A covariate-adjusted classification model for multiple biomarkers in disease screening and diagnosis

by

Suizhi Yu

M.S., George Washington University, 2013

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

The classification methods based on a linear combination of multiple biomarkers have been widely used to improve the accuracy in disease screening and diagnosis. However, it is seldom to include covariates such as gender and age at diagnosis into these classification procedures. It is known that biomarkers or patient outcomes are often associated with some covariates in practice, therefore the inclusion of covariates may further improve the power of prediction as well as the classification accuracy. In this study, we focus on the classification methods for multiple biomarkers adjusting for covariates. First, we proposed a covariate-adjusted classification model for multiple cross-sectional biomarkers. Technically, it is a two-stage method with a parametric or non-parametric approach to combine biomarkers first, and then incorporating covariates with the use of the maximum rank correlation estimators. Specifically, these parameter coefficients associated with covariates can be estimated by maximizing the area under the receiver operating characteristic (ROC) curve. The asymptotic properties of these estimators in the model are also discussed. An intensive simulation study is conducted to evaluate the performance of this proposed method in finite sample sizes. The data of colorectal cancer and pancreatic cancer are used to illustrate the proposed methodology for multiple cross-sectional biomarkers.

We further extend our classification method to longitudinal biomarkers. With the use of a natural cubic spline basis, each subject's longitudinal biomarker profile can be characterized by spline coefficients with a significant reduction in the dimension of data. Specifically, the maximum reduction can be achieved by controlling the number of knots or degrees of freedom in the spline approach, and its coefficients can be obtained by the ordinary least squares method. We consider each spline coefficient as "biomarker" in our previous method, then the optimal linear combination of those spline coefficients can be acquired using Stepwise method without any distributional assumption. Afterward, covariates are

included by maximizing the corresponding AUC as the second stage. The proposed method is applied to the longitudinal data of Alzheimer's disease and the primary biliary cirrhosis data for illustration. We conduct a simulation study to assess the finite-sample performance of the proposed method for longitudinal biomarkers.

A Covariate-Adjusted Classification Model for Multiple Biomarkers in Disease Screening and Diagnosis

by

Suizhi Yu

M.S., George Washington University, 2013

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Dr. Wei-Wen Hsu

Copyright

Abstract

The classification methods based on a linear combination of multiple biomarkers have been widely used to improve the accuracy in disease screening and diagnosis. However, it is seldom to include covariates such as gender and age at diagnosis into these classification procedures. It is known that biomarkers or patient outcomes are often associated with some covariates in practice, therefore the inclusion of covariates may further improve the power of prediction as well as the classification accuracy. In this study, we focus on the classification methods for multiple biomarkers adjusting for covariates. First, we proposed a covariate-adjusted classification model for multiple cross-sectional biomarkers. Technically, it is a two-stage method with a parametric or non-parametric approach to combine biomarkers first, and then incorporating covariates with the use of the maximum rank correlation estimators. Specifically, these parameter coefficients associated with covariates can be estimated by maximizing the area under the receiver operating characteristic (ROC) curve. The asymptotic properties of these estimators in the model are also discussed. An intensive simulation study is conducted to evaluate the performance of this proposed method in finite sample sizes. The data of colorectal cancer and pancreatic cancer are used to illustrate the proposed methodology for multiple cross-sectional biomarkers.

We further extend our classification method to longitudinal biomarkers. With the use of a natural cubic spline basis, each subject's longitudinal biomarker profile can be characterized by spline coefficients with a significant reduction in the dimension of data. Specifically, the maximum reduction can be achieved by controlling the number of knots or degrees of freedom in the spline approach, and its coefficients can be obtained by the ordinary least squares method. We consider each spline coefficient as "biomarker" in our previous method, then the optimal linear combination of those spline coefficients can be acquired using Stepwise method without any distributional assumption. Afterward, covariates are

included by maximizing the corresponding AUC as the second stage. The proposed method is applied to the longitudinal data of Alzheimer's disease and the primary biliary cirrhosis data for illustration. We conduct a simulation study to assess the finite-sample performance of the proposed method for longitudinal biomarkers.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor, Dr. Wei-Wen Hsu, for his invaluable guidance, patience, encouragement, enthusiasm, and support. As an intelligent and knowledgeable statistician and professor, he has shared with me his ideas, experience, and wisdom. Moreover, he has served as an outstanding example of productivity, professionalism, and, most of all, superior work quality.

I am also highly grateful to the committee members: Dr. Gary L. Gadbury, Dr. Juan Du, and Dr. Stefan Bossmann, for their valuable suggestions and helpful comments. A special thank goes to Dr. Gary Gadbury for offering me an opportunity to join the NASA project and guiding me to gain an incredible experience from it.

