X}_\ell,\boldsymbol{\theta})}\right\}. \qquad (3.2.12)$$

By the counting process and martingale theory, we can show that $\hat{\boldsymbol{\theta}}$ is asymptotically multivariate normally distributed around the true value $\boldsymbol{\theta}_0$ with a covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is estimated by an inverse matrix of the observed information matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$ or expected information matrix $\mathcal{F}(\hat{\boldsymbol{\theta}})$. The difference between these matrixes is as follows. Let $r^{(0)}(\boldsymbol{X}_\ell, \boldsymbol{\theta}) = r(\boldsymbol{X}_\ell, \boldsymbol{\theta})$, $r^{(1)}(\boldsymbol{X}_\ell, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} r(\boldsymbol{X}_\ell, \boldsymbol{\theta})$, $r^{(2)}(\boldsymbol{X}_\ell, \boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} r(\boldsymbol{X}_\ell, \boldsymbol{\theta})$ and

$$S^{(m)}(\boldsymbol{\theta}, t) = \sum_{\ell=1}^{N} Y_\ell(t) r^{(m)}(\boldsymbol{X}_\ell, \boldsymbol{\theta}) \tag{3.2.13}$$

for $m = 0, 1, 2$. Then, the observed Fisher information of $\boldsymbol{\theta}$ is

$$\begin{aligned} I(\boldsymbol{\theta}) &= -\sum_{i=1}^{N} \delta_i \left\{ \left(\frac{\frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} r(\boldsymbol{X}_i, \boldsymbol{\theta})}{r(\boldsymbol{X}_i, \boldsymbol{\theta})} - \left\{ \frac{\frac{\partial}{\partial \boldsymbol{\theta}} r(\boldsymbol{X}_i, \boldsymbol{\theta})}{r(\boldsymbol{X}_i, \boldsymbol{\theta})} \right\} \left\{ \frac{\frac{\partial}{\partial \boldsymbol{\theta}} r(\boldsymbol{X}_i, \boldsymbol{\theta})}{r(\boldsymbol{X}_i, \boldsymbol{\theta})} \right\}^\top \right) \right. \\ &- \left. \left(\frac{S^{(2)}(\boldsymbol{\theta}, t_i)}{S^{(0)}(\boldsymbol{\theta}, t_i)} - \left(\frac{S^{(1)}(\boldsymbol{\theta}, t_i)}{S^{(0)}(\boldsymbol{\theta}, t_i)} \right) \left(\frac{S^{(1)}(\boldsymbol{\theta}, t_i)}{S^{(0)}(\boldsymbol{\theta}, t_i)} \right)^\top \right) \right\}. \end{aligned} \tag{3.2.14}$$

The expected Fisher information matrix is derived by predictable variation process of the score function. This is written as

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i \left\{ \frac{S^{(2)}(\boldsymbol{\theta}, t_i)}{S^{(0)}(\boldsymbol{\theta}, t_i)} - \left(\frac{S^{(1)}(\boldsymbol{\theta}, t_i)}{S^{(0)}(\boldsymbol{\theta}, t_i)} \right) \left(\frac{S^{(1)}(\boldsymbol{\theta}, t_i)}{S^{(0)}(\boldsymbol{\theta}, t_i)} \right)^\top \right\} \tag{3.2.15}$$

In the case of Cox's proportional hazard model, (3.2.14) is equal to (3.2.15). In fact, we get that

$$\begin{aligned} \frac{\frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} r(\boldsymbol{X}_i, \boldsymbol{\theta})}{r(\boldsymbol{X}_i, \boldsymbol{\theta})} - \left\{ \frac{\frac{\partial}{\partial \boldsymbol{\theta}} r(\boldsymbol{X}_i, \boldsymbol{\theta})}{r(\boldsymbol{X}_i, \boldsymbol{\theta})} \right\} \left\{ \frac{\frac{\partial}{\partial \boldsymbol{\theta}} r(\boldsymbol{X}_i, \boldsymbol{\theta})}{r(\boldsymbol{X}_i, \boldsymbol{\theta})} \right\}^\top &= \frac{\boldsymbol{X}_i \boldsymbol{X}_i^\top \exp(\boldsymbol{\theta}^\top \boldsymbol{X}_i)}{\exp(\boldsymbol{\theta}^\top \boldsymbol{X}_i)} - \boldsymbol{X}_i \boldsymbol{X}_i^\top \\ &= 0. \end{aligned}$$

However, these may differ for other models. In these case, the expected information would be more stable than the observed one because (3.2.14) depends on each covariate values of the individuals (Aalen *and others*, 2008). Cox's proportional hazard model has a deep connection with Poisson regression model as discussed in McCullagh and Nelder (1989). In fact, it follows that $G(t) = 1 - \exp(-\int_{-\infty}^{t} h_0(s) ds \exp(F(\boldsymbol{X})))$, $S(t) = 1 - G(t) =$

$\exp(-\int_{-\infty}^{t} h_0(s)ds\exp(F(\boldsymbol{X})))$ and $g(t) = h_0(t)\exp(F(\boldsymbol{X})-\int_{-\infty}^{t} h_0(s)ds\exp(F(\boldsymbol{X})))$. Then, the log full-likelihood of Cox's proportional hazard model is written as

$$\begin{aligned}
l^{full} &= \sum_{i=1}^{N} \delta_i \log(g(t_i)) + (1-\delta_i)\log(S(t_i)) \\
&= \sum_{i=1}^{N} \delta_i(\log(h_0(t_i)) + F(\boldsymbol{X}_i)) - \int_{-\infty}^{t_i} h_0(s)ds\exp(F(\boldsymbol{X}_i)) \\
&= \sum_{i=1}^{N} \delta_i \left(\log\left(\int_{-\infty}^{t_i} h_0(s)ds\right) + F\left(\boldsymbol{X}_i\right)\right) - \int_{-\infty}^{t_i} h_0(s)ds\exp\left(F\left(\boldsymbol{X}_i\right)\right) + \delta_i \log\left(\frac{h_0(t_i)}{\int_{-\infty}^{t_i} h_0(s)ds}\right) \\
&= \sum_{i=1}^{N} (\delta_i \log\mu_i - \mu_i) + \sum_{i=1}^{N} \delta_i \log\left(\frac{h_0(t_i)}{\int_{-\infty}^{t_i} h_0(s)ds}\right), \qquad (3.2.16)
\end{aligned}$$

where $\mu_i = \int_{-\infty}^{t_i} h_0(s)ds\exp(F(\boldsymbol{X}_i))$. The log full-likelihood is equivalent to the likelihood of Poisson model with log-linear function when regarding the censoring label δ_i as Poisson distributed with mean μ_i.

Chapter 4

Quasi-linear logistic model

In this chapter, we extend the linear logistic model by the quasi-linear predictor. The optimality of the quasi-linear predictor is assured by the Bayes optimal form in the case of binary classification between normal and mixture-normal samples. In Section 4.1, we give two decision rules in binary classification problem and discuss the optimal setting for normal-normal assumption. In Section 4.2, we derive the quasi-linear logistic model. In Section 4.3, we introduce the restricted quasi-linear logistic model and the numerical calculation method on parameter estimation for the penalized quasi-linear logistic model including cross-L_1, L_1 and L_2 penalties. In Section 4.4 and 4.5, we give simulation and application studies. We close Section 4.6 with the discussion on the quasi-linear logistic model.

4.1 Statistical classification

4.1.1 Statistical decision

Consider a classification problem for binary classes. Let $Y \in \{0, 1\}$ be a class indicator. In accordance with the convention, $Y = 0$ and $Y = 1$ denotes a normal and disease label, respectively. We have a continuous random vector $\boldsymbol{X}$ which has marginal density $P(\boldsymbol{X})$, and we wish to decide from which classes $\boldsymbol{X}$ generated. In the statistical decision theory, there are two most popular rules: Bayes decision rule and Neyman-Pearson decision rule (Webb and Copsey, 2011). These two rules are based on three distributions, namely a prior

distribution: $P(Y)$, a posterior distribution: $P(Y|\boldsymbol{X})$ and a class conditional distribution: $P(\boldsymbol{X}|Y)$.

4.1.2 Bayes decision rule

The Bayes decision rule for the binary classification problem is denoted by two posterior distributions. It is denoted as follows:

$$\begin{cases} P(Y=1|\boldsymbol{X}) \geq P(Y=0|\boldsymbol{X}) & \Rightarrow \quad \widehat{Y}=1, \\ P(Y=1|\boldsymbol{X}) < P(Y=0|\boldsymbol{X}) & \Rightarrow \quad \widehat{Y}=0, \end{cases} \tag{4.1.1}$$

where $\widehat{Y}$ is an estimated class indicator. It is an intuitive natural rule in that the rule simply classify the random sample $\boldsymbol{X}$ as the class with higher conditional probability. By the Bayes' theorem, $P(Y|\boldsymbol{X})$ is written as

$$P(Y=y|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|Y=y)P(Y=y)}{P(\boldsymbol{X})} \quad (y=0,1) \tag{4.1.2}$$

so that the decision rule (4.1.1) is equivalent to the rule

$$\begin{cases} \mathrm{LR}(\boldsymbol{X}) \geq \frac{P(Y=0)}{P(Y=1)} & \Rightarrow \quad \widehat{Y}=1, \\ \mathrm{LR}(\boldsymbol{X}) < \frac{P(Y=0)}{P(Y=1)} & \Rightarrow \quad \widehat{Y}=0, \end{cases} \tag{4.1.3}$$

where $\mathrm{LR}(\boldsymbol{X}) = \frac{P(\boldsymbol{X}|Y=1)}{P(\boldsymbol{X}|Y=0)}$. The class conditional probability $P(\boldsymbol{X}|Y)$ is the probability that the data X is obtained when a class label Y is given and it is called the likelihood. The function $\mathrm{LR}(\boldsymbol{X})$ is therefore called the likelihood ratio.

4.1.3 Neyman-Pearson decision rule

An alternative of the Bayes decision rule is the Neyman-Pearson decision rule. In the statistical decisions, we may have two errors:

$$\epsilon_1 = \int_{\Omega_1} P(\boldsymbol{x}|Y=0)d\boldsymbol{x}, \tag{4.1.4}$$

$$\epsilon_2 = \int_{\Omega_0} P(\boldsymbol{x}|Y=1)d\boldsymbol{x}, \tag{4.1.5}$$

where Ω_y is a region such that if $\boldsymbol{x} \in \Omega_y$ then $\boldsymbol{x}$ belongs to class y. The two errors ϵ_1 and ϵ_2 are called type I error and type II error, respectively. The type I error is the error probability of classifying the sample from class 0 as the class 1, and the type II error is the contrast. In particular for the classification problem, the type I and type II error is called a false positive and negative rate, respectively. The Neyman-Pearson decision rule is derived as the decision boundary which minimizes the ϵ_2 subject to ϵ_1 being equal to a constant error ϵ_0. Such minimizing problem is described by method of Lagrange multiplier as

$$\begin{aligned} & \underset{\Omega_1}{\operatorname{argmin}} \int_{\Omega_1^C} P(\boldsymbol{x}|Y=1)d\boldsymbol{x} + \lambda \left\{ \int_{\Omega_1} P(\boldsymbol{x}|Y=0)d\boldsymbol{x} - \epsilon_0 \right\} \\ = & \underset{\Omega_1}{\operatorname{argmin}} \, (1-\lambda\epsilon_0) + \int_{\Omega_1} \lambda P(\boldsymbol{x}|Y=0) - P(\boldsymbol{x}|Y=1)d\boldsymbol{x}, \end{aligned} \tag{4.1.6}$$

where λ is a Lagrange multiplier. The solution is

$$\begin{aligned} \Omega_1 &= \{\boldsymbol{x} \mid \lambda P(\boldsymbol{x}|Y=0) - P(\boldsymbol{x}|Y=1) \leq 0\} \\ &= \left\{ \boldsymbol{x} \;\middle|\; \frac{P(\boldsymbol{x}|Y=1)}{P(\boldsymbol{x}|Y=0)} \geq \lambda \right\} \\ &= \{\boldsymbol{x} \mid \mathrm{LR}(\boldsymbol{x}) \geq \lambda\}, \end{aligned} \tag{4.1.7}$$

where λ is chosen so that $\int_{\Omega_1} P(\boldsymbol{x}|Y=0)d\boldsymbol{x} = \epsilon_0$. The difference between Bayes and Neyman-Pearson decision rule becomes clear by comparing (4.1.3) and (4.1.7). Both rules are derived by the same likelihood ratio and the generally different threshold term.

4.2 Linear logistic model

In this section, we introduce the linear logistic model. Its optimality is assured in the framework of Bayes risk consistency, which is originally shown in Fisher's linear discriminant analysis.

4.2.1 Bayes risk consistency in normal-normal assumption

As described in the former section, two major decision rules depend on the likelihood ratio. We consider a simple binary classification problem of samples from two normal distributions with different means and the equal variance. Assume that $\boldsymbol{X}|(Y=1) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{X}|(Y=0) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Let $\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a normal density function with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ defined as

$$\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right\}. \tag{4.2.1}$$

It is not strict assumption that we set mean parameter to $\mathbf{0}$ in the normal group density, because there is a proper linear transformation in order to come down to this setting even if $\mathrm{E}[\boldsymbol{X}|Y=0] \neq \mathbf{0}$. Then, the true log likelihood ratio is

$$\begin{aligned} \log \frac{\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\phi(\boldsymbol{x}; \mathbf{0}, \boldsymbol{\Sigma})} &= \log \frac{(2\pi)^{-(p/2)} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right\}}{(2\pi)^{-(p/2)} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \right)} \\ &= \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \end{aligned} \tag{4.2.2}$$

Thus the decision boundary is written as $\Omega_1 = \{\boldsymbol{x} | \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} > c\}$, where c is a proper constant value. This is called Fisher's linear discriminant analysis (LDA) rule. It is well known that the rule has Bayes risk consistency for the setting. We note that the log likelihood ratio is the form of linear predictor: $\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}$, where $\alpha = -\frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ and $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$.

4.2.2 Formulations of linear logistic model

Regardless of whether we can or cannot assume the normal distribution on the two-labeled groups, we assume in the logistic model that the true log likelihood ratio is approximated by the linear predictor as

$$\log\left(\frac{P(Y=1|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}. \tag{4.2.3}$$

The model is equivalent to the generalized linear model in the case that the distribution of class label Y follows Binomial distribution. In fact, we see that the model (4.2.3) is rewritten as

$$\pi = g(F(\boldsymbol{x};\boldsymbol{\theta})), \tag{4.2.4}$$

where $\pi = P(Y=1|\boldsymbol{x})$ and $g(z) = 1/(1+\exp(-z))$ which is called a logistic function. We can regard the class label Y given $\boldsymbol{X}$ as a random sample from Bernoulli distribution with mean parameter π.

4.2.3 Parameter estimation in linear logistic model

If a data set is given as $\{(\boldsymbol{x}_i, y_i); 1 \leq i \leq N\}$, the log-likelihood function of the linear logistic model is written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log \frac{\exp(y_i F(\boldsymbol{x}_i;\boldsymbol{\theta}))}{1+\exp(F(\boldsymbol{x}_i;\boldsymbol{\theta}))} \\ &= \sum_{i=1}^{N} y_i F(\boldsymbol{x}_i;\boldsymbol{\theta}) - \log(1+\exp(F(\boldsymbol{x}_i;\boldsymbol{\theta}))). \end{aligned} \tag{4.2.5}$$

Because the estimator which maximize (4.2.5) cannot be written in analytical formulation, we need numerical optimization to get the maximum likelihood estimator. One of the most widely used optimization strategies is Fisher's scoring method performed by the score function and Fisher information matrix of the log likelihood function. The score function is

written as

$$\begin{aligned}\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{i=1}^{N}(y_i - \hat{\pi}_i)\frac{\partial F(\boldsymbol{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \boldsymbol{W}^\top(\boldsymbol{y} - \hat{\boldsymbol{\pi}}),\end{aligned} \tag{4.2.6}$$

where

$$\boldsymbol{W} = \left(\frac{\partial F(\boldsymbol{x}_1;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \frac{\partial F(\boldsymbol{x}_2;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \cdots, \frac{\partial F(\boldsymbol{x}_N;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top, \tag{4.2.7}$$

$$\boldsymbol{y} = (y_1, y_2, \cdots, y_N)^\top, \quad \hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_N)^\top, \tag{4.2.8}$$

and

$$\hat{\pi}_i = \frac{\exp(F(\boldsymbol{x}_i;\boldsymbol{\theta}))}{1 + \exp(F(\boldsymbol{x}_i;\boldsymbol{\theta}))}. \tag{4.2.9}$$

The Fisher information matrix of the log likelihood is

$$\begin{aligned}\mathrm{E}\left[-\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right] &= \sum_{i=1}^{N}\frac{\partial F(\boldsymbol{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \hat{\pi}_i}{\partial \boldsymbol{\theta}^\top} \\ &= \sum_{i=1}^{N}\frac{\exp(F(\boldsymbol{x}_i))}{(1+\exp(F(\boldsymbol{x}_i)))^2}\frac{\partial F(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}}\frac{\partial F(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}^\top}, \\ &= \sum_{i=1}^{N}\hat{\pi}_i(1-\hat{\pi}_i)\frac{\partial F(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}}\frac{\partial F(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}^\top} \\ &= \boldsymbol{W}^\top \boldsymbol{V}\boldsymbol{W},\end{aligned} \tag{4.2.10}$$

where $\boldsymbol{V} = diag(\hat{\boldsymbol{\pi}})(\boldsymbol{I}_n - diag(\hat{\boldsymbol{\pi}}))$. The maximum likelihood estimator of $\boldsymbol{\theta}$ is calculated by the following update formula from some initial value $\boldsymbol{\theta}^{(0)}$ as

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} + \left\{ \mathrm{E}\left[-\frac{\partial^2 l(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right] \right\}^{-1} \frac{\partial l(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} \\
&= \boldsymbol{\theta}^{(t)} + \left(\boldsymbol{W}^{(t)\top} \boldsymbol{V}^{(t)} \boldsymbol{W}^{(t)} \right)^{-1} \boldsymbol{W}^{(t)\top} (\boldsymbol{y} - \hat{\boldsymbol{\pi}}^{(t)}) \\
&= \left(\boldsymbol{W}^{(t)\top} \boldsymbol{V}^{(t)} \boldsymbol{W}^{(t)} \right)^{-1} \boldsymbol{W}^{(t)\top} \boldsymbol{V}^{(t)} (\boldsymbol{W}^{(t)} \theta^{(t)} + {\boldsymbol{V}^{(t)}}^{-1} (\boldsymbol{y} - \hat{\boldsymbol{\pi}}^{(t)})). \quad (4.2.11)
\end{aligned}$$

In the framework of generalized linear model, $\boldsymbol{Z}^{(t)} = \boldsymbol{W}^{(t)} \boldsymbol{\theta}^{(t)} + {\boldsymbol{V}^{(t)}}^{-1} (\boldsymbol{y} - \boldsymbol{\pi}^{(t)})$ is called the working response, and this algorithm is referred to as the iteratively reweighted least-square method (Nelder and Wedderburn, 1972).

4.3 Quasi-linear logistic model

4.3.1 Bayes risk consistency in normal mixture-normal assumption

As shown in the previous section, the optimality of the linear logistic model is expected when we consider the normal-normal assumption. However, as discussed in Chapter 1, it seems that more natural assumption in the field of biomedical science might be that the covariate vectors of disease samples follow mixture distribution. One of the simplest assumptions is that these are distributed with mixture of normal distributions. In the situation, it holds the following theorem.

Theorem 1. *Assume that* $\boldsymbol{X}|(Y = 1) \sim \sum_{k=1}^{K} p_k \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ *and* $\boldsymbol{X}|(Y = 0) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, *where* K *is a number of mixture components and* p_k *denotes mixing proportion for* $k = 1, 2, \cdots, K$, *satisfying* $\sum_{k=1}^{K} p_k = 1$. *Then, the true log likelihood ratio forms the quasi-linear predictor with tuning parameter* $\tau = 1$.

Proof. Let $\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a normal density function with mean parameter $\boldsymbol{\mu}$ and covariance

matrix $\boldsymbol{\Sigma}$. Then, the true log likelihood ratio is given as

$$\begin{aligned}
\log\frac{\sum_{k=1}^{K} p_k\phi(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma})}{\phi(\boldsymbol{x};\boldsymbol{0},\boldsymbol{\Sigma})} &= \log\frac{\sum_{k=1}^{K} p_k(2\pi)^{-(p/2)}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}}{(2\pi)^{-(p/2)}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)} \\
&= \log\left(\sum_{k=1}^{K} p_k\exp\left(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k-\frac{1}{2}\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\right)\right) \\
&= \log\left(\sum_{k=1}^{K}\exp\left(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k-\frac{1}{2}\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\log p_k\right)\right) \\
&= \log\left(\frac{1}{K}\sum_{k=1}^{K}\exp(\alpha_k+\boldsymbol{\beta}_k^\top\boldsymbol{x})\right) \\
&= F_1(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}), \qquad (4.3.1)
\end{aligned}$$

where $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_K)$ and $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_K)$ with $\alpha_k=-\frac{1}{2}\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\log Kp_k$ and $\boldsymbol{\beta}_k=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$. □

Theorem 1 means that the quasi-linear form is Bayes optimal in the case of the normal mixture-normal assumption. Thus the quasi-linear predictor is expected to be ideal for classifying the two-labeled data with intrinsic heterogeneity as expressed by the normal mixture distribution.

4.3.2 Formulations of quasi-linear logistic model

For a data set $\{(\boldsymbol{x}_i,y_i);1\le i\le N\}$ we define the quasi-linear logistic model as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=F_\tau(\boldsymbol{x}_i;\boldsymbol{\theta}), \qquad (4.3.2)$$

where $\pi_i=P(y_i=1|\boldsymbol{x}_i)$. Here we assume that y_is are independently distributed according to Bernoulli distribution with the mean parameter π_i and the cluster number K of the quasi-linear predictor is some fixed integer. How to determine the number of clusters is described in the next section. Then, the log-likelihood function of the quasi-linear logistic

model is written as

$$l(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n} y_i F_i - \log(1 + \exp(F_i)), \tag{4.3.3}$$

where $F_i = F_\tau(\boldsymbol{x}_i; \boldsymbol{\theta})$. Because the maximum likelihood estimator cannot be written analytically, we performed Fisher's scoring method as the linear logistic model on the parameter estimation. The score function of the log-likelihood of the quasi-linear logistic model is written as

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{\pi}), \tag{4.3.4}$$

where $\boldsymbol{W} = (\partial F_1/\partial \boldsymbol{\theta}, \cdots, \partial F_n/\partial \boldsymbol{\theta})^\top$, $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$ and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)^\top$. The Fisher information matrix is given by

$$\mathrm{E}\left[-\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right] \quad = \quad \boldsymbol{W}^\top \boldsymbol{V} \boldsymbol{W}, \tag{4.3.5}$$

where $\boldsymbol{V} = \mathrm{diag}\{\pi_1(1 - \pi_1), \cdots, \pi_n(1 - \pi_n)\}$. The maximum likelihood estimator of the quasi-linear logistic model is calculated by updating some initial value repeatedly by Fisher's scoring method as

$$\boldsymbol{\theta}^{(t+1)} \quad = \quad \boldsymbol{\theta}^{(t)} + \left(\boldsymbol{W}^{(t)\top} \boldsymbol{V}^{(t)} \boldsymbol{W}^{(t)}\right)^{-1} \boldsymbol{W}^{(t)\top} (\boldsymbol{y} - \boldsymbol{\pi}^{(t)}), \tag{4.3.6}$$

where $\boldsymbol{W}^{(t)} = (\partial F_1/\partial \boldsymbol{\theta}, \cdots, \partial F_n/\partial \boldsymbol{\theta})^\top |_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$, $\boldsymbol{V}^{(t)} = \mathrm{diag}\{\pi_1^{(t)}(1 - \pi_1^{(t)}), \cdots, \pi_n^{(t)}(1 - \pi_n^{(t)})\}$ and $\boldsymbol{\pi}^{(t)} = (\pi_1^{(t)}, \cdots, \pi_n^{(t)})^\top$. This algorithm is the natural extension of the *IRLS* because the equation (4.3.6) is written as $\boldsymbol{\theta}^{(t+1)} = \left(\boldsymbol{W}^{(t)\top} \boldsymbol{V}^{(t)} \boldsymbol{W}^{(t)}\right)^{-1} \boldsymbol{W}^{(t)\top} \boldsymbol{V}^{(t)} \boldsymbol{Z}^{(t)}$. We can combine the L_1 and L_2 regularization method based on the penalized likelihood just as for the penalized linear logistic model (Park and Hastie, 2007; Friedman *and others*, 2010). However, we have no identifiability for the parameters in the model (4.3.3). This fact may cause instability on the parameter estimation. Moreover, we need a simpler model in order to improve the interpretability of the estimated model. The conceptual diagram of such a parsimonious model is described in Figure 4.1. We have two strategies to take it into

consideration as discussed in the next two subsections.

4.3.3 Restricted quasi-linear model

The first strategy is to use the idea of disjoint sets of covariates which were proposed by Omae *and others* (2017). In the strategy, we assume that we know the disjoint decomposition of $\boldsymbol{X}_i$ as $\boldsymbol{X}_{i(1)}, \cdots, \boldsymbol{X}_{i(K)}$ with a fixed cluster size K, and that this is identical among individuals. We denote the size of $\boldsymbol{X}_{i(k)}$ as p_k, where $\sum_{k=1}^{K} p_k = p$. We note that such decomposition is given by beforehand one of clustering methods for $\{\boldsymbol{X}_i; i = 1, \cdots, n\}$ or prior knowledge about the disjoint structure of $\boldsymbol{X}$. The disjoint sets of covariates yield the restricted quasi-linear predictor defined by

$$F_{\tau}^{Res}(\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \cdots, \boldsymbol{X}_{(K)}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{\tau} \log \left(\frac{1}{K} \sum_{k=1}^{K} \exp(\tau \alpha_k + \tau \boldsymbol{\beta}_k^{\top} \boldsymbol{X}_{(k)}) \right). \tag{4.3.7}$$

We can keep the predictor F_{τ}^{Res} to have Bayes risk consistency for the normal mixture-normal assumption if some regularity condition is satisfied as the following theorem.

Theorem 2. *Let* $\boldsymbol{Z}|(Y = 1) \sim \sum_{k=1}^{K} p_k \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ *and* $\boldsymbol{Z}|(Y = 0) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$*, where* K *is a number of mixture components and* p_k *denotes mixing proportion for* $k = 1, 2, \cdots, K$*, satisfying* $\sum_{k=1}^{K} p_k = 1$*. Assume that* $(\boldsymbol{\mu}_1^{\top}, \boldsymbol{\mu}_2^{\top}, \cdots, \boldsymbol{\mu}_K^{\top})$ *are linearly independent. Then, there exists the non-singular matrix* $A \in \mathbb{R}^{p \times p}$ *such that the true log likelihood ratio forms the restricted quasi-linear predictor with tuning parameter* $\tau = 1$ *as* $F_1^{Res}(\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \cdots, \boldsymbol{X}_{(K)}; \boldsymbol{\alpha}, \boldsymbol{\beta})$*, where* $\boldsymbol{X}_{(k)}$ *is the vector of the* k*-th blocked components of* $(A^{\top})^{-1}\boldsymbol{Z}$ *corresponding to any decomposition of* $\boldsymbol{Z}$ *into* $(\boldsymbol{Z}_{(1)}, \cdots, \boldsymbol{Z}_{(K)})$*.*

The proof of the Theorem 2 is given in Appendix A.2. Thus, the quasi-linear predictor is naturally derived when we incorporate the heterogeneity as a normal mixture assumption and modify the optimal form appropriately. We fix $\tau = 1$ below in this section for keep the quasi-linear predictor to be the Bayes risk consistent form.

In this setting, the unknown parameters are fully identifiable because θ_1 is equal to θ_2 if $F_{\tau}^{Res}(\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \cdots, \boldsymbol{X}_{(K)}; \boldsymbol{\theta}_1) = F_{\tau}^{Res}(\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \cdots, \boldsymbol{X}_{(K)}; \boldsymbol{\theta}_2)$ for all $\boldsymbol{X} \in \mathbb{R}^p$. For example, the Figure 4.2 is the diagram used for cluster composition in Omae *and others*

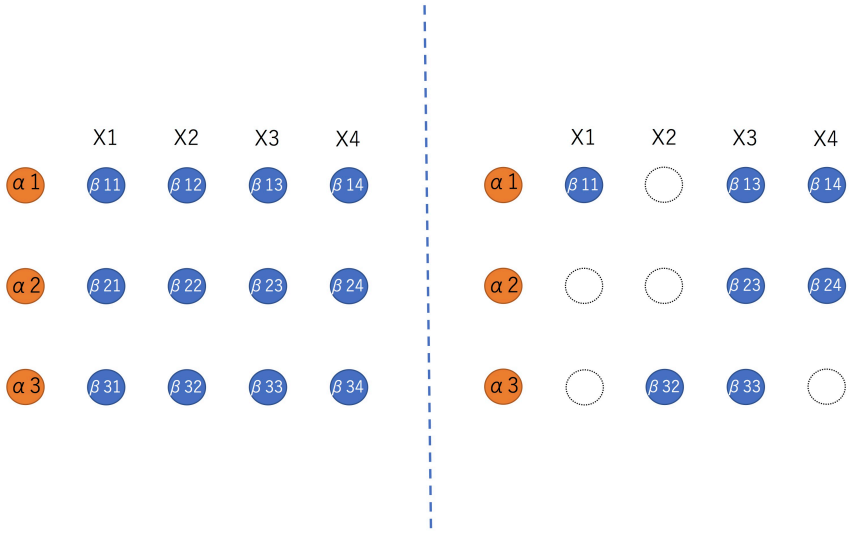

Figure 4.1: The conceptual diagram of the full and parsimonious models in the setting of $p = 4$ and $K = 3$ are drawn. The left figure shows the full model written by the quasi-linear predictor $F_\tau(\boldsymbol{X}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{\tau} \log(\sum_{k=1}^{K} \exp(\alpha_k + \sum_{j=1}^{p} \beta_{kj} X_j))$. The right figure shows the parsimonious model written in the same predictor but has some zero-coefficients.

(2017). This figure would suggest that the appropriate cluster size is 2 or 3.

4.3.4 Cross-penalized quasi-linear model

In the second strategy, we regularize the log-likelihood function by cross-L_1 penalty defined by

$$P^{(c)}(\boldsymbol{\beta}) = \lambda^{(c)} \sum_{\ell \neq m} \sum_{j=1}^{p} |\beta_{\ell j} \beta_{mj}|, \tag{4.3.8}$$

where $\lambda^{(c)}$ is a regularization parameter and β_{kj} is the j-th component of k-th coefficient vector $\boldsymbol{\beta}_k$. When $\lambda^{(c)}$ goes infinity, the estimated parameter $\boldsymbol{\beta}_k$'s would be sparse and $\beta_{kj} = 0$ for any $k \neq \ell$ if $\beta_{\ell j} \neq 0$. Moreover, L_1 and L_2 penalties have an important role for the high dimensional settings as in the generalized linear model. The elastic net (Zou and Hastie, 2005) type of the regularized log-likelihood with cross L_1 penalty is written as

$$l^{pen}(\boldsymbol{\theta}) \quad = \quad l(\boldsymbol{\theta}) - P(\boldsymbol{\beta}) - P^{(c)}(\boldsymbol{\beta}), \tag{4.3.9}$$

where $P(\boldsymbol{\beta}) = \zeta\lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p} |\beta_{kj}| + (1-\zeta)\lambda_2 \sum_{k=1}^{K} \sum_{j=1}^{p} \beta_{kj}^2$. The maximization of equation (4.3.9) is achieved by the full gradient algorithm (Goeman, 2010). The parameter estimation of the quasi-linear logistic model is performed by updating some initial value repeatedly as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \min\{t_{\text{opt}}(\boldsymbol{\theta}^{(t)}), t_{\text{edge}}(\boldsymbol{\theta}^{(t)})\} \boldsymbol{d}(\boldsymbol{\theta}^{(t)}), \tag{4.3.10}$$

where

$$\begin{aligned} \boldsymbol{d}(\boldsymbol{\theta}) &= (d_1(\boldsymbol{\theta}), d_2(\boldsymbol{\theta}), \cdots, d_{K(1+p)}(\boldsymbol{\theta}))^\top \\ &= (d_1^{(0)}(\boldsymbol{\theta}), \cdots, d_K^{(0)}(\boldsymbol{\theta}), d_1^{(1)}(\boldsymbol{\theta}), \cdots, d_p^{(1)}(\boldsymbol{\theta}), \cdots, d_1^{(K)}(\boldsymbol{\theta}), \cdots, d_p^{(K)}(\boldsymbol{\theta}))^\top, \end{aligned}$$

$$t_{\text{edge}}(\boldsymbol{\theta}) = \min_{1+K \leq j \leq K(1+p)} \left(-\frac{\theta_j}{d_j(\boldsymbol{\theta})} : \text{sign}(\theta_j) = -\text{sign}(d_j(\boldsymbol{\theta})) \neq 0 \right)$$

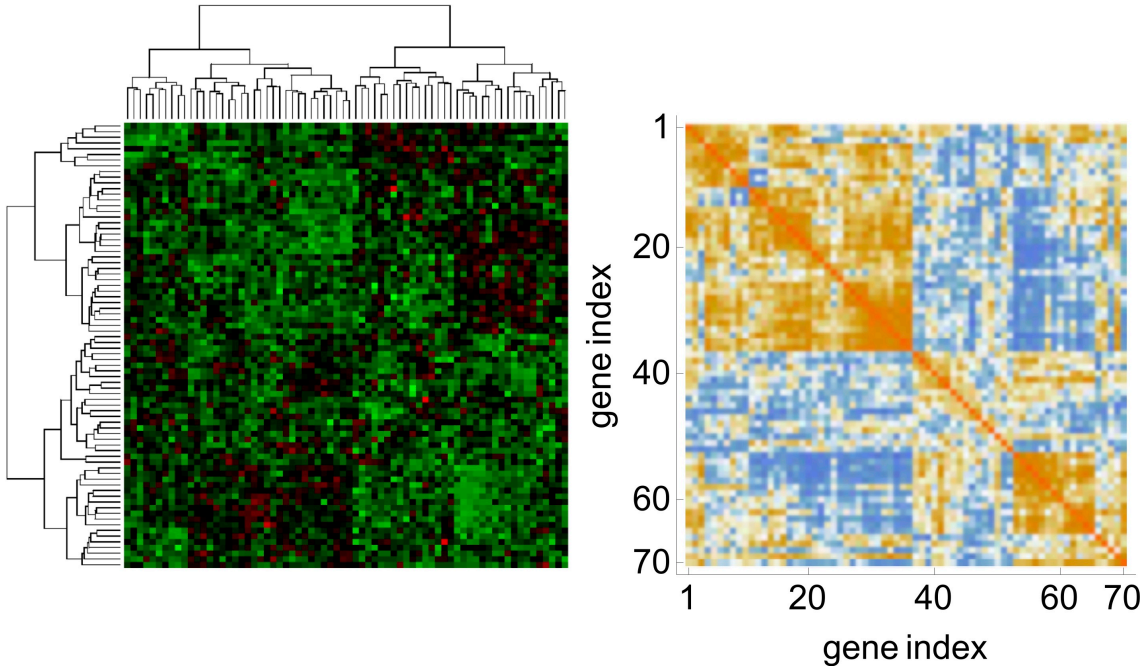

Figure 4.2: The hierarchical clustering and the correlation matrix of 70 genes for the dataset from van't Veer et al. (2002). The left figure shows the clustering result. There are 70 rows representing genes and 78 columns representing samples and the gene expressions ranging from green (negative) to red (positive) are displayed. The right figure shows the corresponding correlation matrix. Elements of the correlation matrix ranging from blue (negative) to yellow (positive) are displayed.

and

$$t_{\mathrm{opt}}(\boldsymbol{\theta}) = \frac{|d(\boldsymbol{\theta})|}{d(\boldsymbol{\theta})^\top \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} d(\boldsymbol{\theta})}.$$

Here $d_j^{(0)}(\boldsymbol{\theta}) = l_j^{(0)}(\boldsymbol{\theta})$ for $j = 1, \cdots, K$ and

$$d_j^{(k)}(\boldsymbol{\theta}) = \begin{cases} h_j^{(k)}(\boldsymbol{\theta}) - \lambda_j^{(k)} \mathrm{sign}(\beta_{kj}) & \text{if } \beta_{kj} \neq 0 \\ h_j^{(k)}(\boldsymbol{\theta}) - \lambda_j^{(k)} \mathrm{sign}(l_j^{(k)}(\boldsymbol{\theta})) & \text{if } \beta_{kj} = 0 \text{ and } |h_j^{(k)}(\boldsymbol{\theta})| > \lambda_j^{(k)} \\ 0 & \text{otherwise} \end{cases} \tag{4.3.11}$$

for $j = 1, \cdots, p$, where $h_j^{(k)}(\boldsymbol{\theta}) = l_j^{(k)}(\boldsymbol{\theta}) - 2(1-\zeta)\lambda_2 \beta_{kj}$, $\mathrm{sign}(z)$ is the sign function, $l_j^{(0)}$ is the j-th component and $l_j^{(k)}$ is the $(K + (k-1)p + j)$-th component of the score function and $\lambda_j^{(k)} = \left(\zeta\lambda_1 + \lambda^{(c)} \sum_{\ell \neq k} |\beta_{\ell j}|\right)$,.

4.4 Simulation studies of quasi-linear logistic model

In this section, we show the efficiency of the restricted quasi-linear logistic models (QL) compared with the linear logistic model (LL) by the simulation studies. All simulations in the section were preformed in Omae *and others* (2017).

We conducted a simulation of five scenarios. In each scenario, we consider two populations of equal sample size. The class labels $Y = 0$ and $Y = 1$ mean the normal population and the disease population, respectively. In subsection 4.4.1, we showed the consistency of the parameters estimation of the restricted quasi-linear logistic model in the low dimensional setting. Here, we checked it in a simple setting that has a Bayes optimal predictor of the quasi-linear form. In this example, the sample size was set to 400 or 1600. Each situation was tried 1000 times by random samples. In subsection 4.4.2, the performance of the restricted quasi-linear logistic model was compared with the linear logistic model by four settings focusing on the test AUC. Covariates of the sample from the normal population and disease population follow normal and normal mixture distribution, respectively. In these examples, we simulated 1,000 random data sets of 400 samples and 200 samples for the training data set and the test data set, respectively. For these settings, we use the L_1

and L_2 penalization method in order to avoid overfitting and hard computation if necessary. The tuning parameters were determined by a cross-validation method. Below, we define $\boldsymbol{r}_p = (r, r, \cdots, r) \in \mathbb{R}^p$ for scalar r for simple notations.

4.4.1 Checking consistency

In this scenario, we assumed normal distribution for the normal group and normal mixture distribution for the disease group as

$$\boldsymbol{X}|(Y=0) \sim \mathrm{N}(\boldsymbol{0}_2^\top, \boldsymbol{I}_2), \quad \boldsymbol{X}|(Y=1) \sim \sum_{g=1}^{2} \tau_g \mathrm{N}(\boldsymbol{\mu}_{1g}, \boldsymbol{I}_2), \quad \sum_{g=1}^{2} \tau_g = 1. \tag{4.4.1}$$

We let $\boldsymbol{\mu}_{11} = (-1, 0)^\top$ and $\boldsymbol{\mu}_{12} = (0, 1.5)^\top$. In this setting, the Bayes optimal form is $\log(\exp(\alpha_1 + \beta_1 X_1) + \exp(\alpha_2 + \beta_2 X_2))$ from (4.3.1). Figure 4.3 shows box plots of estimated parameters in the 1000 trials. The optimal parameter derived from the true likelihood is given as $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (-1.19, -1.82, -1.00, 1.50)$. The mean values of estimated parameters from 1000 trials are $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (-1.28, -1.99, -1.07, 1.61)$ for the 400 samples datasets and $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (-1.21, -1.85, -1.01, 1.53)$ for the 1600 samples datasets. Thus we observed that the parameter estimation was more precise and nearly equal to the true values when the sample size was large.

4.4.2 Evaluation of test AUC

- (a): normal - normal

  In this example we assumed normal distributions for both groups as

  $$\boldsymbol{X}|(Y=y) \sim \mathrm{N}(\boldsymbol{\mu}_y, \boldsymbol{I}_p) \quad (y = 0, 1). \tag{4.4.2}$$

  We had three settings of different parameters: (a)-(1) $p = 2, \boldsymbol{\mu}_0 = \boldsymbol{0}_2^\top, \boldsymbol{\mu}_1 = \boldsymbol{1}_2^\top$, (a)-(2) $p = 100, \boldsymbol{\mu}_0 = \boldsymbol{0}_{100}^\top, \boldsymbol{\mu}_1 = \boldsymbol{0}.1_{100}^\top$, (a)-(3) $p = 100, \boldsymbol{\mu}_0 = \boldsymbol{0}_{100}^\top, \boldsymbol{\mu}_1 = \boldsymbol{0}.5_{100}^\top$. When we used the restricted quasi-linear logistic model, we assumed to misspecify there were heterogeneous structure, as $K = 2$ and $p_1 = p_2 = 1$ for (a)-(1) or $p_1 = p_2 = 50$ for

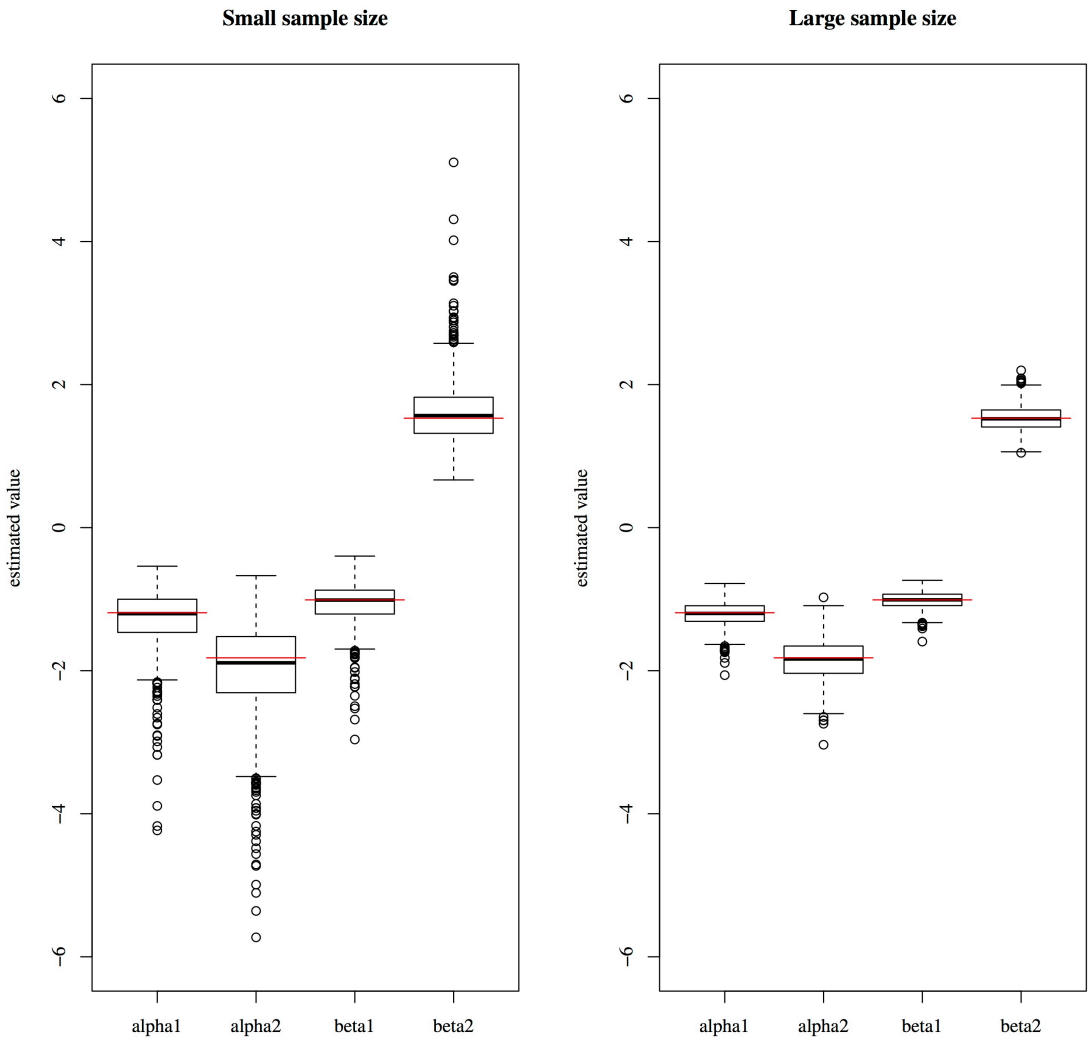

Figure 4.3: Box plots of the estimated parameters in the simulation of checking consistency. The left and right figure show results from 400 samples and 1600 samples, respectively. The red lines mean the optimal parameters derived from the true log-likelihood.

(a)-(2) and (a)-(3).

- (b): normal - normal mixture

  In this example, we assumed normal distributions for the normal group and normal mixture distributions for the disease group as

$$\boldsymbol{X}|(Y=0) \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{I}_p), \quad \boldsymbol{X}|(Y=1) \sim \sum_{g=1}^{G} \tau_g \mathrm{N}(\boldsymbol{\mu}_{1g}, \boldsymbol{I}_p), \quad \sum_{g=1}^{G} \tau_g = 1. \tag{4.4.3}$$

  We had four settings. In (b)-(1) and (b)-(2), we let $G = 2$, $p = 100$, $\tau_1 = \tau_2 = 0.5$, $\boldsymbol{\mu}_0 = \mathbf{0}_{100}^\top$. In (b)-(3) and (b)-(4), we let $G = 3$, $p = 100$, $\tau_1 = \tau_2 = \tau_3 = 1/3$, $\boldsymbol{\mu}_0 = \mathbf{0}_{100}^\top$. The mean parameter of case group was set as (b)-(1) $\boldsymbol{\mu}_{11} = (-1, \mathbf{0}_{99})^\top$, $\boldsymbol{\mu}_{12} = (\mathbf{0}_{50}, 1.5, \mathbf{0}_{49})^\top$, (b)-(2) $\boldsymbol{\mu}_{11} = (-1_{10}, \mathbf{0}_{90})^\top$, $\boldsymbol{\mu}_{12} = (\mathbf{0}_{50}, \mathbf{1}.5_{10}, \mathbf{0}_{40})^\top$, (b)-(3) $\boldsymbol{\mu}_{11} = (-1.5, \mathbf{0}_{99})^\top$, $\boldsymbol{\mu}_{12} = (\mathbf{0}_{34}, 1.5, \mathbf{0}_{65})^\top$, $\boldsymbol{\mu}_{13} = (1, \mathbf{0}_{99})^\top$, (b)-(4) $\boldsymbol{\mu}_{11} = (-\mathbf{1}.5_3, \mathbf{0}_{97})^\top$, $\boldsymbol{\mu}_{12} = (\mathbf{0}_{34}, \mathbf{1}.5_3, \mathbf{0}_{63})^\top$, $\boldsymbol{\mu}_{13} = (\mathbf{0}_{67}, \mathbf{1}_3, \mathbf{0}_{30})^\top$. For the restricted quasi-linear logistic model we assumed to specify correctly there were G heterogeneous structure as $K = G$ and $p_1 = p_2 = 50$ or $p_1 = 34, p_2 = p_3 = 33$.

- (c): normal mixture-normal mixture

  In this example, we assumed normal mixture distributions for both groups.

$$\boldsymbol{X}|(Y=y) \sim \sum_{g=1}^{G} \tau_{yg} \mathrm{N}(\boldsymbol{\mu}_{yg}, \boldsymbol{I}_p), \quad \sum_{g=1}^{G} \tau_{yg} = 1 \ \ (y = 0, 1). \tag{4.4.4}$$

  We used the following settings: $G = 2$, $p = 100$, $\tau_{yg} = 0.5$ ($y = 0, 1$, $g = 1, 2$), $\boldsymbol{\mu}_{01} = \mathbf{0}_{100}^\top$, $\boldsymbol{\mu}_{02} = (\mathbf{0}_{50}, \mathbf{0}.3_{10}, \mathbf{0}_{40})^\top$, $\boldsymbol{\mu}_{11} = (\mathbf{0}.5_{50}, \mathbf{0}_{50})^\top$, $\boldsymbol{\mu}_{12} = (\mathbf{0}_{50}, \mathbf{0}.8_{50})$. For the restricted quasi-linear logistic model, we assumed to specify there are two heterogeneous structure as $K = 2$ and $p_1 = p_2 = 50$.