I owe my deepest gratefulness to my parents, Xiaolong Yu and Ping Yuan, for their unconditional love and generous support throughout my life.

Finally, I also want to thank my friends: Chao Zhang, Yitian Na, Bing Wang, Jiali Zhu and Yiming Li, for helping and encouraging me as always.

Dedication

In dedication to the memory of my grandfather Zhixin Yuan and my grandmother Yi Bai, gentle and diligent people whom I still miss a lot.

Chapter 1

Introduction

In disease diagnosis and screening, a large number of biomarkers have been studied and used for the distinction between disease and healthy people. For instance, prostate-specific antigen (PSA) is a biomarker measured in serum for the screening test of prostate cancer (Catalona et al., 1991). Another example, cerebrospinal fluid (CSF) biomarkers are considered as a potential source of Alzheimer's disease (AD) pathology in many AD studies (Mistur et al., 2009). In practice, using multiple biomarkers can significantly improve the diagnostic and test accuracy and result in higher sensitivity and specificity than using a single biomarker (Pepe and Thompson, 2000; Etzioni et al., 2003).

Sensitivity and specificity are the traditional measures of classification accuracy which are generally reported with the use of the receiver operating characteristic (ROC) curve. The ROC curve is a widely used statistical and graphical tool for continuous biomarkers to assess the performance of disease diagnosis and classification (Swets, 1986; Zweig and Campbell, 1993; Baker, 2003). The ROC curve can be summarized by many indices such as the Youden index and the area under the ROC curve (AUC). Particularly, AUC is the most commonly used measure (Pepe, 2003, Page 77).

Combining multiple biomarkers for classification has been well discussed in the literature. The popular approach is to find a linear combination of biomarkers by maximizing the AUC using parametric and non-parametric approaches. Su and Liu (1993) presented an optimal

linear combination of biomarkers to maximize the corresponding AUC under the assumption of the multivariate normal distribution on biomarkers. To relax the biomarker distributional assumptions, Pepe and Thompson (2000) suggested an empirical approach to search for the linear biomarker combination to achieve the maximum AUC. However, their method only works for combining two biomarkers. Pepe et al. (2006) further considered to find a combination of multiple biomarkers under a generalized linear model. Liu et al. (2011) developed a min-max approach to combine only the smallest and largest values of multiple biomarkers by yielding the largest AUC nonparametrically. Kang et al. (2016) proposed a nonparametric stepwise approach to combine one biomarker each time by maximizing the empirical AUC to include all biomarkers.

However, the above classification procedures rarely discuss and use covariates. Covariates, such as age and gender, are often related to biomarkers or outcomes of patients in practice. For example, the detection of breast cancer by mammography highly depends on the female's age. In the exercise stress study, gender is an important covariate because of the difference in the physical exercise abilities between women and men (Pepe, 2003, Page 48). Janes and Pepe (2008) suggested that covariate adjustment should be included in the classification procedures since covariates may provide the additional information to further improve the classification accuracy.

Some studies have discussed the covariate adjustment on the classification for biomarkers. One of the most popular approaches is to model the ROC curve as a function of covariates. For example, Pepe (2000) suggested to parametrically model the ROC curve with covariates under the framework of the generalized linear models. The similar work can be found in Alonzo and Pepe (2002) and Cai and Pepe (2002). Janes and Pepe (2009) developed a weighted average of the covariate-specific ROC curve for one biomarker and demonstrated the associated asymptotic properties. As an extension of Janes and Pepe's work, Kim and Huang (2017) proposed a classification model conditioning on a discrete covariate. Another common approach is to express biomarkers as a function of covariates. For instance, Schisterman et al. (2004) extended Su and Liu (1993) method to include covariates by representing biomarkers as a regression of covariates under the multivariate normality assumption.