- (d): normal- normal mixture (Correlated)

  In this example, we assumed normal distributions for the normal group and normal mixture distributions for the disease group.

$$\boldsymbol{X}|(Y=0) \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \quad \boldsymbol{X}|(Y=1) \sim \sum_{g=1}^{G} \tau_g \mathrm{N}(\boldsymbol{\mu}_{1g}, \boldsymbol{\Sigma}), \quad \sum_{g=1}^{G} \tau_g = 1. \tag{4.4.5}$$

The variance assumption was based on a real dataset as shown in Figure 4.2. We used the following settings: (1) $G = 2$, $p = 70$, $\tau_1 = \tau_2 = 0.5$, $\mu_0 = \mathbf{0}_{70}^{\top}$, $\boldsymbol{\mu}_{11} = (-\mathbf{0}.5_5, \mathbf{0}_{65})^{\top}$, $\boldsymbol{\mu}_{12} = (\mathbf{0}_{35}, \mathbf{1}_5, \mathbf{0}_{30})^{\top}$, $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 \\ \boldsymbol{\Sigma}_2^{\top} & \boldsymbol{\Sigma}_1 \end{pmatrix}$, where $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{I}_{35} + 0.3\boldsymbol{J}_{35}, \boldsymbol{\Sigma}_2 = -0.15\boldsymbol{J}_{35}$, where $\boldsymbol{J}_m$ is a matrix of size m of which all components are 1. For the restricted quasi-linear logistic model we assumed to specify there are two heterogeneous structure as $K = 2$ and $p_1 = p_2 = 50$.

Table 4.1 - (a) summarizes the AUC values of the test datasets for the (a) settings. We note that the linear logistic model is optimal form in the sense of likelihood ratio under this assumption. However, the restricted quasi-linear logistic model is not less than the simple linear model regardless of misspecifying structure. This is because the quasi-linear predictor locally includes the linear boundary, and almost of all data points are fitted on it. As a result, the predictions based on the quasi-linear predictor were not so mismatched.

Table 4.1 - (b) summarizes the AUC values of the test datasets for the (b) settings. We note that the restricted quasi-linear logistic model is Bayes optimal form under this assumption. Unlike the situation in 4.1, the quasi-linear predictor succeeded in making a difference in performance relative to the ordinary linear predictor. As the numbers of effective explanatory valuables increased, the difference in predictive performance between the quasi-linear and linear predictors also grew. In these setting, the L_1 shrinkage method performed well, because the number of effective explanatory variables was small compared to the number of noisy variables.

Table 4.1 - (c) summarizes the AUC values of test datasets for the (c) setting. When we assumed the normal mixture assumption for both groups, the optimal form of the predictor is no longer simple, and differs both from the linear and the quasi-linear forms. However, the quasi-linear predictor also worked well in this setting. This result indicates that the quasi-

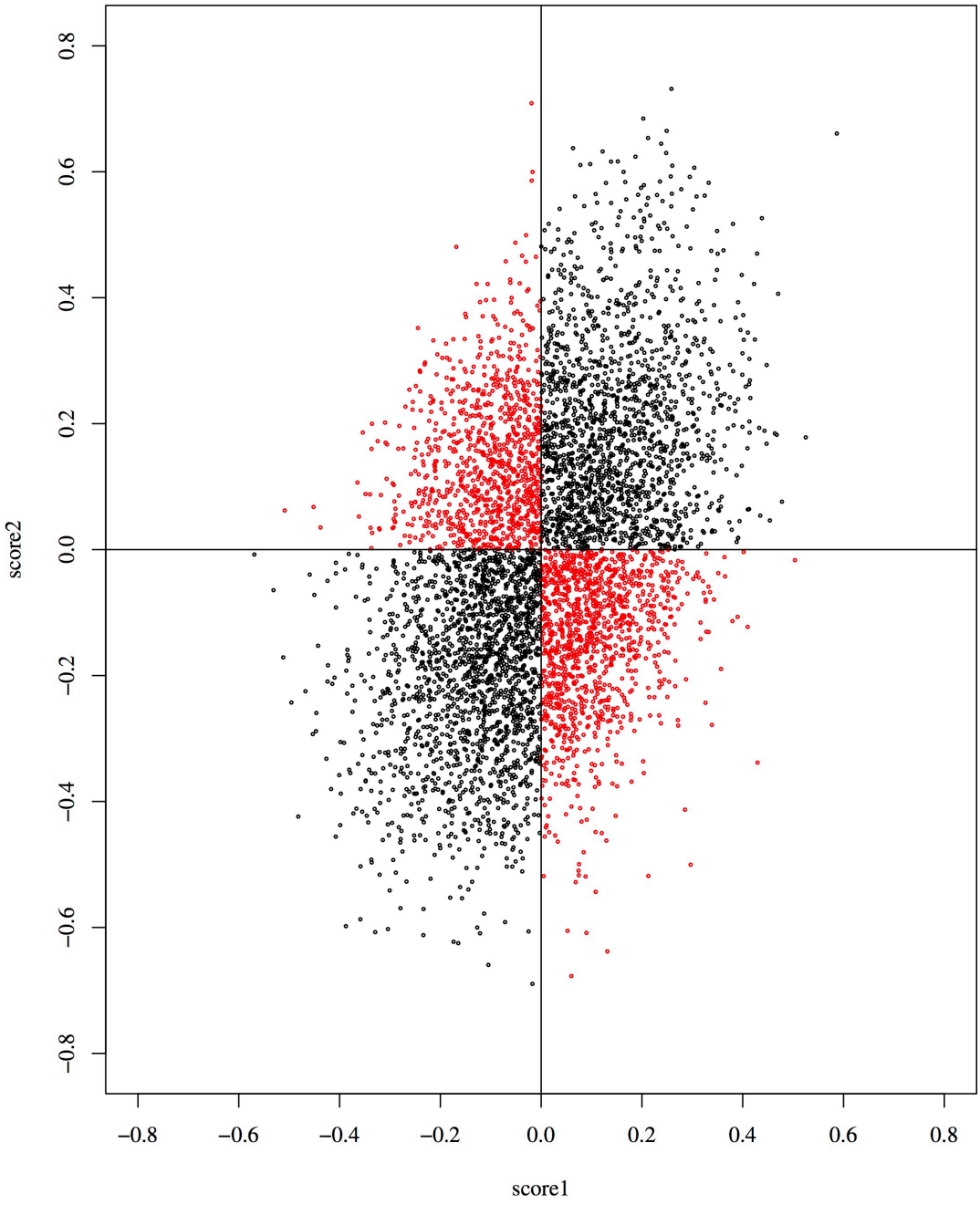

Figure 4.4: t-statistic values for two datasets from van't Veer et al. (2002). The red points show the genes with sign mismatched t-values for these data.

linear predictor should have good predictive performance relative to the linear predictor in complex heterogeneous setting which seems to be often in real datasets.

Table 4.1 - (d) summarizes the AUC values of test datasets for the (d) setting. The quasi-linear predictor also worked well in this setting.

Table 4.1: Estimated AUC (standard deviation) of 1000 repetitions. **LR** denotes the discrimination function derived from the true log-likelihood ratio for each scenario.

			LL		**QL**		**LR**
			ridge	lasso	ridge	lasso	no penalty
(*a*)	homo-homo	(1)	0.841 (0.027)	0.840 (0.027)	0.818 (0.029)	0.818 (0.029)	0.842 (0.027)
		(2)	0.690 (0.039)	0.665 (0.040)	0.679 (0.041)	0.685 (0.029)	0.760 (0.033)
		(3)	0.999 (0.001)	0.997 (0.002)	0.999 (0.001)	0.999 (0.001)	0.999 (0.001)
(*b*)	homo-hetero	(1)	0.641 (0.040)	0.675 (0.038)	0.659 (0.039)	0.725 (0.036)	0.754 (0.034)
		(2)	0.953 (0.014)	0.960 (0.013)	0.985 (0.007)	0.963 (0.016)	0.986 (0.006)
		(3)	0.616 (0.040)	0.642 (0.040)	0.634 (0.040)	0.668 (0.040)	0.740 (0.035)
		(4)	0.757 (0.033)	0.796 (0.032)	0.817 (0.029)	0.827 (0.029)	0.890 (0.022)
(*c*)	hetero-hetero	(1)	0.713 (0.039)	0.697 (0.047)	0.766 (0.035)	0.752 (0.039)	0.824 (0.029)
(*d*)	correlated	(1)	0.762 (0.034)	0.741 (0.037)	0.781 (0.033)	0.736 (0.055)	0.841 (0.024)

Table 4.2: Estimated AUC (95% confidence interval) by elastic net shrinkage; training dataset from van't Veer *and others* (2002), test dataset from Buyse *and others* (2006). A parameter ϵ denotes the proportion of ridge regularization to lasso regularization.

	LL	**QL**$(K = 2)$
$\epsilon = 0.25$	0.732 (0.665, 0.796)	0.755 (0.691, 0.814)
$\epsilon = 0.50$	0.723 (0.655, 0.788)	0.754 (0.691, 0.813)
$\epsilon = 0.75$	0.707 (0.636, 0.776)	0.748 (0.684, 0.807)

4.5 Applications of quasi-linear logistic model

In this section, we shows the Application studies from Omae *and others* (2017). In the paper, we compared the predictive performance between the restricted quasi-linear logistic model and the other common methods for binary classification using real dataset. We applied our method for two datasets, namely breast cancer and prostate cancer data. For both types of datasets, two independent datasets were used as training and testing to evaluate the predictive ability by test AUC. First, we compared the test AUC among decision tree (DT), random forest (RF), support vector machine (SVM), naive Bayes (NB), group lasso (GL), neural network (NN), L_1 or L_2 penalized linear logistic (LL1, LL2) and L_1 or L_2 penalized restricted quasi-linear logistic (QL1, QL2) model. Performance was evaluated by the test AUC and the 95% CIs of the test AUC based on 2000 bootstrapping sampling, as described in Yan *and others* (2015). The tuning parameters were determined by a grid search and resampling method as needed. Second, the stability for marker selection was compared among LL1, QL1 and GL. We used a similarity index proposed by Kalousis *and others* (2005) defined by $S(A, B) = |A \cap B|/|A \cup B|$, where A and B are subsets of marker index set, and $|A|$ is a cardinality of the set A. S takes a value between 0 and 1 whose high value means high stability. We evaluated the stability measure by $\frac{2}{R(R-1)} \sum_{i=1}^{R-1} \sum_{j=i+1}^{R} S(M_i, M_j)$, where $M_1, \cdots, M_R$ are sets of the selected marker for R bootstrap sample sets from the training data set. R was set to 100 below.

Breast Cancer data

We used the dataset from van't Veer *and others* (2002) as the training dataset and the dataset from Buyse *and others* (2006) as the test dataset. We focused on the 70 genes detected by van't Veer *and others* (2002) for relevant genes to the prognosis of breast cancer patients. These datasets include 78 patients and 307 patients. In the study of van't Veer *and others* (2002), the linear predictor was evaluated to classify metastatic events. Because the estimated parameters of the linear model are related to the t-statistic values, we checked the t-statistics directly for the purpose of visualization. If the data have heterogeneous structure, it can be clarified by observing the difference between two divided, independent

datasets. We therefore divided the data into two independent datasets, data1 and data2. Figure 4.4 shows the correspondence of the t-statistics for both datasets. Some genes had no consistency in the signs of their t-statistics values. It indicates that gene expressions of some patients from the metastatic group had higher expression, whereas gene expressions of the other patients had lower expression, relative to the non-metastatic group. This phenomenon may be caused by heterogeneous factors (Omae *and others*, 2016, 2017). In fact, due to the existence of multiple subtypes of breast cancer, the disease is known to exhibit heterogeneity (Sørlie *and others*, 2001). We therefore used the dataset to compare the performance between the restricted quasi-linear logistic model and the other methods.

Figure 4.5 displays the estimated AUC for the test dataset. The restricted quasi-linear logistic model with L_1 (QL1) and L_2 (QL2) penalty, performed better than the linear logistic model with L_1 (LL1) and L_2 (LL2) penalty and any other non-linear methods. The highest test AUC was obtained when we used QL1 based on two clusters. The test AUC of the restricted quasi-linear logistic model did not change for different cluster sizes ($K = 2$ and $K = 3$). The numbers of selected markers in LL1, QL1 ($K = 2$), QL1 ($K = 3$) and GL were 14, 14, 24 and 70, respectively. Similarly, the stability measures were 0.323, 0.320, 0.399 and 0.960, respectively. The stability did not differ between LL1 and QL1 greatly. We note that GL almost did not shrink any coefficients to zero in this setting.

When we use the linear logistic model, the absolute value of the coefficients of each marker reflects the order of importance of all markers for prediction. Therefore, the linear logistic model is understandable in the sense that we can recognize strong markers. This is no longer a consideration when we use a generalized non-linear predictor. However, the restricted quasi-linear logistic model enables us to compare coefficients within the same cluster. An example is shown in Figure 4.6, which displays the ranking of the absolute values of the estimated coefficients by the ridge regularization method based on the existence of two clusters. The gene labels are arranged in order of the rankings. We observed that the restricted quasi-linear logistic model and the linear logistic model gave quite different rankings. This result shows that the quasi-linear predictor would produce different interpretations for the relationship between the markers.

Figure 4.7 shows learning and fitting of the quasi-linear predictor using the L_1 penalty regularization method. The estimated score distributions in the training and test datasets were quite well-matched. Figure 4.7 shows that the restricted quasi-linear logistic model of two clusters with L_1 regularization will work well if we give a cut-off value for binary decisions. For example, the test error rates of the restricted quasi-linear and linear logistic models were 37.8% and 45.0%, respectively, when we used the Youden-index (Youden, 1950).

Although the quasi-linear predictor is approximately equivalent to the maximum function, the two are numerically different. In fact, the test AUC of the restricted quasi-linear score with the L_1 penalty regularization method when we assumed two clusters was 0.752, and the corresponding predictor with maximum score F_∞ is 0.745, so that the smooth non-linearity of the quasi-linear form produced good predictive performance.

The elastic net shrinkage method (Zou and Hastie, 2005), which combines the lasso and ridge shrinkage methods, is among the most frequently used. When we combined the restricted quasi-linear model and the elastic net regularization, the number of the tuning parameters was inflated. Although we used the elastic net experimentally for the application for some selected parameters, the predictive performance was not significantly different from the performance obtained with either the L_1 or L_2 penalty. Detailed results are summarized in Table 4.2. Moreover, to check the utility of the unsupervised clustering, we randomly divided the 70 genes into two subsets of 36 and 34 genes, and applied QL2 for the test dataset (2000 times). Figure 4.8 shows that clustered subsets (red line) performs better than randomly divided subsets. Thus, unsupervised clustering naturally benefits supervised learning via the quasi-linear form.

Prostate Cancer data

The data set was taken from Setlur *and others* (2008) which contains expression data for 6144 genes obtained from 455 prostate cancer tumors. The tumors were from 103 subjects determined to be fusion status-positive and 352 subjects determined to be fusion status-negative. We randomly divided the whole dataset into two independent datasets with the same number of tumor samples (training and test data) while maintaining the ratio of

positive to negative statuses. First, we selected 100 relevant genes, which had top 100 absolute value of t-statistic between the two statuses using only the training dataset. Such marker preselection has been performed in many studies (Dettling and Bühlman, 2003). For QL, grouping of 100 genes was based on the Ward's clustering method only by training dataset. We had two options for dividing all the genes into clusters. In the first option, the 100 genes were divided into two clusters, one with 81 and the other with 19 genes. In the second option, the 100 genes were divided into three clusters of 25, 56, and 19 genes. For GL, We used two clusters option. We then compared the test AUC among all comparative methods. Figure 4.9 displays the estimated AUC for the test dataset. As well as the application for breast cancer data, QL1 and QL2 performed better than any other comparative methods. The numbers of selected markers in LL1, QL1 ($K = 2$), QL1 ($K = 3$) and GL were 31, 38, 67 and 100, respectively. Similarly, the stability measures were 0.361, 0.993, 0.982 and 1.00, respectively. The stability of QL1 was higher than LL1. We note that GL almost did not shrink any coefficients to zero as application for breast cancer data set.

4.6 Discussion on quasi-linear logistic model

We focused on the optimal prediction function formulated by the log likelihood ratio based on the Bayes risk consistency. At first, it was confirmed that the optimality of linear predictor are assured when we assume that the covariates of disease and normal samples follow normal distribution with equal variance. Then we showed that the quasi-linear predictor is the Bayes optimal predictor when we assume that the covariates of disease and normal samples follow normal mixture and normal distribution with equal variance respectively, assuming that the simplest assumption of disease heterogeneity would be denoted by such normal mixture formulation. In this chapter, we thus focused on heterogeneous structure and investigated how to reflect such heterogeneity in the prediction function. For this purpose, the quasi-linear predictor was used for extension of the linear logistic model. The quasi-linear predictor was derived as the generalized mean called the Kolmogorov-Nagumo average. The quasi-linear form is also called a soft maximum function or the log-sum-exp function (Boyd and Vandenberghe, 2004). In the context of machine learning theory, the soft maximum

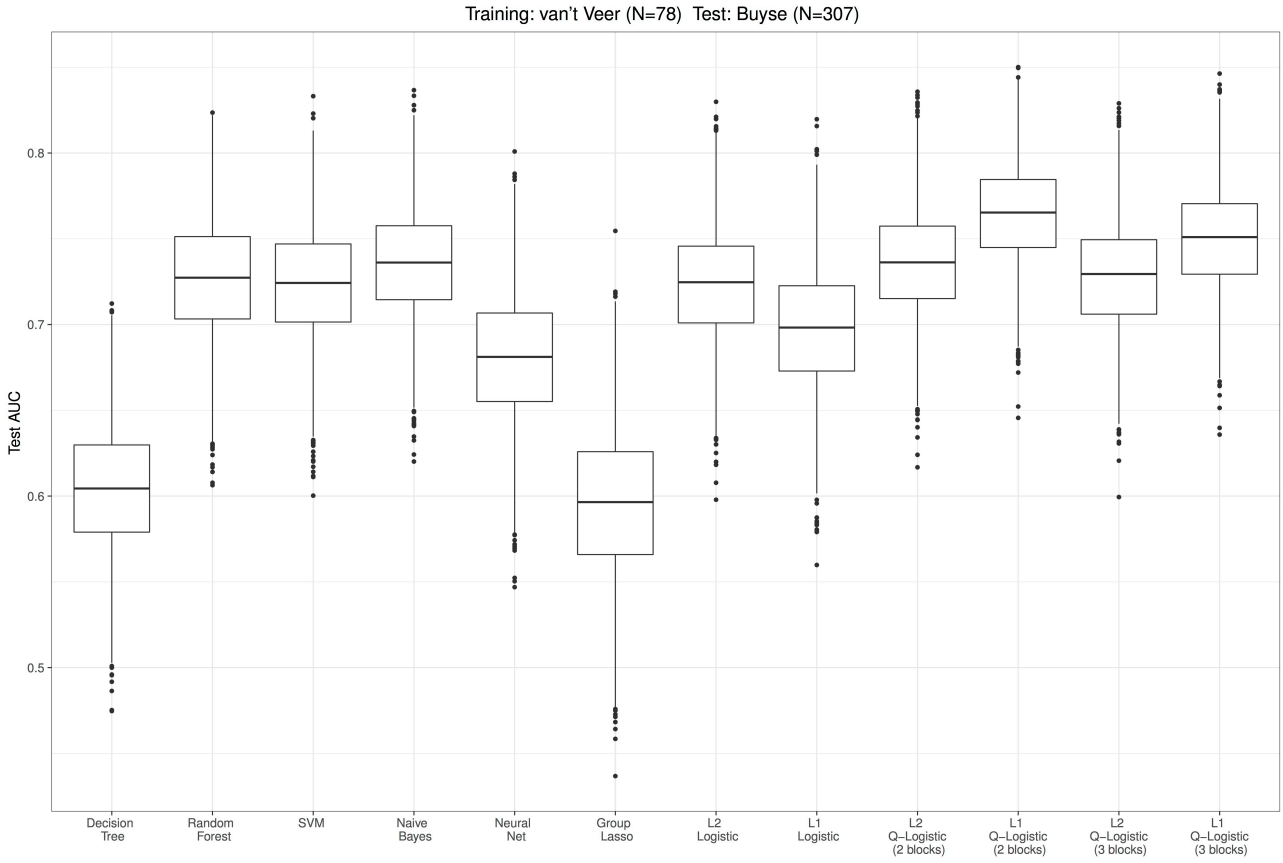

Figure 4.5: Box plots of the test AUC for all comparative methods in breast cancer data. The vertical axis corresponds to the test AUC values. These values are estimated by bootstrap sampling from the test dataset.

function intends to the differentiability and approximation of the maximum function. In the context of computer science, the log-sum-exp function is used to avoid computational problems such as overflow. The non-linearity of the quasi-linear predictor is explained by the soft maximum function. The maximum score of all separated cluster achieves cluster selection. We developed two strategy to avoid the loss of parameter identifiability and difficulty in parameter estimation. The first strategy is to use the restricted quasi-linear model which is defined by the quasi-linear predictor with disjoint clusters of markers. The restricted formulation gave understandability of the parameters. Moreover, this formula does not need any prior information or assumption to separate features to some clusters because it can be easily obtained from the results of unsupervised learning. The simulation and application studies indicated that the restricted quasi-linear logistic model has good performance in the assumption of the heterogeneity, while the performance is comparable with the linear logistic model also in the classical settings whose optimal predictor is the linear form. The second strategy is to add the cross-penalty to the log-likelihood function of the quasi-linear model. Such penalization method is discussed in more detail and applied for the quasi-linear relative risk model in the next section.

The quasi-linear predictor is based on the idea of combining predictors, which is related with several ideas in the literature. For example, a mixture of expert model discusses the idea of decomposing input space Jacobs *and others* (1991), where the model divides the problem space probabilistically and the predictors learned in all sub-spaces are combined. The restricted quasi-linear predictor utilizes the information given by the clustering method to reflect heterogeneity of markers and combines the linear predictors of all clusters, thus it relies on the disjoint decomposition of the markers. The method of combining linear predictors was also discussed in Thompson and Baker (1981), known as composite links, which assumes that the score is formed by a weighted sum of block-wise markers. Unlike the generalized linear model, the composite link model does not restrict single link function to use. In the special case, the composite link logistic model corresponds to the restricted quasi-linear logistic model. However, these ideas differ in that the composite link considers the sum of the linked linear predictors whereas the restricted quasi-linear predictor considers a link

of the summarization of linear predictors in all clusters. The key in our proposal is to model the heterogeneity utilizing the information from the clustering method in the connection of the supervised learning with the unsupervised learning without any assumptions via the change of the predictor form from the simple average to the Kolmogorov-Nagumo average, which mean the linear and quasi-linear, respectively.

For future work, we would extend some fixed settings in the thesis; the choice of the clustering method in the restricted quasi-linear model, the size of the markers, the set of the tuning parameters, the type of the outcomes and the format of targeted data. Because the restricted quasi-linear predictor can be defined by any decomposition ideas, the performance should be evaluated by any other clustering methods than the Ward's method, such as the k-means (McQueen, 1967). Moreover, we need to investigate the size of markers and the number of candidate sets of tuning parameters adding to the parameter τ, to get more flexible form of the quasi-linear function. The quasi-linear predictor would be also applicable in a case of the continuous outcomes and in a regression model although we focused on the binary outcomes and logistic model in the thesis. Regarding this point, we discuss the survival time analysis with the quasi-linear predictor is discussed in the next section. The performance of the quasi-linear predictor would be exhibited in the mixed-up large dataset, which would play an important role in near future biomedical studies, because such data must be heterogeneous. Furthermore, our method is not limited to biomedical data, and could be also beneficial for analysis of any data which has heterogeneous structure.

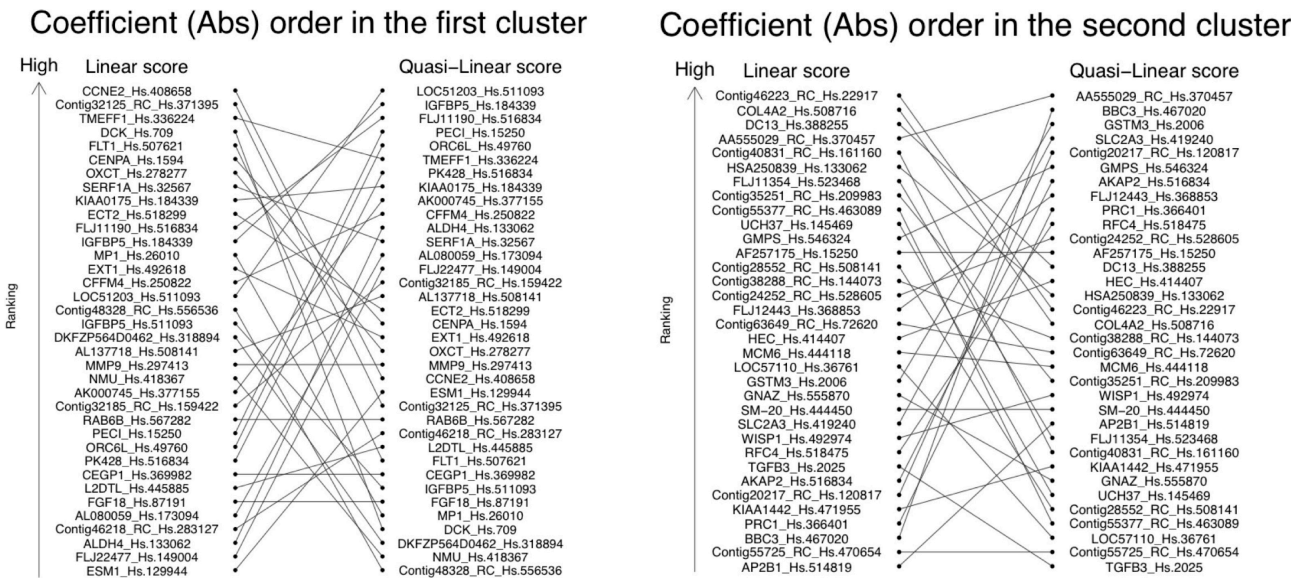

Figure 4.6: Ranking of the absolute values of the coefficients within the cluster with ridge regularization.

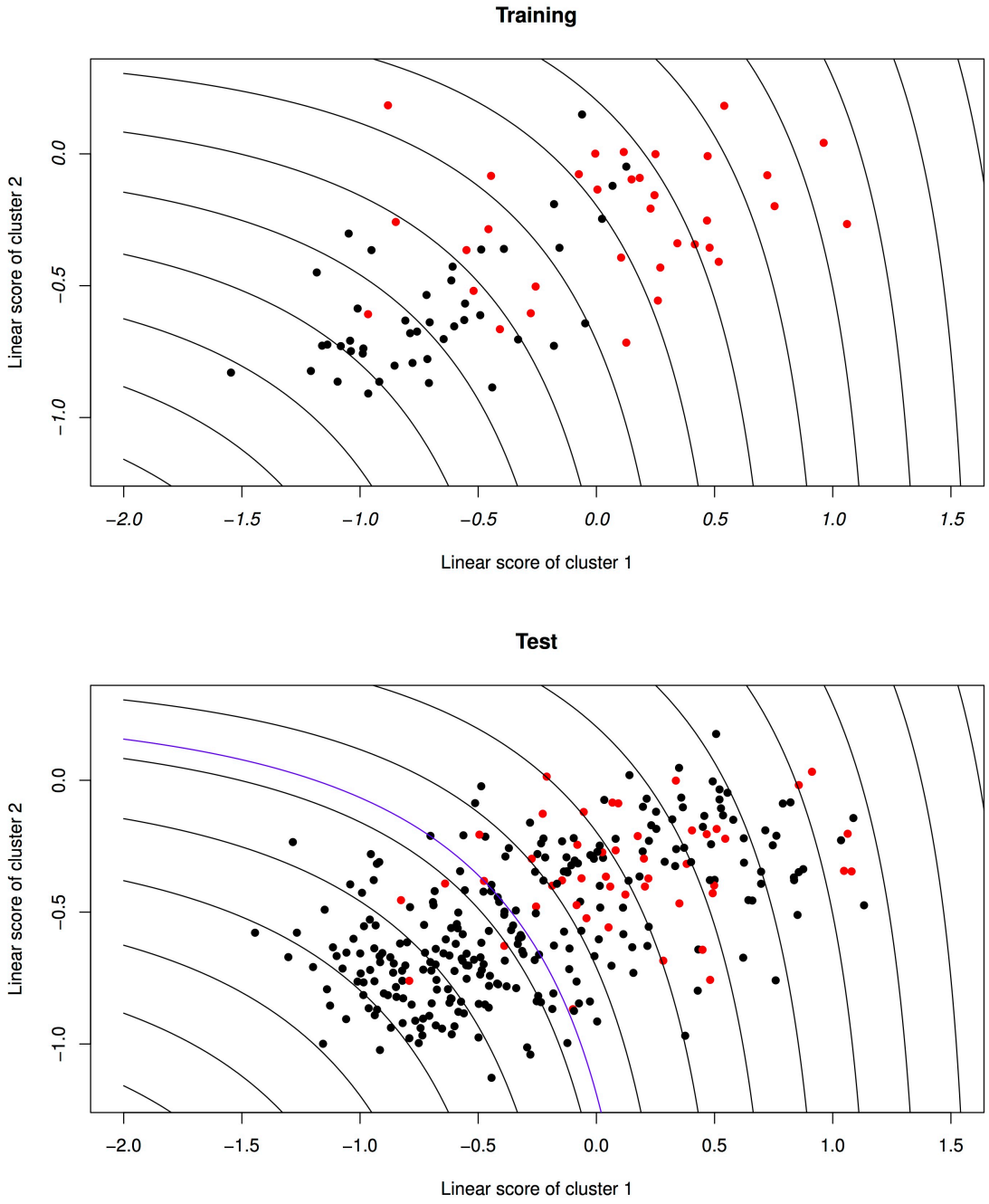

Figure 4.7: Learning and fitted plot for the training and test dataset when using the quasi-linear predictor of two clusters with L_1 penalty regularization. The horizontal and vertical axes are the linear scores of the first and second clusters. Red points indicate the metastatic group and black points indicate the control group. Curve lines are contours of the quasi-linear score and blue line shows cut-off value based on Youden-index.

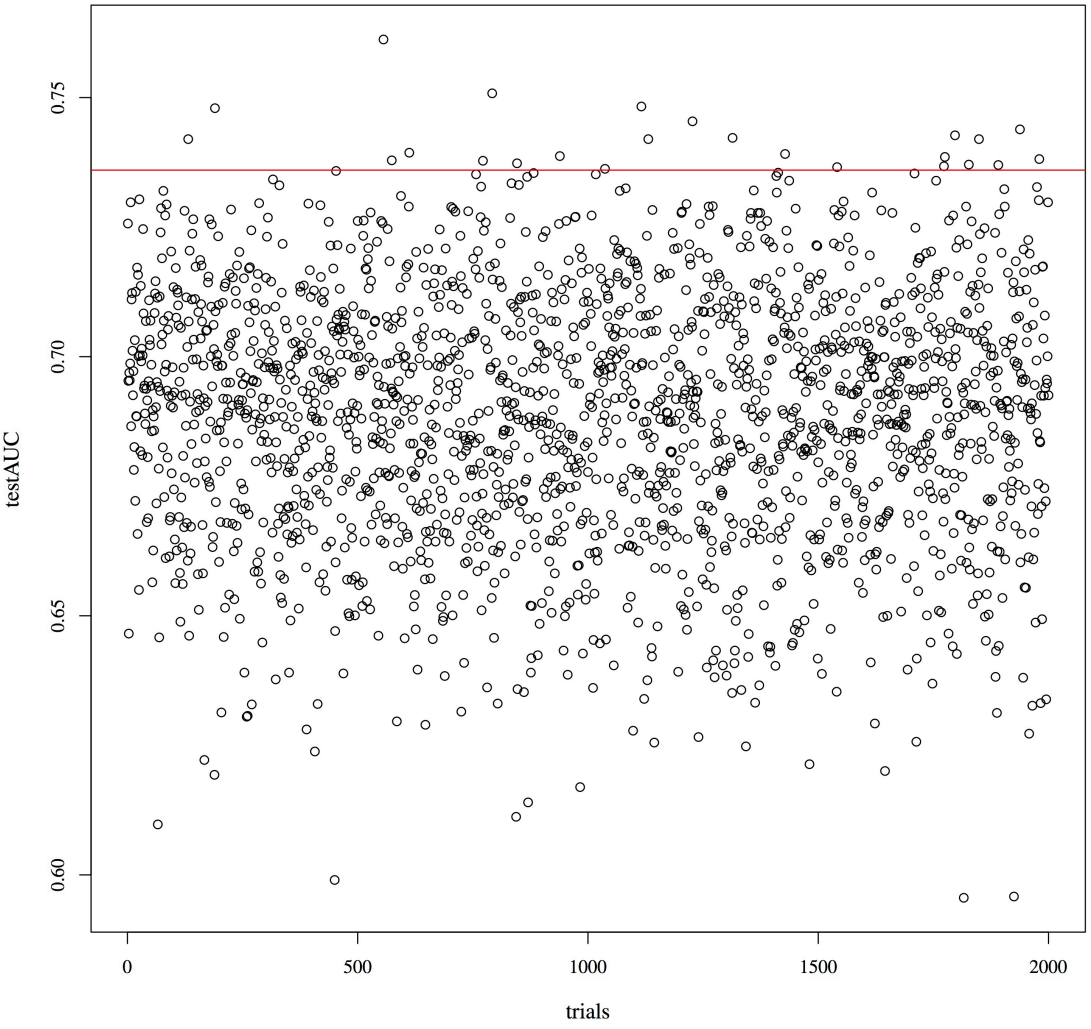

Figure 4.8: Test AUCs by the quasi-linear score for the dataset from Buyse et al. (2006). The score is learning by randomly divided genes subsets for the dataset from van't Veer et al (2002). The red line is the test AUC by the quasi-linear score, which consists of subsets of genes clustered by unsupervised learning.

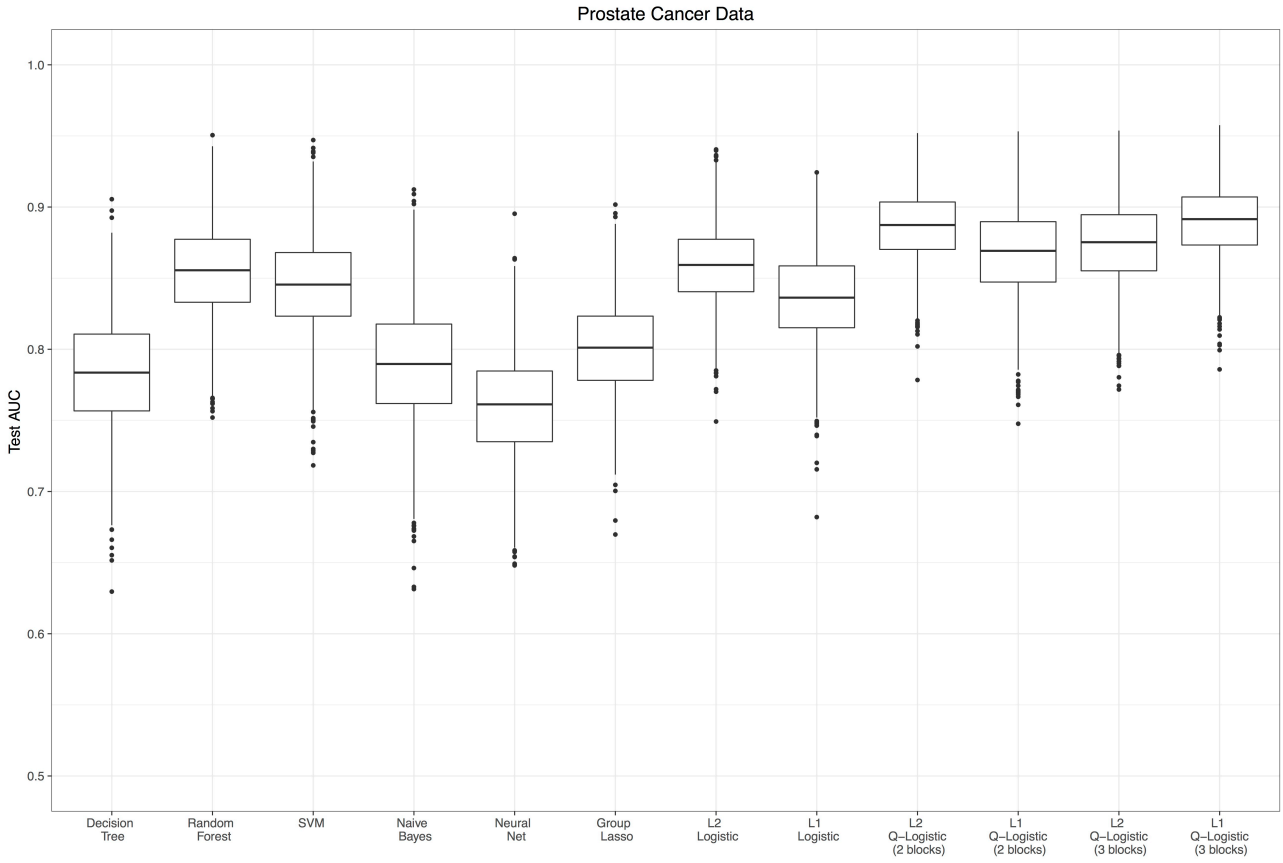

Figure 4.9: Box plots of the test AUC for all comparative methods for prostate cancer data. The vertical axis corresponds to the test AUC values. These values are estimated by bootstrap sampling of from test dataset.

Chapter 5

Quasi-linear relative risk model

In a relative risk model, log hazard and covariates are connected via a linear predictor. In this chapter, we consider an extension of the relative risk model by the quasi-linear predictor. In Section 5.1, we briefly derive the quasi-linear relative risk model. More theoretical discussions are given in Appendix A.3. In Section 5.2, we discuss the normalizing expression of the quasi-linear relative risk model. In Section 5.3 and 5,4, we derive the partial likelihood of the quasi-linear relative risk model and develop the parameter estimation algorithm in the basis of the maximum partial likelihood estimator as Cox's proportional hazard model. Moreover, as is the case of the quasi-linear logistic model, we discuss the L_1, L_2 and cross-L_1 penalized method. In Section 5.5 and 5.6, we give simulation and application studies of the quasi-linear relative risk model. We close Section 5.6 with the discussion on the quasi-linear relative risk model.

5.1 Formulation of relative risk model

As noted in Section 3.2, the relative risk model is written as

$$h(t|\boldsymbol{X},\boldsymbol{\theta}) = h_0(t)r(\boldsymbol{X},\boldsymbol{\theta}), \tag{5.1.1}$$

where $\boldsymbol{X}$ and $\boldsymbol{\theta}$ is a vector of covariates and parameters. The term $h_0(t)$ and $r(\boldsymbol{X},\boldsymbol{\theta})$ is often called baseline hazard and relative risk function. Cox (1972) especially focused on the

linear predictor for a log relative risk model as $\log r(\boldsymbol{X}, \theta) = F(\boldsymbol{X}; 0, \boldsymbol{\beta})$. We substitute the linear predictor for the quasi-linear predictor to define a quasi-linear relative risk model:

$$\begin{aligned} h(t|\boldsymbol{X}, \boldsymbol{\theta}) &= h_0(t) \exp(F_\tau) \\ &= h_0(t) \left(\frac{1}{K} \sum_{k=1}^{K} \exp(\tau\alpha_k + \tau\beta_k^\top \boldsymbol{X}) \right)^{\frac{1}{\tau}}. \end{aligned} \tag{5.1.2}$$

5.2 Normalized expression of quasi-linear relative risk model

Cox's proportional hazard model usually described as the intercept-excluded form:

$$h_C(t|\boldsymbol{X}, \boldsymbol{\beta}) = h_0(t) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}). \tag{5.2.1}$$

This is because the intercept-included form is verbose in parameterization as

$$\begin{aligned} h_C(t|\boldsymbol{X}, \boldsymbol{\beta}) &= h_0'(t) \exp(\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}) \\ &= \{h_0'(t) \exp(\alpha - c)\} \exp(c + \boldsymbol{\beta}^\top \boldsymbol{X}) \\ &= h_0(t) \exp(c + \boldsymbol{\beta}^\top \boldsymbol{X}), \end{aligned} \tag{5.2.2}$$

where $h_0(t) = h_0'(t) \exp(\alpha)$. Then the intercept parameter is not identifiable. Often we regard $c = 0$ in equation (5.2.2) which is equivalent to the equation (5.2.1). Thus we call $h_0(t)$ as "baseline hazard" because $h(t|\mathbf{0}, \boldsymbol{\beta}) = h_0(t)$, where $\boldsymbol{X} = 0$ is regarded as the reference values of all covariates. In a similar meaning, the intercept parameter of the quasi-linear relative risk model should have the property that

$$\begin{aligned} h(t|\mathbf{0}, \boldsymbol{\theta}) &= h_0(t) \left(\frac{1}{K} \sum_{k=1}^{K} \exp(\tau\alpha_k) \right)^{1/\tau} \\ &= h_0(t) \end{aligned} \tag{5.2.3}$$

for any $\tau \neq 0$. It follows that $\sum_{k=1}^{K} \exp(\tau\alpha_k) = K$ from the equation (5.2.3). Thus we have the normalized expression of the quasi-linear relative risk model as

$$\begin{aligned} h(t|\boldsymbol{X}, \boldsymbol{\theta}) &= h_0(t) \left(\sum_{k=1}^{K} \left(\frac{\exp(\tau\alpha_k)}{\sum_{k=1}^{K} \exp(\tau\alpha_k)} \right) \exp\left(\tau\beta_k^\top \boldsymbol{X}\right) \right)^{\frac{1}{\tau}} \\ &= h_0(t) \left(\sum_{k=1}^{K} \pi_k \exp\left(\tau\beta_k^\top \boldsymbol{X}\right) \right)^{\frac{1}{\tau}} \end{aligned} \tag{5.2.4}$$

where π_k's are non-zero probability weights with $\sum_{k=1}^{K} \pi_k = 1$.

Consider the case of the tuning parameter τ is equal to 1. Then, the quasi-linear relative risk model is written as

$$\begin{aligned} h(t|\boldsymbol{X}, \boldsymbol{\theta}) &= h_0(t) \left(\sum_{k=1}^{K} \pi_k \exp\left(\beta_k^\top \boldsymbol{X}\right) \right) \\ &= \sum_{k=1}^{K} \pi_k h_0(t) \exp\left(\beta_k^\top \boldsymbol{X}\right) \\ &= \sum_{k=1}^{K} \pi_k h_k(t), \end{aligned} \tag{5.2.5}$$

where $h_k(t) = h_0(t) \exp\left(\beta_k^\top \boldsymbol{X}\right)$. We therefore regard the quasi-linear hazard model as the mixture of hazard models when $\tau = 1$. Rosen and Tanner (1999) derived the mixtures of proportional hazards regression models (5.2.5) in the context of the mixture of experts model. They therefore assumed that the weight parameters π_k $(k = 1, 2, \cdots, K)$ follow the multinomial distribution.

Below, for easy notations, we parameterize the quasi-linear relative risk model as

$$h(t|\boldsymbol{X}, \boldsymbol{\theta}) = h_0(t) \left(\sum_{k=1}^{K} \exp\left(\alpha_k + \tau\boldsymbol{\beta}_k^\top \boldsymbol{X}\right) \right)^{1/\tau}, \tag{5.2.6}$$

where $\sum_{k=1}^{K} \exp(\alpha_k) = 1$. In this formulation, we get that the quasi-linear relative risk model is characterized by the modification of the relative risk model whose relative risk

function is replaced by

$$r(\boldsymbol{X}, \boldsymbol{\theta}) \quad = \quad \exp(F_\tau(\boldsymbol{X}, \boldsymbol{\alpha}/\tau, \boldsymbol{\beta})). \tag{5.2.7}$$

5.3 Partial likelihood of quasi-linear relative risk model

Consider the data $(\boldsymbol{X}_i, t_i, \delta_i)$ $(i = 1, 2, \cdots, N)$, where $\boldsymbol{X}_i$ is a p-dimensional covariates vector, t_i is observed survival time to some event, and δ_i is an event indicator which takes value 1 if the sample experiences the event by $t = t_i$ and value 0 otherwise. We assume that t_i and δ_i are independent for all $i = 1, 2, \cdots, N$. Then, the partial log likelihood of the quasi-linear relative risk model is written as

$$l \quad = \quad \sum_{i=1}^{N} \delta_i \left\{ \frac{1}{\tau} \log \left(\sum_{k=1}^{K} \tilde{\eta}_{ik} \right) - \log \left(\sum_{\ell \in R(t_i)} \left(\sum_{k=1}^{K} \tilde{\eta}_{i\ell} \right)^{1/\tau} \right) \right\} \tag{5.3.1}$$

from the equation (3.2.11), where $R(t_i) = \{l \in \{1, \cdots, N\} | t_i \leq t_\ell\}$ and $\tilde{\eta}_{ik} = \exp\left(\alpha_k + \tau \boldsymbol{\beta}_k^\top \boldsymbol{X}_i\right)$. Because the maximum partial likelihood estimator of (5.3.1) cannot be calculated analytically, we need the numerical optimization method as gradient method or newton-raphson method. We get from (5.2.7) that

$$\begin{aligned} \frac{\partial r}{\partial \boldsymbol{\theta}} \quad &= \quad \frac{1}{\tau} \left\{ \sum_{k=1}^{K} \exp\left(\alpha_k + \tau \boldsymbol{\beta}_k^\top \boldsymbol{X}\right) \right\}^{\frac{1}{\tau} - 1} \begin{pmatrix} \exp\left(\alpha_1 + \tau \boldsymbol{\beta}_1^\top \boldsymbol{X}\right) \\ \vdots \\ \exp\left(\alpha_K + \tau \boldsymbol{\beta}_K^\top \boldsymbol{X}\right) \\ \tau \boldsymbol{X} \exp\left(\alpha_1 + \tau \boldsymbol{\beta}_1^\top \boldsymbol{X}\right) \\ \vdots \\ \tau \boldsymbol{X} \exp\left(\alpha_K + \tau \boldsymbol{\beta}_K^\top \boldsymbol{X}\right) \end{pmatrix} \\ &= \quad \frac{1}{\tau} \left\{ \sum_{k=1}^{K} \tilde{\eta}_k \right\}^{\frac{1}{\tau} - 1} \boldsymbol{R}, \end{aligned} \tag{5.3.2}$$

where $\boldsymbol{R} = \left(\tilde{\eta}_1, \cdots, \tilde{\eta}_K, \tau \boldsymbol{X}^\top \tilde{\eta}_1, \cdots, \tau \boldsymbol{X}^\top \tilde{\eta}_K\right)^\top$, and

$$\begin{aligned}
\frac{\partial^2 r}{\partial\theta\partial\theta^\top} &= \frac{1}{\tau}\left(\frac{1}{\tau}-1\right)\left(\sum_{k=1}^{K}\tilde{\eta}_k\right)^{\frac{1}{\tau}-2}\boldsymbol{R}\boldsymbol{R}^\top \\
&+ \frac{1}{\tau}\left\{\sum_{k=1}^{K}\tilde{\eta}_k\right\}^{\frac{1}{\tau}-1}\begin{pmatrix} 1 & \tau\boldsymbol{x}^\top \\ \tau\boldsymbol{x} & \tau^2\boldsymbol{x}\boldsymbol{x}^\top \end{pmatrix}\otimes \mathrm{diag}(\boldsymbol{R}). \qquad (5.3.3)
\end{aligned}$$

Therefore, the score function, observed Fisher information matrix and expected Fisher information matrix are written as

$$\begin{aligned}
\mathcal{U}(\boldsymbol{\theta}) &= \sum_{i=1}^{N}\delta_i\left\{\frac{\frac{\partial}{\partial\boldsymbol{\theta}}r(\boldsymbol{X}_i,\boldsymbol{\theta})}{r(\boldsymbol{X}_i,\boldsymbol{\theta})}-\frac{\sum_{\ell\in R(t_i)}\frac{\partial}{\partial\boldsymbol{\theta}}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})}{\sum_{\ell\in R(t_i)}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})}\right\} \\
&= \sum_{i=1}^{N}\delta_i\frac{1}{\tau}\left(\frac{R}{\sum_{k=1}^{K}\tilde{\eta}_{ik}}-\frac{\sum_{\ell\in R(t_i)}(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{\frac{1}{\tau}-1}R}{\sum_{\ell\in R(t_i)}(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{\frac{1}{\tau}}}\right) \qquad (5.3.4) \\
\mathcal{I}(\boldsymbol{\theta}) &= -\sum_{i=1}^{N}\delta_i\left\{\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}r(\boldsymbol{X}_i,\boldsymbol{\theta})\{r(\boldsymbol{X}_i,\boldsymbol{\theta})\}-\frac{\partial}{\partial\boldsymbol{\theta}}r(\boldsymbol{X}_i,\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}^\top}r(\boldsymbol{X}_i,\boldsymbol{\theta})}{\{r(\boldsymbol{X}_i,\boldsymbol{\theta})\}^2}\right. \\
&- \left.\frac{\sum_{\ell\in R(t_i)}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\{r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\}-\frac{\partial}{\partial\boldsymbol{\theta}}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}^\top}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})}{\{\sum_{\ell\in R(t_i)}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\}^2}\right\} \\
&= \sum_{i=1}^{N}\delta_i\frac{1}{\tau}\left[\left\{(\frac{1}{\tau}-1)(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{-2}RR^\top+(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{-1}\boldsymbol{P}\right\}\right. \\
&- \left.\frac{\sum_{\ell\in R(t_i)}-(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{2/\tau-2}\boldsymbol{R}\boldsymbol{R}^\top+(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{2/\tau-1}\boldsymbol{P}}{(\sum_{\ell\in R(t_i)}(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{1/\tau})^2}\right] \qquad (5.3.5) \\
\mathcal{F}(\boldsymbol{\theta}) &= \sum_{i=1}^{N}\delta_i\left\{\frac{\sum_{\ell\in R(t_i)}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\{r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\}-\frac{\partial}{\partial\boldsymbol{\theta}}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}^\top}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})}{\{\sum_{\ell\in R(t_i)}r(\boldsymbol{X}_\ell,\boldsymbol{\theta})\}^2}\right\} \\
&= \sum_{i=1}^{N}\delta_i\frac{1}{\tau}\left\{\frac{-\sum_{\ell\in R(t_i)}(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{2/\tau-2})\boldsymbol{R}\boldsymbol{R}^\top+(\sum_{k=1}^{K}\tilde{\eta}_{ik})^{2/\tau-1}\boldsymbol{P}}{(\sum_{\ell\in R(t_i)}r(\boldsymbol{X}_\ell,\boldsymbol{\theta}))^2}\right\} \qquad (5.3.6)
\end{aligned}$$

from the equation (3.2.12), (3.2.14) and (3.2.15), where $\boldsymbol{P} = \begin{pmatrix} 1 & \tau\boldsymbol{x}^\top \\ \tau\boldsymbol{x} & \tau^2\boldsymbol{x}\boldsymbol{x}^\top \end{pmatrix}\otimes \mathrm{diag}(\boldsymbol{R})$. These formulations are common among general relative risk models except for the difference in relative risk function. If the relative risk function is the exponential relative risk (3.2.4), then the observed information matrix and the expected information matrix coincides while these differ in other cases. Aalen *and others* (2008) insist that the expected information

matrix tends to be stable among the two because it depends only on the summary amount over the risk sets, while the observed information matrix depends on the covariate values of the individuals who experience events. We therefore use the expected one when we need the second derivatives of the log-likelihood on each parameter estimation step in iteratively method as Newton-Raphson algorithm.

5.4 Parameter estimation in quasi-linear relative risk model

5.4.1 Parameter estimation without regularization

The parameter estimation without regularization is performed by Newton-Raphson method with Lagrange multiplier. The objective function is given as

$$Q = l(\boldsymbol{\theta}) - \lambda g(\boldsymbol{\theta}), \tag{5.4.1}$$

where $\lambda \in \mathbb{R}$ and $g(\boldsymbol{\theta}) = \sum_{k=1}^{K} \exp(\alpha_k) - 1$. By a Taylor expansion, we get that

$$\begin{aligned}
Q &\approx l(\boldsymbol{\theta}^{(t)}) + \frac{\partial l(\boldsymbol{\theta}^{(t)})^\top}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \frac{\partial^2 l(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) \\
&- \lambda^{(t)} \left\{ g(\boldsymbol{\theta}^{(t)}) + \frac{\partial g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \frac{\partial^2 g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) \right\} \\
&- (\lambda - \lambda^{(t)}) \left\{ g(\boldsymbol{\theta}^{(t)}) + \frac{\partial g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \frac{\partial^2 g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) \right\}.
\end{aligned}$$

The first derivatives of the objective functions are given as

$$\begin{aligned}
\frac{\partial Q}{\partial \boldsymbol{\theta}} &\approx \frac{\partial l(\boldsymbol{\theta}^{(t)})^\top}{\partial \boldsymbol{\theta}} + \frac{\partial^2 l(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) - \lambda^{(t)} \left\{ \frac{\partial g(\boldsymbol{\theta}^{(t)})^\top}{\partial \boldsymbol{\theta}} + \frac{\partial^2 g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) \right\} \\
&- (\lambda - \lambda^{(t)}) \left\{ \frac{\partial g(\boldsymbol{\theta}^{(t)})^\top}{\partial \boldsymbol{\theta}} + \frac{\partial^2 g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) \right\}, \qquad (5.4.2) \\
\frac{\partial Q}{\partial \lambda} &\approx - \left\{ g(\boldsymbol{\theta}^{(t)}) + \frac{\partial g(\boldsymbol{\theta}^{(t)})^\top}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) \right\}. \qquad (5.4.3)
\end{aligned}$$

We get the maximum partial likelihood estimator by updating parameters from some proper initial values $\boldsymbol{\theta}^{(0)}$ by

$$\begin{pmatrix} \boldsymbol{\theta}^{(t+1)} \\ \lambda^{(t+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}^{(t)} \\ \lambda^{(t)} \end{pmatrix} + B^{-1}b, \tag{5.4.4}$$

where

$$B = \begin{pmatrix} \mathcal{F}(\boldsymbol{\theta}^{(t)}) - \lambda^{(t)} \frac{\partial^2 g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} & -\frac{\partial g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} \\ \frac{\partial g(\boldsymbol{\theta}^{(t)})^\top}{\partial \boldsymbol{\theta}} & 0 \end{pmatrix}, \tag{5.4.5}$$

$$b = \begin{pmatrix} -\mathcal{U}(\boldsymbol{\theta}^{(t)}) + \lambda^{(t)} \frac{\partial g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} \\ -g(\boldsymbol{\theta}^{(t)}) \end{pmatrix}. \tag{5.4.6}$$

5.4.2 Parameter estimation with L_1 and L_2 penalty

When the dimension of covariates vector is very high, the parameter estimation of the quasi-linear relative risk model often becomes unstable as the learning in ordinary models. This problem can be avoided by the regularization by L_2-norm, ridge penalty. With the additional penalty for equation (5.4.1), the object function is written as

$$Q_2 = l(\boldsymbol{\theta}) - \lambda g(\boldsymbol{\theta}) - \frac{1}{2} \sum_{k=1}^{K} \epsilon_k \boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k. \tag{5.4.7}$$

Then, the update formula of parameters are the same with the equation (5.4.4) other than that the equation (5.4.5) and (5.4.6) are replaced by

$$B = \begin{pmatrix} \mathcal{F}_2(\boldsymbol{\theta}^{(t)}) - \lambda^{(t)} \frac{\partial^2 g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} & -\frac{\partial g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} \\ \frac{\partial g(\boldsymbol{\theta}^{(t)})^\top}{\partial \boldsymbol{\theta}} & 0 \end{pmatrix}, \tag{5.4.8}$$

$$b = \begin{pmatrix} -\mathcal{U}_2(\boldsymbol{\theta}^{(t)}) + \lambda^{(t)} \frac{\partial g(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} \\ -g(\boldsymbol{\theta}^{(t)}) \end{pmatrix}. \tag{5.4.9}$$

where

$$\mathcal{U}_2(\boldsymbol{\theta}^{(t)}) = \mathcal{U}(\boldsymbol{\theta}^{(t)}) - \begin{pmatrix} \mathbf{0}_K \\ \epsilon_1 \boldsymbol{\beta}_1 \\ \vdots \\ \epsilon_K \boldsymbol{\beta}_K \end{pmatrix}, \tag{5.4.10}$$

$$\mathcal{F}_2(\boldsymbol{\theta}^{(t)}) = \mathcal{F}(\boldsymbol{\theta}^{(t)}) - \begin{pmatrix} \mathbf{0}_{KK} & \\ & \mathrm{diag}(\boldsymbol{\epsilon}) \end{pmatrix} \tag{5.4.11}$$

with $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_K)^\top$.

For the sparse estimation, we add a L_1 penalty which is known as *lasso* in the linear model. Moreover, L_1 and L_2 penalty can be combined. The combined penalty is called elastic net penalty (Zou and Hastie, 2005). The object function with these penalties is written as

$$Q = l(\boldsymbol{\theta}) - \rho \sum_{k=1}^{K} \epsilon_{1k} ||\boldsymbol{\beta}_k||_1 - (1-\rho)\frac{1}{2}\sum_{k=1}^{K} \epsilon_{2k} \boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k, \tag{5.4.12}$$

where $0 \leq \rho \leq 1$, ϵ_{1k} and ϵ_{2k} $(k = 1, \cdots, K)$ are positive parameters. We use the parameter estimation procedure of the full gradient algorithm proposed by (Goeman, 2010). The update formulae is written as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \min\{t_{\mathrm{opt}}(\boldsymbol{\theta}^{(t)}), t_{\mathrm{edge}}(\boldsymbol{\theta}^{(t)})\}\boldsymbol{d}(\boldsymbol{\theta}^{(t)}), \tag{5.4.13}$$

where $\boldsymbol{d}(\boldsymbol{\theta}) = (d_1(\boldsymbol{\theta}), \cdots, d_{p+K}(\boldsymbol{\theta}))^\top$,

$$t_{\mathrm{edge}}(\boldsymbol{\theta}) = \min_{1+K \leq j \leq p+K} \left(-\frac{\theta_j}{d_j(\boldsymbol{\theta})} : \mathrm{sign}(\theta_j) = -\mathrm{sign}(d_j(\boldsymbol{\theta})) \neq 0 \right)$$

and

$$t_{\mathrm{opt}}(\boldsymbol{\theta}) = \frac{|d(\boldsymbol{\theta})|}{d(\boldsymbol{\theta})^\top \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} d(\boldsymbol{\theta})}.$$

Here $d_j(\boldsymbol{\theta}) = \dot{l}_j(\boldsymbol{\theta})$ for $j = 1, \cdots, K$ and

$$d_j(\boldsymbol{\theta}) = \begin{cases} \dot{l}_j(\boldsymbol{\theta}) - \rho\epsilon_{1k}\text{sign}(\theta_j) - (1-\rho)\epsilon_{2k}\theta_j & \text{if } \theta_j \neq 0 \\ \dot{l}_j(\boldsymbol{\theta}) - \rho\epsilon_{1k}\text{sign}(\dot{l}_j(\boldsymbol{\theta})) - (1-\rho)\epsilon_{2k}\theta_j & \text{if } \theta_j = 0 \text{ and } |\dot{l}_j(\boldsymbol{\theta}) - (1-\rho)\epsilon_{2k}\theta_j| > \rho\epsilon_{1k} \\ 0 & \text{otherwise} \end{cases}$$

for $j = K+1, \cdots, p+K$, where sign(z) is a sign function, $\dot{l}_j$ is the j-th component of the score function and k denotes the cluster number that the j-th covariate belongs to. In each step, t_{opt} provides the optimal solution of the gradient descent algorithm and t_{edge} controls the direction of the gradient so as not to change the signs of parameters. The vector of the tuning parameters $(\epsilon_1, \cdots, \epsilon_K)^\top$ is determined by Bayes Information Criteria (BIC) defined as

$$\text{BIC} = -2l(\boldsymbol{\theta}) + \left\{ (K-1) + \sum_{k=1}^{K}\sum_{j=1}^{p} I(\beta_{kj} \neq 0) \right\} \log N \tag{5.4.14}$$

5.4.3 Parameter estimation with cross-L_1 penalty

To get the parsimonious expression as discussed in the development of the quasi-linear logistic model, we derive the log likelihood function with cross-L_1 penalty for the quasi-linear relative risk model. It is defined as

$$l^{pen}(\boldsymbol{\theta}) \quad = \quad l(\boldsymbol{\theta}) - P(\boldsymbol{\beta}) - P^c(\boldsymbol{\beta}), \tag{5.4.15}$$

where $P(\boldsymbol{\beta})$ is an elastic net penalty and $P^{(c)}(\boldsymbol{\beta})$ is penalty function defined by (4.3.8). The maximization of the objective function (5.4.15) is achieved by the full gradient algorithm ((Goeman, 2010)) as the quasi-linear logistic model.

5.5 Simulation study of quasi-linear relative risk model

In this section, we show the results of the simulation studies for the quasi-linear relative risk model with cross-L_1 penalty. In all simulation studies introduced here, the inverse function

method was used for data generation process.

First, the covariates $\boldsymbol{X}$ were generated from the multivariate normal distribution with mean $\mathbf{0}$ and variance matrix $2\boldsymbol{I}$, the apparent censored time T_1 was generated from the exponential distribution with mean 1000 and the random variable U was generated from the uniform distribution on $[0, 1]$. Let the baseline survival time be followed the exponential distribution with mean 100. Then, the true survival time corresponding to the log relative risk function $F(\boldsymbol{X}; \boldsymbol{\theta})$ was given as $T_2 = -(\log(U)/100) \exp(F(\boldsymbol{X}; \boldsymbol{\theta}))$. Based on T_1 and T_2, let $T = \min(T_1, T_2)$ be the observational survival time and $\delta = I(T_1 < T_2)$ be the censored indicator before event time.

The simulation studies are roughly divided into two scenarios. In the first scenario, we assumed that the log relative risk function was expressed by the quasi-linear form of the disjoint covariates combination. In the next scenario, we assumed that log relative risk function was expressed by the quasi-linear form of combination of covariates with overlap.

The sample size were set to $N = 100$ or $N = 200$ in all scenarios. The number of the clusters were determined from 2 or 3 for all settings by the BIC with the tuning parameter λ of cross-L_1 penalty. The tuning parameter λ was determined from some candidates. We note that the maximum candidate value of the tuning parameter λ was controlled sufficiently to achieve the restricted quasi-linear form for every setting.

- *Simulation − Disjoint*

  We performed simulations of 4 scenarios as follows. The true relative risk function forms the restricted quasi-linear function as discussed in the Chapter 4.

  1. $K = 2$, $\tau = 1$, $\boldsymbol{\theta} = (\log(0.7), \log(0.3), (0, 0, 1, 1), (1, 1, 0, 0))^\top$.

  2. $K = 2$, $\tau = 3$, $\boldsymbol{\theta} = (\log(0.7), \log(0.3), (0, 0, 1, 1), (1, 1, 0, 0))^\top$.

  3. $K = 3$, $\tau = 1$, $\boldsymbol{\theta} = (\log(0.2), \log(0.3), \log(0.5), (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 1))^\top$.

  4. $K = 3$, $\tau = 3$, $\boldsymbol{\theta} = (\log(0.2), \log(0.3), \log(0.5), (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 1))^\top$.

- *Simulation − Overlap*

  We performed simulations of 4 scenarios as follows.

  1. $K = 2$, $\tau = 1$, $\boldsymbol{\theta} = (\log(0.7), \log(0.3), (0, 0, 1, 1), (1, 1, 1, 0))^\top$.

  2. $K = 2$, $\tau = 3$, $\boldsymbol{\theta} = (\log(0.7), \log(0.3), (0, 0, 1, 1), (1, 1, 1, 0))^\top$.

  3. $K = 3$, $\tau = 1$, $\boldsymbol{\theta} = (\log(0.2), \log(0.3), \log(0.5), (0, 0, 1, 1), (1, 1, 1, 0), (1, 0, 0, 1))^\top$.

  4. $K = 3$, $\tau = 3$, $\boldsymbol{\theta} = (\log(0.2), \log(0.3), \log(0.5), (0, 0, 1, 1), (1, 1, 1, 0), (1, 0, 0, 1))^\top$.

First, we checked whether the tuning parameter was correctly selected in each scenario. The results of the simulation-disjoint and simulation-overlap are summarized in Table 5.1 and 5.2, respectively. In all scenarios with small sample size ($N = 100$), the true tuning parameters τ and cluster size K were correctly selected in most cases (50%-84%). The accuracy was improved when the sample size was larger ($N = 200$, 71%-89%). Moreover, misspecifying of cluster sizes K and tuning parameter τ rarely happened in large sample

settings. These results empirically indicate that we have the consistency in the selection of tuning parameter by BIC. Second, we checked the consistency of the parameter estimation in disjoint and large sample settings. Figure 5.1, 5.2, 5.3 and 5.4 are box plots of the estimated parameters in Disjoint-1, 2, 3 and 4 settings for $N = 200$ samples. We can show that the parameter estimation procedure worked very well if the true tuning parameter τ and cluster size K were determined by BIC. Moreover, the bias for the coefficient vector was little even if the wrong tuning parameter τ were selected although the proportion parameters α's are relatively largely biased.

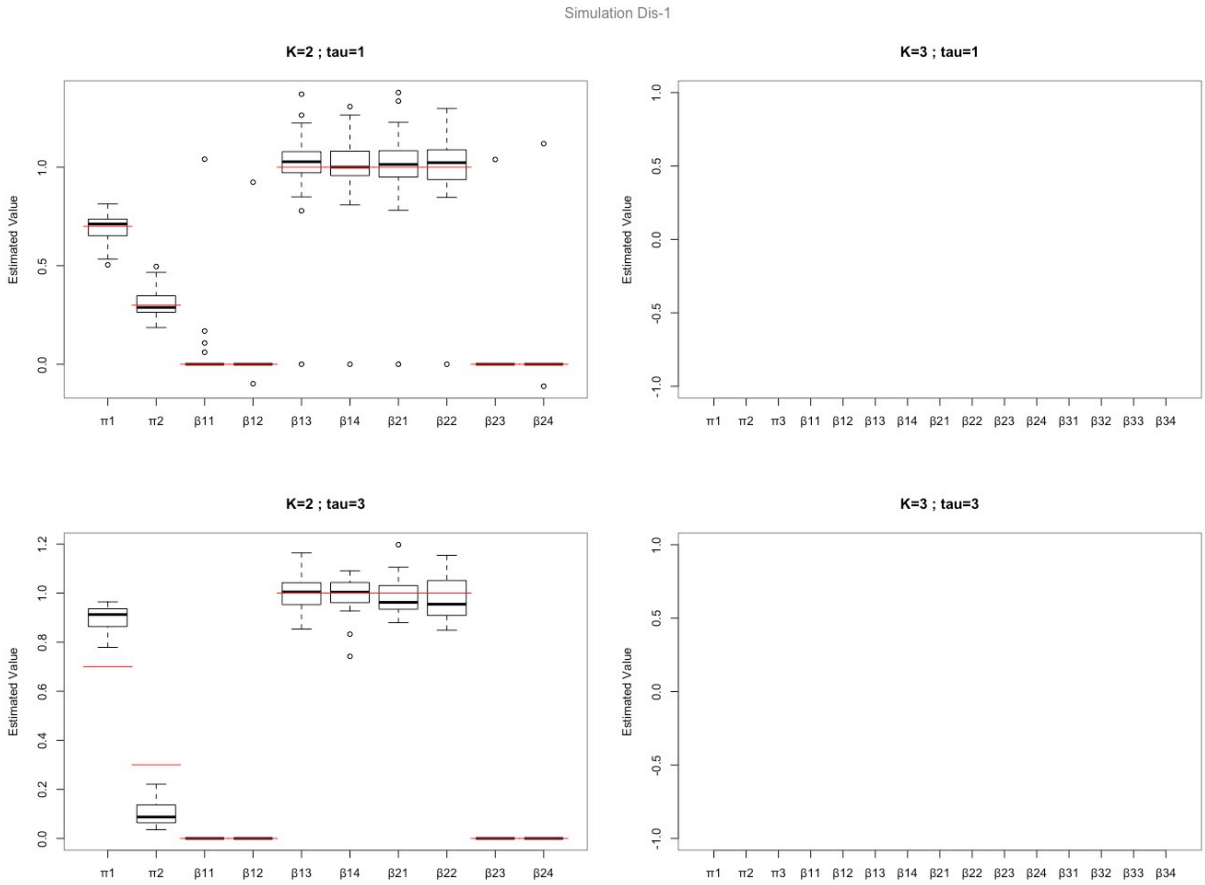

Figure 5.1: Box plot of the estimated parameters in simulation Disjoint-1. There are no plots in the right side figure because misspecifying of the cluster size did not occur. The red lines denotes the true value getting from data generation process.

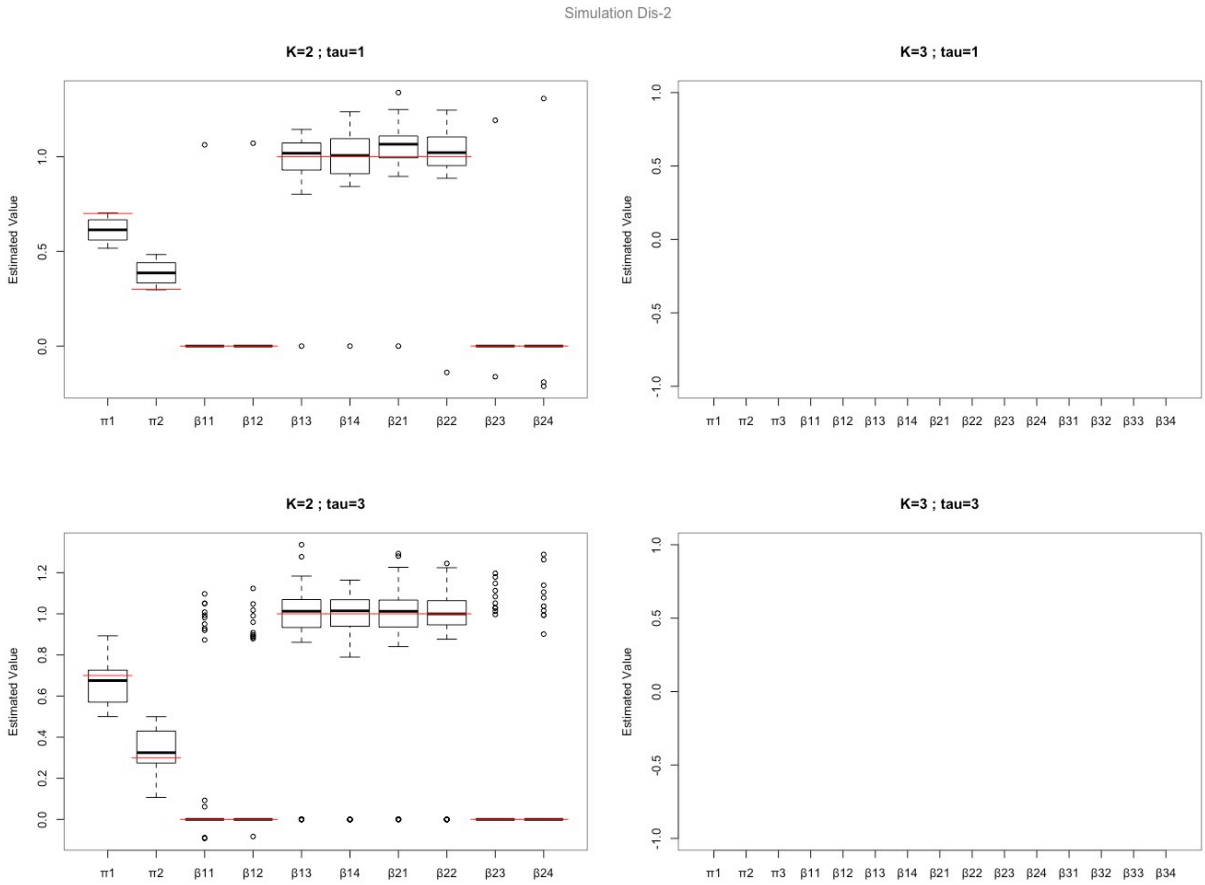

Figure 5.2: Box plot of the estimated parameters in simulation Disjoint-2. There are no plots in the right side figure because misspecifying of the cluster size did not occur. The red lines denotes the true value getting from data generation process.

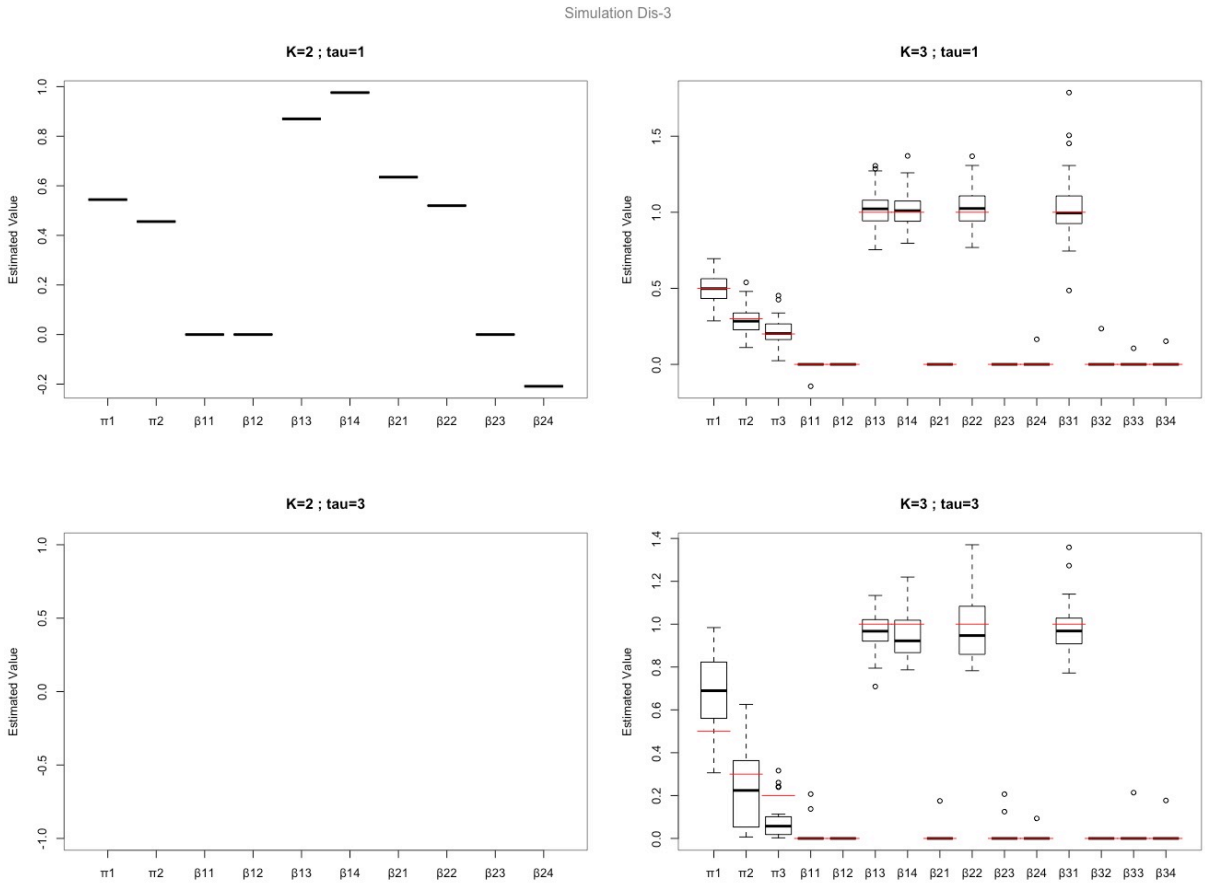

Figure 5.3: Box plot of the estimated parameters in simulation Disjoint-3. There are no plots in the figure of $K = 2$ and $\tau = 3$ because misspecifying as such tuning parameters did not occur. The red lines denotes the true value getting from data generation process.

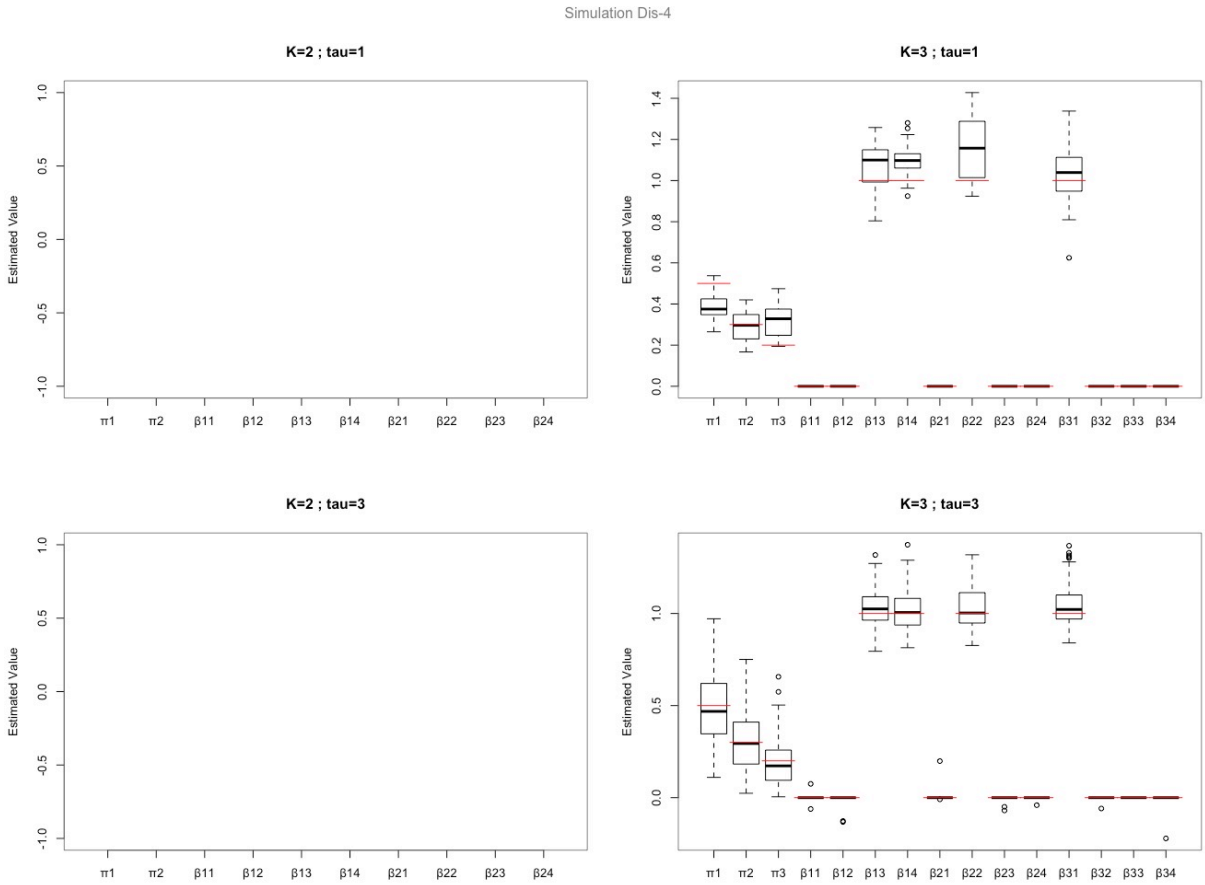

Figure 5.4: Box plot of the estimated parameters in simulation Disjoint-4. There are no plots in the left side figure because misspecifying of the cluster size did not occur. The red lines denotes the true value getting from data generation process.

5.6 Applications of quasi-linear relative risk model

In this section, we show the results of the application studies for two breast cancer datasets in order to evaluate the performance of the quasi-linear relative risk model. For all analysis in this section, the initial values of parameters were set to the equal probability weighting parameters $\exp(\alpha_k) = 1/K$ and the coefficients vector of Cox's proportional hazard models estimated from random K samples sets on the parameter estimation of the quasi-linear relative risk model. To evaluate the predictive ability of the learned model, we calculated the AUC of time dependent ROC (Heagerty *and others*, 2000) using test dataset. The predictive performance was compared between Cox's proportional hazard model and the quasi-linear relative risk model.

1. GBCS dataset

   The first dataset is German breast cancer research data (Hosmer and Lemeshow, 1989). For 680 breast cancer patients, number of progesterone receptor and estrogen receptor and overall survival time data with some censors are obtained in the dataset. First of all, all samples were randomly divided into two datasets for training and test. Learning of Cox's proportional hazard model and the quasi-linear relative risk model were carried out using the training dataset. The number of clusters K is 2 with 2 markers because of the purpose of visualization. The candidate value of τ were 1, 2, 3, 4, 5 and 6, and the candidate value for the tuning parameter λ of cross-L_1 penalty were 1, 2, 4, 8, 16, 32, 64. These tuning parameters were determined by BIC.

   As a result, the tuning parameter $\tau = 5$ and $\lambda = 1$ were selected by BIC. The test AUCs for some time points are drawn in Figure 5.5. For every time points, the test AUC of the quasi-linear relative risk model is higher than Cox's proportional hazard model. The learning curves are displayed in Figure 5.6. The linear and quasi-linear predictors give almost proportional risk scores for the high risk patients although the definitely different trends in the lower risk patients.

Table 5.1: Model selected probability (estimated by 100 simulations, N=100)

	$\tau = 1,\ K = 2$	$\tau = 3,\ K = 2$	$\tau = 1,\ K = 3$	$\tau = 3,\ K = 3$
Dis-1	72%	28%	0%	0%
Dis-2	37%	63%	0%	0%
Dis-3	16%	1%	59%	24%
Dis-4	0%	0%	44%	56%
Ove-1	66%	33%	1%	0%
Ove-2	38%	60%	1%	1%
Ove-3	29%	12%	50%	9%
Ove-4	0%	8%	8%	84%

Table 5.2: Model selected probability (estimated by 100 simulations, N=200)

	$\tau = 1,\ K = 2$	$\tau = 3,\ K = 2$	$\tau = 1,\ K = 3$	$\tau = 3,\ K = 3$
Dis-1	84%	16%	0%	0%
Dis-2	17%	83%	0%	0%
Dis-3	1%	0%	74%	25%
Dis-4	0%	0%	17%	83%
Ove-1	80%	19%	1%	0%
Ove-2	29%	71%	0%	0%
Ove-3	9%	2%	79%	10%
Ove-4	0%	0%	11%	89%

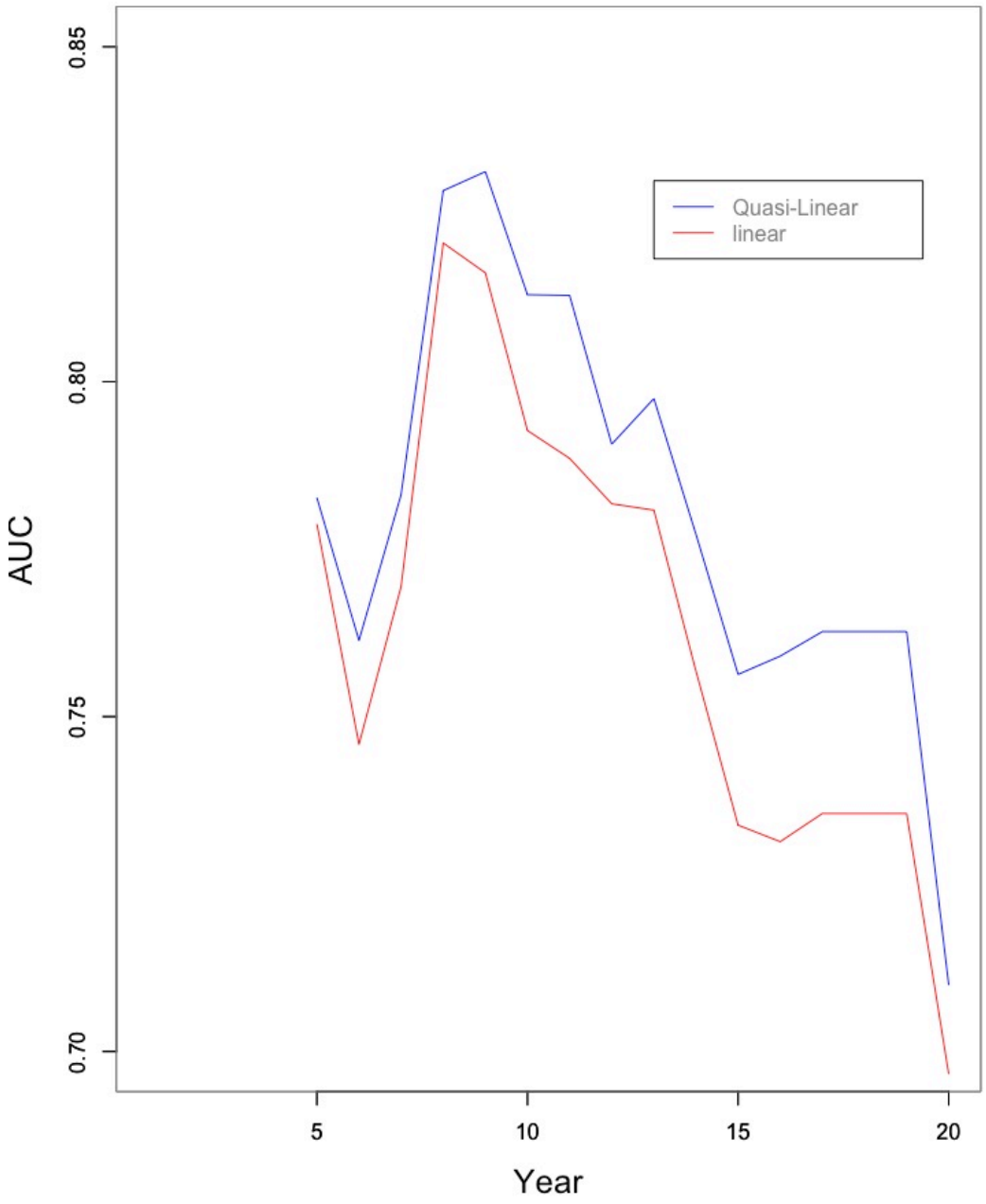

Figure 5.5: The test AUC changes in German breast cancer research dataset. Two line graphs show the AUC values at each time calculated from time dependent ROC for the linear (red) and quasi-linear (blue) predictor.

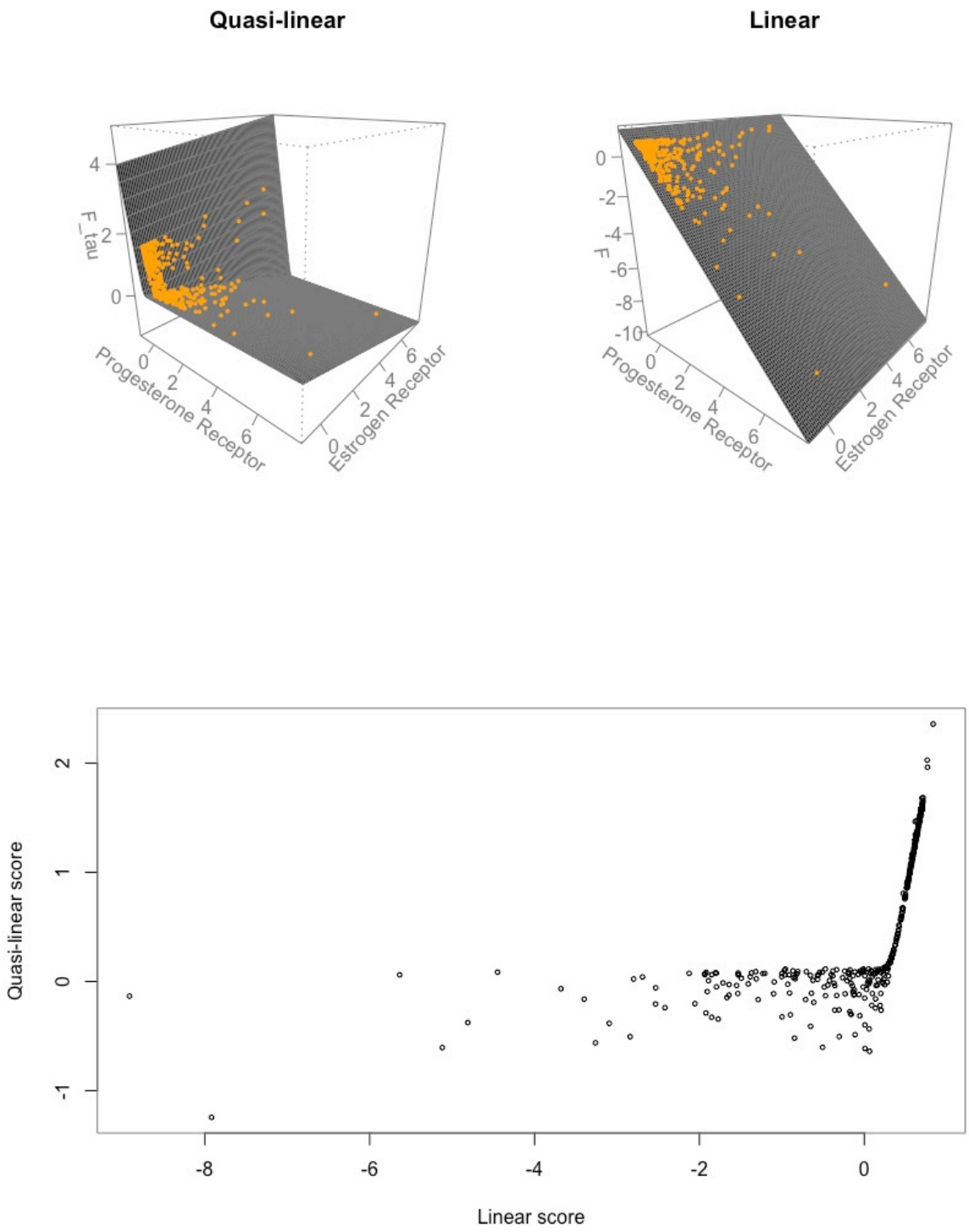

Figure 5.6: The upper figures show the learning surfaces and fitted score plots of the quasi-linear predictor and linear predictor. The lower figure shows the relationship between these fitted scores.

2. Seventy genes dataset

   The second dataset is the same dataset that we applied the quasi-linear logistic model in Section 4.5. We used the dataset from van't Veer *and others* (2002) as the training data and the dataset from Buyse *and others* (2006) as the test data. These include the gene expressions of 70 genes and survival time with some censors. Except for samples with missing values, there are 75 samples in training data and 220 samples in test data, respectively. In this application, we extracted the top 10 relevant genes to evaluate the model performance without L_1 and L_2 regularizations. Such marker preselection has been performed in many studies (Dettling and Bühlman, 2003). The candidate values of tuning parameter τ were 1, 2, 3, 4, 5 and 6, and the candidate values for the penalty coefficient λ of cross-L_1 penalty were 1, 2, 4, 8, 16. These tuning parameters were determined by BIC.

   As a result, the tuning parameter $\tau = 1$, the penalty coefficient $\lambda = 1$ and cluster size $K = 2$ were selected. The test AUCs for some time points are drawn in Figure 5.7. For every time points, the test AUC of the quasi-linear relative risk model is higher than Cox's proportional hazard model. The estimated coefficients for $K = 1$ (Linear), $2, 3$ and 4 (Quasi-linear) are displayed in Figure 5.8.

5.7 Discussion on quasi-linear relative risk model

We extended Cox's proportional hazard model by the quasi-linear predictor. We saw that the quasi-linear relative risk model is the generalization of the mixture hazard model proposed in the literatures. Thus it would have better performance for the modeling of the dataset with heterogeneous structure than Cox's proportional hazard model. Like the quasi-linear logistic model, the non-linearity of the quasi-linear predictor yielded the flexibility of the model.

In order to introduce the parsimonious expression in the quasi-linear relative risk model, we applied the model with cross-L_1 penalty while the disjoint expression were used in the application of the quasi-linear logistic model as discussed in section 4.5. We therefore need

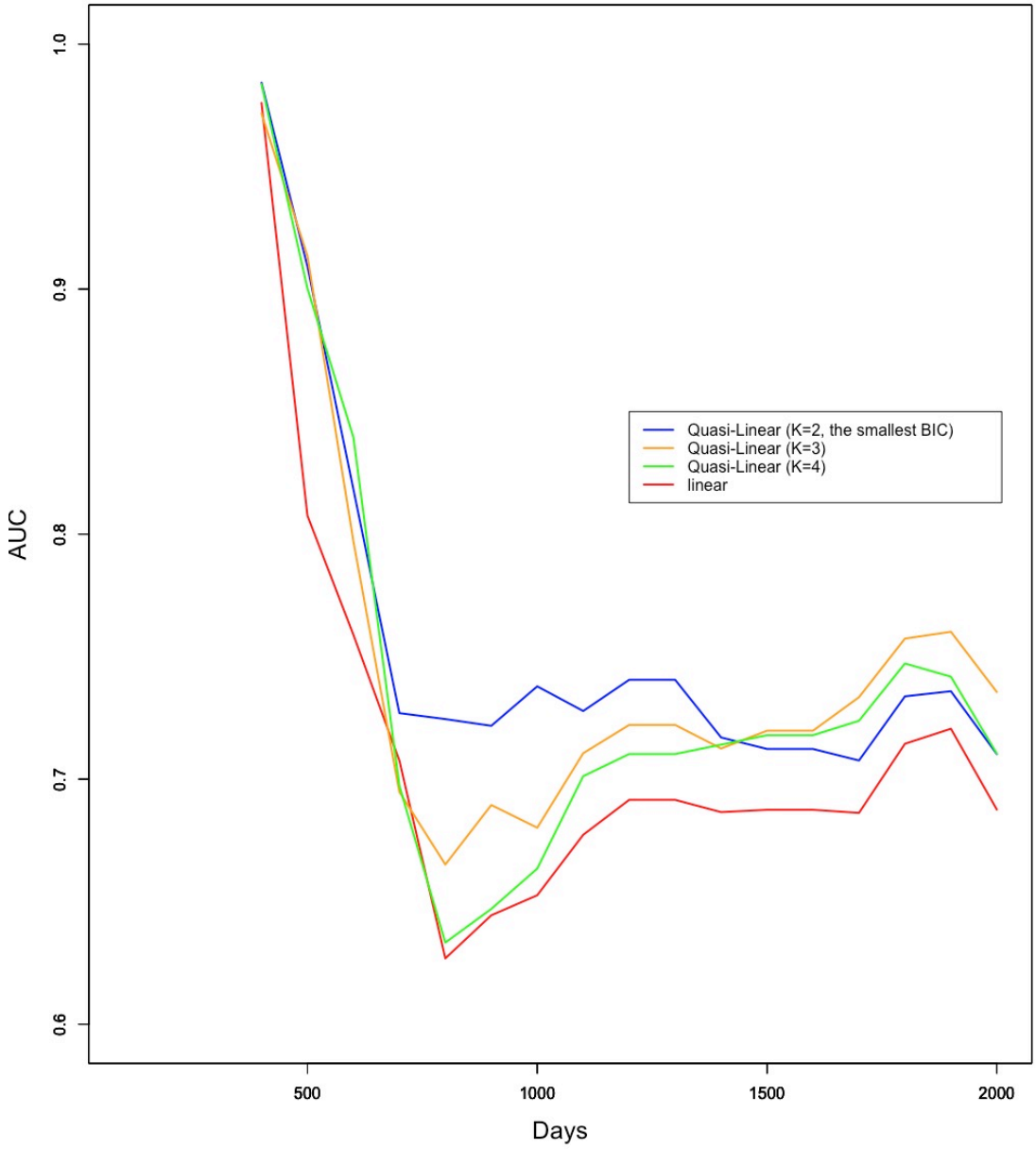

Figure 5.7: The test AUC changes in the dataset from Buyse *and others* (2006). Four line graphs show the AUC values at each time calculated from time dependent ROC for the linear (red) and quasi-linear (blue, K=2; yellow, K=3; green, K=4) predictor.

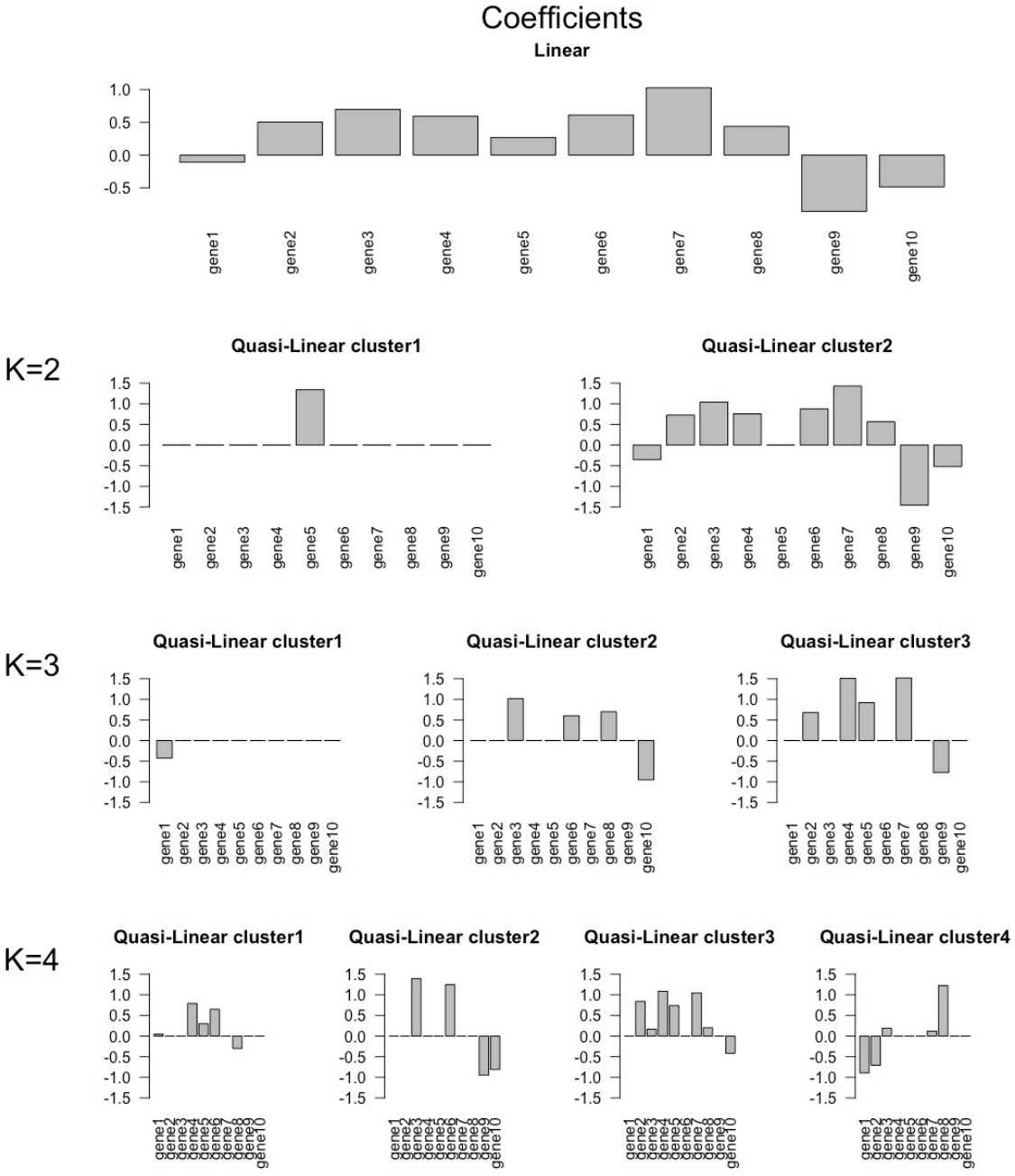

Figure 5.8: The estimated coefficients in the dataset from van't Veer *and others* (2002). In order from the top, the bar plots of the estimated values are displayed for Cox's proportional hazard model, the quasi-linear relative risk model ($K = 2$, 3 and 4).

the tuning of the regularization parameter of the cross-L_1 penalty. Our simulation studies showed that BIC is useful for the tuning of this parameter. It is good property that the misspecifying of cluster size K rarely happens in model selection. Although the misspecifying of tuning parameter τ sometimes occurs, the estimated score did not differ from the estimated score with true tuning parameter so much. These facts show that moderate large sample size sufficiently results in the inference of the appropriate non-linearity in the parameter family $\{F_\tau, 0 < \tau < \infty\}$.

The cross-L_1 penalty works well in the simulation and application studies. In fact, the true model were correctly selected and achieved consistent parameter estimation in the simulation study by cross-L_1 penalty and BIC. Although the application study showed that the different cluster sizes K result in the different trend in the estimated parameter as show in Figure 5.8. Such unstable phenomenon may be caused by small sample size in the training dataset. In fact, the dependency of the estimated values on the initial values on the estimation algorithm was observed in the small sample size situations in the simulation studies. It is the future task to solve the problem of the initial value dependency in the small sample size learning. However, while having such instability, the quasi-linear relative risk model achieved higher performance than Cox's proportional hazard model about predictive ability for the independent dataset. It seem to be the reason why the proposed method worked well that the form of the combination of the quasi-linear predictor, cross-L_1 penalty and BIC succeeded in the learning of the heterogeneous structure in the dataset.

Chapter 6

The roles, relationships and future works of quasi-linear form among ordinary methods

In this chapter, we discuss the roles of the quasi-linear form in the clustering methods and the relationships among quasi-linear, mixture of experts and neural network models. Moreover, we also refer to the future works related to the proposed method.

6.1 K-means and maximum entropy clustering

Clustering algorithms are divers. Fahad *and others* (2014) gave an overview of clustering taxonomy and they divide them into 5 categories: partitioning, hierarchical, density, grid and model based clustering algorithms. One of the most widely used and studied formulations is K-means algorithm (Kanungo *and others*, 2002). The K-means algorithm is one of the partitioning-based methods, that is, the method that all clusters are determined promptly. In particular, the K-means algorithm is called a hard clustering because it assigns all samples to one cluster from candidate clusters definitely. Let $\boldsymbol{x}_i$ be a covariate vector for i-th sample ($i = 1, 2, \cdots, N$). Then, the K-means clustering of K clusters is given as a

minimization problem of

$$\underset{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_K}{\operatorname{argmin}} \sum_{i=1}^{N} \min_k ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2, \tag{6.1.1}$$

where $\boldsymbol{\mu}_k$ is a center point for k-th cluster ($k = 1, 2, \cdots, K$). This is a minimization problem of loss function

$$L = \sum_{i=1}^{N} \min_k ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2. \tag{6.1.2}$$

There are some variants to solve the problem (6.1.1) while the most common algorithm is referred to as Lloyd's algorithm as follows (Kanungo *and others*, 2002). The optimization of (6.1.2) only takes the center point of the nearest cluster into account for each observation.

Set cluster centers $\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \cdots, \boldsymbol{\mu}_K^{(0)}$. For $t = 1, 2, \cdots$, repeat (1) and (2) until the algorithm converges for all $k = 1, 2, \cdots, K$.

(1) Assign each sample to the nearest cluster in the sense of Euclidean distance:

$$M_k^{(t)} = \{\boldsymbol{x}_i; ||\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t)}|| < ||\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t)}|| \ (\forall j \neq k)\} \tag{6.1.3}$$

(2) Update the center

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{|M_k^{(t)}|} \sum_{x_j \in M_k^{(t)}} \boldsymbol{x}_j \tag{6.1.4}$$

A fuzzy C-means algorithm (Dunn, 1973) is known as the soft clustering method based on the K-means clustering. The objective function of the C-means is defined as follows (Fahad *and others*, 2014):

$$E = \sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik}^m ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2, \tag{6.1.5}$$

where $\mu_{ik} = 1/\sum_{\ell=1}^{K} (\frac{|\boldsymbol{x}_i - \boldsymbol{\mu}_k|}{|\boldsymbol{x}_i - \boldsymbol{\mu}_\ell|})^{2/(m-1)}$ and $m(> 1)$ is a fuzzyness factor. Let consider an

opposite extreme case of the K-means:

$$E_0 = \sum_{i=1}^{N} \sum_{k=1}^{K} ||x_i - \mu_k||^2. \tag{6.1.6}$$

Minimizing equation (6.1.6), we get $\mu_k = \frac{1}{N} x_i$ for all k, corresponding to the smallest energy function of Euclidean distance when taking all cluster centroids for each observation into consideration, and produces one point center as the global mean. These two problems are connected by the generalized exp-mean:

$$E_\tau = \sum_{i=1}^{N} -\frac{1}{\tau} \log \left\{ \sum_{k=1}^{K} \exp \left\{ -\tau ||x_i - \mu_k||^2 \right\} \right\}, \tag{6.1.7}$$

where τ is a positive parameter. Applying Proposition 1, the equation (6.1.7) is equivalent to the equation (6.1.2) when τ goes to infinity. The energy function (6.1.7) was used as the objective function of the maximum entropy clustering proposed by Rose *and others* (1990). In contrast to the hard clustering, the soft clustering method gives assignments in probability to each cluster for all samples. The quasi-linear form was thus used as the energy function in the clustering method. We can extend the energy function (6.1.7) by the generalized average form (2.2.1). Then we define a generalized energy function by

$$G_\tau = \sum_{i=1}^{N} \phi^{-1} \left(\sum_{k=1}^{K} \pi_k \phi(||x_i - \mu_k||^2) \right). \tag{6.1.8}$$

Different choice of the function ϕ seem to yield the different property of the resultant clustering. It is one of the future works to evaluate the property of the energy function (6.1.8).

6.2 Mixture of experts and neural network model

A mixture of experts model was introduced by Jacobs *and others* (1991). The mixture of experts model is composed by experts and gating functions. Let $\boldsymbol{x}$ be a covariate vector

and y be a response valuable. Then, the mixture of experts model is described as

$$\begin{aligned} P(y|\boldsymbol{x};\boldsymbol{\theta}) &= \sum_{k=1}^{K} P(y,k|\boldsymbol{x},\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} P(k|\boldsymbol{x},\boldsymbol{\theta}^g)P(y|\boldsymbol{x},\boldsymbol{\theta}_k^e), \end{aligned} \tag{6.2.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}^{g\top}, \boldsymbol{\theta}_1^{e\top}, \cdots, \boldsymbol{\theta}_K^{e\top})^\top$ with $\boldsymbol{\theta}^g = (\boldsymbol{\alpha}^{g\top}, \boldsymbol{\beta}^{g\top})^\top$ and $\boldsymbol{\theta}_k^e = (\alpha_k^{e\top}, \boldsymbol{\beta}_k^{e\top})^\top$. In the equation (6.2.1), the first term: $P(k|\boldsymbol{x},\boldsymbol{\theta}^g)$ is called a gating function and the second term: $P(y|\boldsymbol{x},\boldsymbol{\theta}^e)$ is called a experts model. The gate function is defined by

$$P(k|\boldsymbol{x},\boldsymbol{\theta}^g) = \eta_k(\boldsymbol{x};\boldsymbol{\alpha}^g,\boldsymbol{\beta}^g), \tag{6.2.2}$$

where η_k is the softmax function introduced in (2.1.10). There are some variants but the experts in the regression model was originally defined as

$$P(y|\boldsymbol{x},\boldsymbol{\theta}_k^e) = \phi_N(\alpha_k^e + \boldsymbol{\beta}_k^{e\top}\boldsymbol{x}, \boldsymbol{\Sigma}_k), \tag{6.2.3}$$

and the predicted response valuable is given by

$$\hat{y} = \sum_{k=1}^{K} P(k|\boldsymbol{x},\hat{\boldsymbol{\theta}}_k^g)\hat{\boldsymbol{w}}_k^\top\boldsymbol{x}. \tag{6.2.4}$$

In the case of the binary classification model for the binary class label $y = 0, 1$, the experts are defined by

$$P(y|\boldsymbol{x},\boldsymbol{\theta}_k^e) = \tilde{\eta}_0(\boldsymbol{x};\boldsymbol{\alpha}_k^e,\boldsymbol{\beta}_k^e)^{(1-y)}\tilde{\eta}_1(\boldsymbol{x};\boldsymbol{\alpha}_k^e,\boldsymbol{\beta}_k^e)^y, \tag{6.2.5}$$

where $\tilde{\eta}_y(\boldsymbol{x};\boldsymbol{\alpha}_k^e,\boldsymbol{\beta}_k^e) = \frac{\exp(\alpha_{ky}^e + \boldsymbol{\beta}_{ky}^{e\top}\boldsymbol{x})}{\sum_{c=0}^{1}\exp(\alpha_{kc}^e + \boldsymbol{\beta}_{kc}^{e\top}\boldsymbol{x})}$. The class label y is predicted as

$$\hat{y} = \underset{y}{\operatorname{argmax}} \sum_{k=1}^{K} P(k|\boldsymbol{x},\hat{\boldsymbol{\theta}}_k^g)\tilde{\eta}_y(\boldsymbol{x};\hat{\boldsymbol{\alpha}}_k^e,\hat{\boldsymbol{\beta}}_k^e). \tag{6.2.6}$$

The parameters of mixture of experts model are estimated by EM algorithm (Dempster *and others*, 1977). Focusing on the mixture of experts model is regarded as a mixture model, consider that the each response value (y_i) is generated from one distribution and it is indicated by hidden variables

$$z_{ik} = \begin{cases} 1 & \text{if } y_i \text{ is from k-th expert} \\ 0 & \text{else.} \end{cases} \tag{6.2.7}$$

Then the complete log likelihood of the data $\{(\boldsymbol{x}_i, y_i); i = 1, \cdots, N\}$ is written as

$$l(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left(\log P(k|\boldsymbol{x}_i, \boldsymbol{\theta}^g) + \log P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k^e) \right). \tag{6.2.8}$$

The object function to optimize for each step in EM algorithm is given as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathrm{E}[z_{ik}|\boldsymbol{x}_i, y_i, \boldsymbol{\theta}^{(t-1)}] \left(\log P(k|\boldsymbol{x}_i, \boldsymbol{\theta}^g) + \log P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k^e) \right). \tag{6.2.9}$$

Here, the conditional expectation of the hidden indicator is calculated as

$$\begin{aligned} \mathrm{E}[z_{ik}|\boldsymbol{x}_i, y_i, \boldsymbol{\theta}^{(t-1)}] &= P(z_{ik} = 1|\boldsymbol{x}_i, y_i, \boldsymbol{\theta}^{(t-1)}) \\ &= \frac{P(y_i|z_{ik} = 1, \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}) P(z_{ik} = 1|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})}{P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})} \\ &= \frac{P(y_i|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k^{c(t-1)}) P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{g(t-1)})}{\sum_{k=1}^{K} P(y_i|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k^{c(t-1)}) P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{g(t-1)})}. \end{aligned} \tag{6.2.10}$$

From an initial value, computing $\mathrm{E}[z_{ik}|\boldsymbol{x}_i, y_i, \boldsymbol{\theta}^{(t-1)}]$ and updating $\boldsymbol{\theta}$ is alternately repeated until convergence.

The mixture of experts model and the hierarchical modified version (hierarchical mixture of experts model, Jordan and Jacobs (1993)) have been used in numerous regression and classification applications in the healthcare and the other areas (Yuksel *and others*, 2012). These models are parts of the neural network models. The neural network model is based on the divide and conquer principle. In the model, the problem space is divided between

some neural networks, and supervised by a gating network (Masoudnia and Ebrahimpour, 2014). The aim is to separate complex pattern into some local pattern easier to solve for the specific learner or neuron, and achieve effective learning.

As discussed in Chapter 5, the mixture of experts model was used as mixtures of proportional hazard regression models (Rosen and Tanner, 1999). Such a mixture form is similar but differ to the quasi-linear models. Although the mixture of experts model is basically the mixture model for the outcome, the quasi-linear model is the mixture model for the regression or prediction function. On the extension of the relative risk model, the quasi-linear relative risk model is equivalent form to the mixture model only when the tuning parameter $\tau = 1$.

An neural networks model, or artificial neural networks is widely used in many science areas (Amato *and others*, 2013). We give a brief introduction of the feedforward neural network. This is often called a multi-layer perceptron because it has a original source in a single layer perceptron proposed by Rosenblatt (1958), which is a vanilla type of neural networks (Hastie *and others*, 2001). The network consists of several units. Each unit receives multiple inputs and computes one output. The input value y is linear combination of the outputs from all previous units. The input value y is characterized by the linear predictor as

$$y = F(\boldsymbol{X}, \alpha, \boldsymbol{\beta}). \tag{6.2.11}$$

The intercept parameter α is often called a bias in the neural network model. Then, the unit outputs activated value of input as

$$z = f(y), \tag{6.2.12}$$

where f is called an activation function. A number of activation functions have been proposed up to date. We give examples of the activation functions as follows.

- identity function (Rosenblatt, 1958)

$$f(u) = u \tag{6.2.13}$$

- Logistic function

$$f(u) = \frac{1}{1 + \exp(-u)} \tag{6.2.14}$$

- hyperbolic tangent function (LeCun *and others*, 1998)

$$f(u) = \tanh(u) \tag{6.2.15}$$

- rectified linear function (Glorot *and others*, 2011)

$$f(u) = \max(u, 0) \tag{6.2.16}$$

- approximated logistic function

$$f(u) = \begin{cases} -1 & u \leq -1 \\ u & -1 \leq u < 1 \\ 1 & 1 \leq u \end{cases} \tag{6.2.17}$$

- softplus (Glorot *and others*, 2011)

$$f(u) = \log(1 + \exp(u)) \tag{6.2.18}$$

The multilayer perceptron connects the two or more layers each of which is composed of several units. If we focus on 2 layers with I and J units, the output is calculated as

$$y_j = \sum_{i=1}^{I} \beta_{ji} x_i + \alpha_j \tag{6.2.19}$$

$$z_j = f(y_j) \tag{6.2.20}$$

for $j = 1, 2, \cdots, J$. Such relationships are handed down to the last layer. We assume that the final output is a scalar and write it $y(\boldsymbol{x}_i; \alpha, \boldsymbol{B})$, where $\boldsymbol{B}$ is a vector of all coefficients. The training of the neural network model is performed by minimizing some error function. For example, let $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$ be the training data, where $\boldsymbol{x}_i$ is a covariates

vector and y_i is a true binary label from i-th sample. In this case, the logistic activation function (6.2) is often used to model $P(y = 1|\boldsymbol{x})$. Then the error function is similar to the negative log likelihood of the linear logistic model. It is defined as

$$E(\alpha, \boldsymbol{B}) = -\sum_{i=1}^{N} \left(y_i \log(y(\boldsymbol{x}_i; \alpha, \boldsymbol{B})) + (1 - y_i) \log(1 - y(\boldsymbol{x}_i; \alpha, \boldsymbol{B}))\right). \tag{6.2.21}$$

In general, when there are many layers of neurons, it is very difficult to calculate the derivative of the error function by the parameters of the layer closer to the input side. This problem can be avoided by the back propagation algorithm proposed by Rumelhart *and others* (1986).

The neural network model is similar to the quasi-linear model with respect to combining some linear predictors. However, the neural network model is unclear in relation to the linear model as with the quasi-linear model, and it is difficult to interpret the parameters. The restricted quasi-linear model and the quasi-linear model with cross L_1 penalizations achieves the parsimonious expression and it results in an easy interpretation for the estimated model. Conversely, the proposed method in the thesis can be combined with the neural network model. First, the activate function (6.2.15) and (6.2.18) are connected by the quasi-linear form as $f(u) = \log(\exp(\tau \cdot 0) + \exp(\tau u))/\tau$. Then, the tuning of parameter τ may yield better activation function according to the training data. Second, the cross L_1 penalty with the neural network model would result in the parsimonious network model. It may improve the generalization ability of the complicated network and result in better predictive performance.

Bibliography

Aalen, O, Borgan, Ornulf and Gjessing, H. (2008). *Survival and event history analysis*. New York City: Springer.

Amato, F, Lopez, A, Pena-Mendez, E M, Vanhara, P, Hampl, A and Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine* **11**, 47–58.

Anderson, P K and Gill, R D. (1982). Cox's regression model for counting process: A large sample study. *Annals of Statistics* **10**, 1100–1120.

Barlow, W E and Prentice, R L. (1988). Residuals for relative risk regression. *Biometrika* **75**, 65–74.

Belisle, F, Bengio, Y, Dugas, C, Garcia, R and Nadeau, C. (2002). Incorporating second-order functional knowledge for better option pricing. *CIRANO Working Papers*, CIRANO.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. *Journal of the American Statistical Association* **48**, 565–599.

Berkson, J. (1955). Maximum likelihood and minimum χ^2 estimates of the logistic function. *Journal of the American Statistical Association* **50**, 130–162.

Boyd, S and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Brimacombe, M. (2014). High-dimensional data and linear models: a review. *Open Access Medical Statistics* **4**, 17–27.

Buyse, M, Loi, S, van't Veer, L, Viale, G, Delorenzi, M, Glas, AM and *et al.* (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute* **98**, 1183–1192.

Casanova, R, Whitlow, C T, Wagner, B, Williamson, J, Shumaker, S A, Maldjian, J A. and Espeland, M A. (2011). High dimensional classification of structural mri alzheimer's disease data based on large scale regularization. *Fronters in Neuroinformatics* **5**, 22.

Cox, D R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B* **20**, 215–242.

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

Dempster, A P, Laird, N M and Rubin, D B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B* **39**, 1–38.

Dettling, M and Bühlman. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19(9)**, 1061–1069.

Di Camillo, B, Sanavia, T, Martini, M, Jurman, G, Sambo, F, Barla, A, Squillario, M, Furlanello, C, Toffolo, G and Cobelli, C. (2012). Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment. *PLOS ONE* **7**, 1–8.

Dunn, J C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics* **3**, 32–57.

Eguchi, S and Komori, O. (2015). Path connectedness on a space of probability density functions. *Lecture Notes in Computer Science* **9389**, 615–624.

Ein-Dor, L, Kela, I, Getz, G, Givol, D and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171–178.

Elmahdy, E E and Aboutahoun, A W. (2013). A new approach for parameter estimation of finite weibull mixture distributions for reliability modeling. *Applied Mathematical Modelling* **37**, 1800–1810.

Fahad, A, Alshatri, N, Tari, Z, Alamri, A, Khalil, I, Zomaya, A Y, Foufou, S and Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing* **2**, 267–279.

Foster, K. R, koprowski, R and Skufca, J. D. (2014). Machine learning, medical diagnosis, and biomedical engineering research - commentary. *Biomedical Engineering Online* **13**, 94.

Friedman, J, Hastie, T and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Glorot, X, Bordes, A and Bengio, Y. (2011). Deep sparse rectifier neural networks. In: Gordon, Geoffrey J. and Dunson, David B. (editors), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, Volume 15. Journal of Machine Learning Research - Workshop and Conference Proceedings. pp. 315–323.

Goeman, J J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* **52**, 70–84.

Hastie, T, Tibshirani, R and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Heagerty, P J, Lumley, T and Pepe, M S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics* **56**(2), 337–344.

Hinton, G E, Sejnowski, T J and H, Ackley D. (1984). Boltzmann machines: constraint satisfaction networks that learn. *Technical Report* CMS-CS-84-119, CMU Computer Science Department.

Hosmer, D.W. and Lemeshow, S. (1989). *Applied logistic regression*, Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Jacobs, R A., Jordan, M I, Nowlan, S J and Hinton, G E. (1991). Adaptive mixtures of local experts. *Neural Computation* **3**, 79–87.

Jordan, M I and Jacobs, R A. (1993). Hierarchical mixtures of ex- perts and the em algorithm. *Technical Report AIM-1440*.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. John Wiley & Sons.

Kalousis, A, Prados, J and Hilario, M. (2005). Stability of feature selection algorithms. *In Proc. 5th IEEE International Con- ference on Data Mining (ICDM '05)*, 218–225.

Kanungo, T, Mount, D M, Netanyahu, N S, Piatko, C D, Silverman, R and Wu, A. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 881–892.

Klein, J P and Moeschberger, M L. (2003). *Survival Analysis Techniques for Censored and Truncated Data*, Second edition. Springer.

Kolmogorov, A. (1930). On the notion of mean. *In Tikhomirov (ed, 1991), Selected works of A.N. Kolmogorov, Vol. I: Mathematics and Mechanics, Kluwer Academic Publishers*, 144–146.

Komori, O, Pritchard, M and Eguchi, S. (2013). Multiple suboptimal solutions for prediction rules in gene expression data. *Computational and mathematical methods in medicine* **2013**, 798189.

Kravitz, R L, Duan, N and Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank quarterly* **82**, 661–687.

LeCun, Y, Bottou, L, Orr, G B and Muller, K R. (1998). Effiicient backprop. In: *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*. London, UK, UK: Springer-Verlag. pp. 9–50.

Louzada-Neto, F, Mazucheli, J and Achcar, J A. (2002). Mixture hazard models for lifetime data. *Biometrical Journal* **44**, 3–14.

Masoudnia, S and Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review* **42**, 275–293.

McCullagh, P and Nelder, J A. (1989). *Generalized linear models (Second edition)*. London: Chapman & Hall.

McCulloch, C E, Searle, S R and Neuhaus, J M. (2008). *Generalized, linear, and mixed models, 2nd edition*. New Jersey: Wiley.

McQueen, J. (1967). Some methods for classification and analysis of multivariate observartions. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 281–297.

Murphy, K P. (2012). *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. Cambridge: The MIT Press.

Nagumo, M. (1930). *ü*ber eine klasse der mittelwerte. *Japanese journal of mathematics :transactions and abstracts* **7**, 71–79.

Nelder, J A and Wedderburn, R W M. (1972). Generalized linear models. *Journal of the Royal Statistical Society* **125**, 370–384.

Oghabian, A, Kilpinen, S, hautaniemi, S and Czeizler, E. (2014). Biclustering methods: Biological relevance and application in gene expression analysis. *PLoS ONE* **9**, e90801.

Omae, K, Komori, O and Eguchi, S. (2016). Reproducible detection of disease-associated markers from gene expression data. *BMC Medical Genomics* **9:53**.

Omae, K, Komori, O and Eguchi, S. (2017). Quasi-linear score for capturing heterogeneous structure in biomarkers. *BMC Bioinformatics* **18:308**.

Park, M Y and Hastie, T. (2007). l_1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society* **69**, 659–677.

Rose, K, Gurewitz, E and Fox, G C. (1990). Statistical mechanics and phase transitions in clustering. *Physical Review Letters* **65**, 945–948.

Rosen, O and Tanner, M. (1999). Mixtures of proportional hazards regression models. *Statistics in Medicine* **18**(9), 1119–1131.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65–386.

Rumelhart, D E, Hinton, G E and Williams, R J. (1986). Learning representations by back-propagating errors. *Nature* **323**, 533–536.

Setlur, S, Mertz, K, Hoshida, Y, Demichelis, F and *et al.*, Lupien M. (2008). Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *Journal of the National Cancer Institute* **100**, 815–825.

Sørlie, T, Perou, C M, Tibshirani, R, Aas, T, Geisler, S and *et al.*, Johnsen. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10869–10874.

Suzuki, K, Zhou, L and Wang, Q. (2017). Machine learning in medical imaging. *Pattern Recognition* **63**, 465–467.

Thompson, B R and Baker, R J. (1981). Composite link functions in generalized linear models. *Journal of the Royal Statistical Society* **30**, 125–131.

Tsiatis, A A. (1981). A large sample study of cox's regression model. *The annals of statistics* **9**, 93–108.

van't Veer, L J, Dai, H, van de Vijver, M J, He, Y D, Hart, A A M, Mao, M and *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

Vaupel, J W and Yashin, A I. (1985). The deviant dynamics of death in heterogeneous populations. *Sociological Methodology* **15**, 179–211.

Wallstrom, G, Anderson, K S and LaBaer, J. (2013). Biomarker discovery for heterogeneous diseases. *Cancer Epidemiol Biomarkers Prev* **22**, 747–755.

Wang, Y, Kijin, J G, Zhang, Y, Sieuwerts, A M, Look, M P, Yang, F and *et al.* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679.

Webb, A R and Copsey, K D. (2011). *Statistical Pattern Recognition, Third Edition*. Chichester: John Wiley & Sons.

Yan, L, Tian, L and Liu, S. (2015). Combining large number of weak biomarkers based on auc. *Statistics in Medicine* **34**, 3811–3830.

Youden, W J. (1950). Index for rating diagnostic tests. *Cancer* **3**, 32–35.

Yuksel, S E, Wilson, J N and Gader, P D. (2012). Twenty years of mixture of experts. *IEEE Trans. Neural Netw. Learning Syst* **23**, 1177–1193.

Zhang, Q, Hua, C and Xu, G. (2014). A mixture weibull proportional hazard model for mechanical system failure prediction utilising lifetime and monitoring data. *Mechanical Systems and Signal Processing* **43**, 103–112.

Zou, H and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* **67**, 301–320.

Appendix

A.1 *proof of Proposition 1*

(1) It is easy to see the following inequality:

$$K \min_{k=1,\cdots,K}\{\exp(z_k)\} \leq \sum_{k=1}^{K} \exp(z_k) \leq K \max_{k=1,\cdots,K}\{\exp(z_k)\}$$

for $z_k \in \mathbb{R}(k = 1, 2, \cdots, K)$. Divide by K and take the logarithms and divide by τ, then we get that

$$\frac{1}{\tau} \min_{k=1,\cdots,K}\{z_k\} \leq \frac{1}{\tau} \log\left(\frac{1}{K}\sum_{k=1}^{K} \exp(z_k)\right) \leq \frac{1}{\tau} \max_{k=1,\cdots,K}\{z_k\} \tag{A.1}$$

Let $z_k = \tau(\alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x})$. Then (1) of Proposition 1 follows.

(2) Let $\alpha_m + \boldsymbol{\beta}_m^\top \boldsymbol{x} = \min_{k=1,\cdots,K}\{z_k\}$. Then

$$\begin{aligned} F_\tau(\boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}) - (\alpha_m + \boldsymbol{\beta}_m^\top \boldsymbol{x}) &= \frac{1}{\tau} \log\left(\frac{1}{K}\sum_{k=1}^{K} \exp(\tau\alpha_k + \tau\boldsymbol{\beta}_k^\top \boldsymbol{x})\right) \\ &- \frac{1}{\tau} \log\left(\exp(\tau\alpha_m + \tau\boldsymbol{\beta}_m^\top \boldsymbol{x})\right) \\ &= \frac{1}{\tau} \log\left(\frac{1}{K}\sum_{k=1}^{K} \exp\left((\tau\alpha_k + \tau\boldsymbol{\beta}_k^\top \boldsymbol{x}) - \left(\tau\alpha_m + \tau\boldsymbol{\beta}_m^\top \boldsymbol{x}\right)\right)\right) \\ &= \frac{1}{\tau} \log\left(\frac{1}{K}\sum_{k=1}^{K} \exp\left(\tau(\gamma_k)\right)\right), \end{aligned}$$

where $\gamma_k = (\alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}) - \left(\alpha_m + \boldsymbol{\beta}_m^\top \boldsymbol{x}\right)$. Here, since $\gamma_k \geq 0$ for all $k = 1, 2, \cdots, K$,

$$\lim_{\tau \to \infty} F_\tau(\boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}) - (\alpha_m + \boldsymbol{\beta}_m^\top \boldsymbol{x}) = 0 \tag{A.2}$$

(3) When τ goes to 0, we see by l'Hôpital's rule that

$$\begin{aligned}
\lim_{\tau \to 0} F_\tau(\boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \lim_{\tau \to 0} \frac{\log\left(\frac{1}{K}\sum_{k=1}^K \exp(\tau\alpha_k + \tau\boldsymbol{\beta}_k^\top \boldsymbol{x})\right)}{\tau} \\
&= \lim_{\tau \to 0} \frac{\frac{1}{K}\sum_{k=1}^K (\alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}) \exp(\tau\alpha_k + \tau\boldsymbol{\beta}_k^\top \boldsymbol{x})}{\frac{1}{K}\sum_{k=1}^K \exp(\tau\alpha_k + \tau\boldsymbol{\beta}_k^\top \boldsymbol{x})} \\
&= \frac{1}{K}\sum_{k=1}^K \alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x} \\
&= \frac{1}{K}\sum_{k=1}^K F(\boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}).
\end{aligned} \tag{A.3}$$

Therefore, (3) of Proposition 1 follows.

(4) It is easy to see the following inequality:

$$\max_{k=1,\cdots,K}\{\exp(z_k)\} \leq \sum_{k=1}^K \exp(z_k) \leq K \max_{k=1,\cdots,K}\{\exp(z_k)\}$$

for $z_k \in \mathbb{R} (k = 1, 2, \cdots, K)$. Divide by K and take the logarithms and divide by τ for each side of the inequality, then we get that

$$\frac{1}{\tau}\max_{k=1,\cdots,K}\{z_k\} - \frac{\log K}{\tau} \leq \frac{1}{\tau}\log\left(\frac{1}{K}\sum_{k=1}^K \exp(z_k)\right) \leq \frac{1}{\tau}\max_{k=1,\cdots,K}\{z_k\}. \tag{A.4}$$

Let $z_k = \tau(\alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x})$ in the inequality (A.4). Then, F_τ converges to F_∞ by a sandwich theorem when τ goes to infinity.

A.2 *proof of Theorem 2*

Let $\boldsymbol{\gamma}_k = \Sigma^{-1}\boldsymbol{\mu}_k$ for $k = 1, 2, \cdots, K$. By the assumption we get that $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_K$ are linearly independent. Then, there exists the non-singular matrix $A \in \mathbb{R}^{p \times p}$ such that $A\boldsymbol{\gamma}_k = (0, \ldots, 0, \boldsymbol{\beta}_k^\top, 0, \cdots, 0)^\top$ for any $k \in \{1, 2, \cdots, K\}$, where $\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}$. For example,

we set

$$A = B(R^\top R)^{-1}R^\top + C(I_p - R(R^\top R)^{-1}R^\top), \tag{A.5}$$

where

$$B = \begin{pmatrix} \boldsymbol{\beta}_1 & 0_{p_1} & \cdots & 0_{p_1} \\ 0_{p_2} & \boldsymbol{\beta}_2 & \cdots & 0_{p_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{p_K} & \cdots & \cdots & \boldsymbol{\beta}_K \end{pmatrix}, \tag{A.6}$$

$R = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_K)$ and C is any square matrix of size p. Then $AR = B$, or equivalently $A\boldsymbol{\gamma}_k = (0, \cdots, 0, \boldsymbol{\beta}_k^\top, 0, \cdots, 0)^\top$ for any $k \in \{1, 2, \cdots, K\}$. Let C be fixed so as to satisfy that A is non-singular. Then,

$$\begin{aligned} \boldsymbol{\gamma}_k^\top \boldsymbol{Z} &= (A\boldsymbol{\gamma}_k)^\top (A^\top)^{-1}\boldsymbol{Z} \\ &= (0, \cdots, 0, \boldsymbol{\beta}_k^\top, 0, \cdots, 0)(\boldsymbol{X}_{(1)}, \cdots, \boldsymbol{X}_{(K)})^\top \\ &= \boldsymbol{\beta}_k^\top \boldsymbol{X}_{(k)}. \end{aligned}$$

By the Theorem 1 and the existence of such a transformation A confirms that the true log likelihood forms $F(\boldsymbol{X}) = \log\{\sum_{k=1}^K \exp(\alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{X}_{(k)})\}$.

A.3 *Counting Process and Martingale theory in Cox's proportional hazard model*

This part is written referencing Kalbfleisch and Prentice (2002); Aalen *and others* (2008). We prepare the definitions and notations as follows.

1. Counting process $\{N(t), t \geq 0\}$ is defined as a nonnegative discrete, nondecreasing and right-continuous with left-hand limit, say *cadlag* , stochastic process with $N(0) = 0$.

2. At-risk process $\{Y(t), t \geq 0\}$ is defined as a left continuous process which takes 1 or 0.

3. Covariate process $\{\boldsymbol{X}(t), t \geq 0\}$ is defined as a left continuous process, where $\boldsymbol{X}(t)$ is a covariates vector.

4. A history or *filtration* of n individuals between a interval $[0, t)$ is defined as

$$\mathcal{F}_t = \sigma\{N_i(u), Y_i(u^+), \boldsymbol{X}_i(u^+),\ i = 1, \cdots, n, 0 \leq u < t\}, \tag{A.7}$$

where σ denotes a sigma field of all events.

5. A stochastic process $U = \{U(t), t \geq 0\}$ is said to be *adapted* to the filtration $\mathcal{F}_t$ if for each t, $U(t)$ is a function of $\mathcal{F}_t$, that is, $\mathcal{F}_t$-measurable.

6. A stochastic process $U = \{U(t), t \geq 0\}$ is said to be *predictable* with respect to $\mathcal{F}_t$ if for each t, the value of $U(t)$ is a function of $\mathcal{F}_{t-}$.

7. A stochastic process $\{M(t), 0 \leq t \leq \tau\}$ is a martingale with respect to the filtration $\mathcal{F}_t$ if it is cadlag, adapted to $\mathcal{F}_t$ and satisfies the martingale property

$$\mathrm{E}[M(t)|\mathcal{F}_s] = M(s) \quad \text{for all} \quad s \leq t \leq \tau. \tag{A.8}$$

Especially, it is called as zero-mean martingale if $\mathrm{E}[M(0)] = 0$.

8. A stochastic process $\{\bar{M}(t), 0 \leq t \leq \tau\}$ is a submartingale with respect to the filtration $\mathcal{F}_t$ if it is cadlag, adapted to $\mathcal{F}_t$ and satisfies the submartingale property

$$\mathrm{E}[\bar{M}(t)|\mathcal{F}_s] \geq \bar{M}(s) \quad \text{for all} \quad s \leq t \leq \tau. \tag{A.9}$$

9. A martingale is said to be *square-integrable* if $\mathrm{E}[M^2(t)] < \infty$ for all $t \leq \tau$.

10. A *predictavle variation process* of a square-integrable martingale M is defined as

$$\langle M \rangle(t) = \int_0^t \mathrm{var}[dM(u)|\mathcal{F}_{u-}]. \tag{A.10}$$

Now we are consciously about the stochastic process so that let $\boldsymbol{X}_i$ be a vector of covariates for the i-th individual, and admit (1) $\boldsymbol{X}_i$'s depend on time; $\boldsymbol{X}_i = \boldsymbol{X}_i(t)$, (2) to replace the covariates to the p-dimensional predictive function of covariates and time; $\boldsymbol{Z}_i = \boldsymbol{Z}_i(\boldsymbol{X}_i, t)$. We assume that the process $\boldsymbol{Z}_i(t)$ is at least left continuous. For simplicity,

consider the only case of no event ties and independent censoring. Cox proportional hazard model assume that

$$\lambda_i(t) = \lambda_0(t)\exp(\boldsymbol{Z}_i^\top \boldsymbol{\beta}). \tag{A.11}$$

For i-th individual, let $N_i(t) = \mathbf{1}(t_i \le t, c_i \le t)$ be the right continuous counting process, where each $N_i(t)$ counts the number of observed event on $(0, t]$, and let $Y_i(t) = \mathbf{1}(t_i \ge t, c_i \ge t)$ be the left continuous at-risk process which shows the observation status at time t, where c_i and t_i are censoring and true survival time. Define the filtration as $\mathcal{F}_t = \{N_i(u), Y_i(u^+), \boldsymbol{X}_i(u^+); i = 1, \cdots, n; 0 \le u \le t\}$. Then the corresponding intensity process of $N_i(t)$ is defined by

$$\Lambda_i(t)dt = P(dN_i(t) = 1|\mathcal{F}_{t\_}). \tag{A.12}$$

Cox proportional hazard model is described by such counting process formulations as

$$P(dN_i(t) = 1|\mathcal{F}_{t\_}) = Y_i(t)\lambda_0(t)\exp\{\boldsymbol{Z}_i(t)^\top \boldsymbol{\beta}\}dt, \tag{A.13}$$

where $dN_i(t) = N_i(t^- + dt) - N_i(t^-)$. The partial log-likelihood function, the score function $U(\boldsymbol{\beta})$ and the Fisher information matrix $I(\boldsymbol{\beta})$ are written as

$$\begin{aligned} l^p &= \sum_{i=1}^n \int_0^\infty \{\boldsymbol{Z}_i(u)^\top \boldsymbol{\beta}\} dN_i(u) - \int_0^\infty \log\left\{\sum_{j=1}^n Y_j(u)\exp\{\boldsymbol{Z}_j(u)^\top \boldsymbol{\beta}\}\right\} dN_\cdot(u) \\ &= \sum_{i=1}^n \int_0^\infty \left\{\boldsymbol{Z}_i(u)^\top \boldsymbol{\beta} - \log\left(\sum_{j=1}^n Y_j(u)\exp\left(\boldsymbol{Z}_j(u)^\top \boldsymbol{\beta}\right)\right)\right\} dN_i(u), \qquad \text{(A.14)} \\ U(\boldsymbol{\beta}) &= \sum_{i=1}^n \int_0^\infty \{\boldsymbol{Z}_i(u) - \mathcal{E}(\boldsymbol{\beta}, u)\} dN_i(u), \qquad \text{(A.15)} \\ I(\boldsymbol{\beta}) &= \int_0^\infty \sum_{i=1}^n \{\boldsymbol{Z}_i(u) - \mathcal{E}(\boldsymbol{\beta}, u)\}\{\boldsymbol{Z}_i(u) - \mathcal{E}(\boldsymbol{\beta}, u)\}^\top p_i(\boldsymbol{\beta}, u) dN_i(u) \qquad \text{(A.16)} \end{aligned}$$

where $N_\cdot = \sum_{i=1}^n N_i(u)$ and $\mathcal{E}(\boldsymbol{\beta}, u) = \sum_{\ell=1}^n \boldsymbol{Z}_\ell(u) p_\ell(\boldsymbol{\beta}, u)$ with

$$p_\ell(\boldsymbol{\beta}, u) = \frac{Y_\ell(u) \exp\{\boldsymbol{Z}_\ell(u)^\top \boldsymbol{\beta}\}}{\sum_{j=1}^n Y_j(u) \exp\{\boldsymbol{Z}_j(u)^\top \boldsymbol{\beta}\}}. \tag{A.17}$$

Consider the score function based on data available up to time t,

$$U(\boldsymbol{\beta}, t) \quad = \quad \sum_{i=1}^n \int_0^t \{\boldsymbol{Z}_i(u) - \mathcal{E}(\boldsymbol{\beta}, u)\} dN_i(u), \tag{A.18}$$

Doob-Meyer decomposition theorem mentions that any sub-martingale $\bar{M}(t)$ is uniquely decomposed into a mean zero martingale $M(t)$ and an increasing cadlag predictable process $C(t)$, called *compensator*, which satisfies $C(0) = 0$ and $dC(t) = \mathrm{E}[d\bar{M}(t)|\mathcal{F}_{t-}]$ as $\bar{M}(t) = M(t) + C(t)$. Thus $N_i(t)$ is decomposed into compensator $A_i(t)$ and $M_i(t)$, which is a zero-mean martingale with respect to $\mathcal{F}_t$. Thus we derive a process $\{M_i; i = 1, \cdots, n\}$ as

$$M_i(t) = N_i(t) - A_i(t), \tag{A.19}$$

where $A_i(t) = \int_0^t P(dN_i(t) = 1|\mathcal{F}_{t-})$. Then,

$$\begin{aligned} U(\beta, t) \quad &= \quad \sum_{i=1}^n \int_0^t \{\boldsymbol{Z}_i(u) - \mathcal{E}(\boldsymbol{\beta}, u)\} dM_i(u) + \sum_{i=1}^n \int_0^t \{\boldsymbol{Z}_i(u) - \mathcal{E}(\boldsymbol{\beta}, u)\} dA_i(u) \quad \text{(A.20)} \\ &= \quad \sum_{i=1}^n \int_0^t \{\boldsymbol{Z}_i(u) - \mathcal{E}(\boldsymbol{\beta}, u)\} dM_i(u) \quad \text{(A.21)} \end{aligned}$$

since the second term of $(A.20)$ is equal to 0. Let $\boldsymbol{U}(\boldsymbol{\beta}, t) = [U_1(\boldsymbol{\beta}, t), \cdots, U_p(\boldsymbol{\beta}, t)]$ and define the vector of predictable variation process as

$$\langle U \rangle(t) = \int_0^t \mathrm{E}[d\boldsymbol{U}(t) d\boldsymbol{U}(t)^\top | \mathcal{F}_{t-}] \tag{A.22}$$

Focus on the form as $(A.21)$: $U(t) = \sum_{i=1}^n \int_0^t G_i(u) dM_i(u)$. When $M_1, \cdots, M_n$ are orthogonal (i.e. $\langle M_i, M_j \rangle(t) = \int_0^t \mathrm{Cov}[dM_i(u), dM_j(u)|F_{u-}]$=0 for all $i \neq j$) mean zero martingales and $G_i(u)$ is a p-dimensional vector of predictable processes with respect to the same filtration $\mathcal{F}_t$, it is known that the predictable variation process is written as

$$\langle U\rangle(t) = \sum_{i=1}^{n} \int_0^t G_i(u) G_i(u)^\top d\langle M_i\rangle(u). \tag{A.23}$$

Since we assumed that there is no ties of events, for $i \neq j$

$$\begin{aligned} \mathrm{Cov}[dM_i(u), dM_j(u)|F_{u-}] &= \mathrm{Cov}[dN_i(u), dN_j(u)|F_{u-}] \\ &= 0 \end{aligned}$$

and since

$$\begin{aligned} \mathrm{Var}[dM(t)|F_{u-}] &= \mathrm{Var}[dN(t)|F_{u-}] \\ &= \Pr[dN(t) = 1|F_{u-}], \end{aligned}$$

so that

$$\begin{aligned} \langle U(\boldsymbol{\beta})\rangle(t) &= \sum_{i=1}^{n} \int_0^t G_i(u) G_i(u)^\top d\langle M_i\rangle(u) \\ &= \sum_{i=1}^{n} \int_0^t G_i(u) G_i(u)^\top \mathrm{var}[dM_i(u)|\mathcal{F}_{u-}] \\ &= \sum_{i=1}^{n} \int_0^t G_i(u) G_i(u)^\top Y_i(u) \lambda_0(t) \exp\{\boldsymbol{Z}_i(u)^\top \boldsymbol{\beta}\} du, \end{aligned} \tag{A.24}$$

where $G_i(u) = \boldsymbol{Z}_i(u) - \mathcal{E}(\boldsymbol{\beta}, u)$. Let

$$S^{(0)}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_i(t) \exp\{\boldsymbol{Z}_i(t)^\top \boldsymbol{\beta}\} \tag{A.25}$$

$$\begin{aligned} S^{(1)}(\boldsymbol{\beta}, t) &= \frac{\partial}{\partial \boldsymbol{\beta}} S^{(0)}(\boldsymbol{\beta}, t) \\ &= \sum_{i=1}^{n} Y_i(t) Z_i(t) \exp\{\boldsymbol{Z}_i(t)^\top \boldsymbol{\beta}\} \end{aligned} \tag{A.26}$$

$$\begin{aligned} S^{(2)}(\boldsymbol{\beta}, t) &= \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} S^{(0)}(\boldsymbol{\beta}, t) \\ &= \sum_{i=1}^{n} Y_i(t) \boldsymbol{Z}_i(t) \boldsymbol{Z}_i(t)^\top \exp\{\boldsymbol{Z}_i(t)^\top \boldsymbol{\beta}\} \end{aligned} \tag{A.27}$$

and $||A||$ be the maximum component of vector or matrix A. We denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_0$. Under the existence of an open neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ and functions $s^{(j)}(\boldsymbol{\beta}, t), j = 0, 1, 2$ defined on $\mathcal{B} \times [0, \tau]$ which satisfy the following conditions, partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$.

1. $\sup_{\boldsymbol{\beta} \in \mathcal{B}, t \in [0,\tau]} ||n^{-1} S^{(j)}(\boldsymbol{\beta}, t) - s^{(j)}(\boldsymbol{\beta}, t)|| \xrightarrow{p} 0$ as $n \to \infty$
2. $s^{(0)}(\boldsymbol{\beta}, t)$ is bounded away from 0 for $t \in [0, \tau]$
3. For $j = 0, 1, 2$, $s^{(j)}(\boldsymbol{\beta}, t)$ is a continuous function of $\boldsymbol{\beta}$ uniformly in $t \in [0, \tau]$, where $s^{(1)}$ and $s^{(2)}$ are the first and second derivative of $s^{(0)}$
4. $\Sigma(\boldsymbol{\beta}, \tau) = \int_0^\tau v(\boldsymbol{\beta}, u) s^{(0)}(\boldsymbol{\beta}, u) \lambda_0(u) du$ is positive definite for all $\boldsymbol{\beta} \in \mathcal{B}$, where $v(\boldsymbol{\beta}, t) = s^{(2)}/s^{(0)}(\boldsymbol{\beta}, t) - e(\boldsymbol{\beta}, t) e(\boldsymbol{\beta}, t)^\top$ with $e(\boldsymbol{\beta}, t) = s^{(1)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t)$.

proof. We give an outline of the proof. As for details, refer to Anderson and Gill (1982). Consider the process

$$\begin{aligned} X(\boldsymbol{\beta}, t) &= n^{-1}[l(\boldsymbol{\beta}, t) - l(\boldsymbol{\beta}_0, t)] \\ &= n^{-1} \sum_{i=1}^{n} \int_0^t \left\{ Z_i(u)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \log \frac{S^{(0)}(\boldsymbol{\beta}, u)}{S^{(0)}(\boldsymbol{\beta}_0, u)} \right\} dN_i(u). \end{aligned} \tag{A.28}$$

This is a submartingale with compensator

$$\tilde{X}(\boldsymbol{\beta}, t) = n^{-1} \int_0^t \left\{ S^{(1)}(\boldsymbol{\beta}, u)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \log \frac{S^{(0)}(\boldsymbol{\beta}, u)}{S^{(0)}(\boldsymbol{\beta}_0, u)} S^{(0)}(\boldsymbol{\beta}_0, u) \right\} \lambda_0(u) du.$$

It can be shown that $X(\boldsymbol{\beta}, t) - \tilde{X}(\boldsymbol{\beta}, t)$ is a square integrable martingale. $\tilde{X}(\boldsymbol{\beta}, \tau)$ for $\boldsymbol{\beta} \in \mathcal{B}$ converges in probability to

$$f(\boldsymbol{\beta}) = \int_0^\tau \left\{ s^{(1)}(\boldsymbol{\beta}, u)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \log \frac{s^{(0)}(\boldsymbol{\beta}, u)}{s^{(0)}(\boldsymbol{\beta}_0, u)} s^{(0)}(\boldsymbol{\beta}_0, u) \right\} \lambda_0(u) du. \tag{A.29}$$

By Lenglart's inequality, $X(\boldsymbol{\beta}, \tau)$ converges in probability to the same value with the limit of $\tilde{X}(\boldsymbol{\beta}, \tau)$: $f(\boldsymbol{\beta})$. The first derivative of the $f(\boldsymbol{\beta})$ is zero at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and the second derivative of the $f(\boldsymbol{\beta})$ is a negative definite matrix at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Thus $X(\boldsymbol{\beta}, \tau)$ converges in probability

to a concave function of $\boldsymbol{\beta}$ with a unique solution $\boldsymbol{\beta}_0$. From the definition, $\hat{\boldsymbol{\beta}}$ maximize $X(\boldsymbol{\beta}, \tau)$. Then we get that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ from the concavity of the function $f(\boldsymbol{\beta})$.

Acknowledgements

I am grateful to all the friends, colleagues and professors who have supported me during my doctoral course.

I would like to express my sincerest thanks to Professor Shinto Eguchi. He taught me that there are so many perspectives to interpret things, that having a long-term perspective is often better than raising immediate profits, and that research is nothing but fun. I am proud to be his disciple and to have a doctorate under him.

I appreciate Dr. Shuhei Mano for very careful check in the dissertation and his many valuable comments. I am convinced that his advice will have a positive impact not only on improving this paper but on my future research.

I also appreciate for two sub-advisors, Dr. Masayuki Henmi and Dr. Hisashi Noma, and my two seniors of the laboratory, Dr. Osamu Komori and Dr. Akifumi Notsu. They continuously gave me much advice and encouragement.

My colleagues in Shizuoka Cancer Center also supported and encouraged me in every way. I cannot write all the names but give great thanks to all staff in clinical research center including Dr. Keita Mori and Dr. Toshiaki Takahashi, and all collaborators including Dr. Tateaki Naito and Mr. Taro Okayama.

Finally, I would like to express my deepest gratitude to my parents, Kouichirou and Mutsuko Omae for their understanding and sincere support.