However, it is still unclear how to adjust for continuous or discrete covariates without any distributional assumption. In this study, a new covariate-adjusted classification method for multiple biomarkers is proposed with no assumption on joint distributions of biomarkers and covariates. Technically, it is a two-stage method that biomarkers are first linearly combined by using Su and Liu (1993) method or Stepwise method (Kang et al., 2016), and then adjusting for covariates with the use of the maximum rank correlation estimators. Specifically, these parameter coefficients associated with covariates can be obtained by attaining the largest AUC. The asymptotic properties of these estimators are also studied thoroughly. We conduct an intensive simulation study to assess the performance of the proposed method in finite sample sizes. Our methodology is applied to the data of colorectal cancer and pancreatic cancer data for illustration. The proposed method can significantly improve the classification accuracy after adjusting for covariates. It is an easy-implemented approach, which is also robust against the distributional assumptions on biomarkers and covariates. In addition, we extend the proposed method to longitudinal biomarkers. The longitudinal biomarker profile of each subject can be represented by the spline coefficients using the method of the natural cubic spline. Those spline coefficients are treated as "biomarkers" in our previous method. As the first stage, the spline coefficients are optimally combined by Stepwise method (Kang et al., 2016). After that, covariates are incorporated by maximizing the AUC in the second stage. We evaluate the finite-sample performance of our method for longitudinal biomarkers through a simulation study. Alzheimer's disease data and the longitudinal data of primary biliary cirrhosis are used as real data applications. In sum, this extended approach can significantly reduce dimensions of longitudinal biomarker data and work well particularly when each subject is measured at different time points or has a different number of measurements.

The rest of this dissertation is organized as follows. In Chapter 2, we briefly introduce the area under the receiver operating characteristic curve and two popular classification methods for multiple biomarkers. In Chapter 3, our covariate-adjusted classification method for multiple cross-sectional biomarkers is proposed. The asymptotic properties of the parameter estimators in the proposed model are established. The performance of the proposed method

is evaluated by an intensive simulation study. For real data applications, we apply our method to the data of colorectal cancer and pancreatic cancer data. In Chapter 4, the proposed method is extended to longitudinal biomarkers. The longitudinal data of Alzheimer's disease and primary biliary cirrhosis data are used to illustrate our extended method for longitudinal biomarkers. We conduct a simulation study to assess the performance of the proposed method. Discussion and conclusions are given in Chapter 5.

Chapter 2

The AUC and classification methods for multiple biomarkers

In this chapter, the area under the receiver operating characteristic (ROC) curve is introduced concisely. We also review two popular classification methods for multiple biomarkers, which are used to find an optimal linear combination of biomarkers in our approach.

2.1 The area under the receiver operating characteristic curve

A brief introduction about the area under the ROC curve (AUC) is given in this section. We assume a continuous biomarker is measured on n_d subjects of the disease group and n_h subjects of the health group. Let Y_{D_p} be the biomarker value of the p^{th} subject in the disease group and Y_{H_q} be the biomarker value of the q^{th} subject in the health group with the corresponding cumulative distribution function F_D and F_H, where $p = 1, ..., n_d$ and $q = 1, ..., n_h$. For a threshold c ($c \in R$), the sensitivity and specificity of the biomarker are defined as $P(Y_D > c) = 1 - F_D(c)$ and $P(Y_H \leq c) = F_H(c)$, which are the correct detection rates for the disease and health groups. The ROC curve of the biomarker is generated in a plot of $\{1 - F_H(c), 1 - F_D(c)\}$ by given all possible values of c. The area under the ROC

curve (AUC) is a popular index to evaluate the classification rate. It is defined as

$$AUC = \int_0^1 \Big[1 - F_D(F_H^{-1}(1-t))\Big] dt,$$

where t = (1 - specificity) and $0 \leq t \leq 1$. The value of AUC close to 1 indicates a better classification performance. Bamber (1975) showed that AUC is equivalent to the probability of $P(Y_D > Y_H)$ mathematically. In practice, the computation of AUC is performed much faster by using the Mann-Whitney U statistic. The AUC can be estimated empirically as

$$\widehat{AUC} = \frac{1}{n_d n_h} \sum_{p=1}^{n_d} \sum_{q=1}^{n_h} I(Y_{D_p} > Y_{H_q}),$$

where $I(\cdot)$ is the indicator function.

2.2 The classification methods for multiple biomarkers

Let $Y_{D_p} = (Y_{D_{p1}}, Y_{D_{p2}}, ..., Y_{D_{pk}})$ and $Y_{H_q} = (Y_{H_{q1}}, Y_{H_{q2}}, ..., Y_{H_{qk}})$ be the values of k biomarkers for the p^{th} subject in the disease group and the q^{th} subject in the health group, where $p = 1, ..., n_d$ and $q = 1, ..., n_h$. It is often of great interest to look for a vector of linear combination coefficients $\alpha = (\alpha_1, ..., \alpha_k)'$ that can optimally combine multiple biomarkers to achieve the corresponding maximum AUC. Many classification methodologies have been proposed to find such linear combination of multiple biomarkers in the literature. Among them, we present two popular methods, Su and Liu (1993) method and Stepwise method proposed by Kang et al. (2016) in this section.

2.2.1 Su and Liu method

Su and Liu (1993) suggested a linear combination of biomarkers to attain the largest AUC comparing all other possible linear combinations with assuming the multivariate normality biomarkers.

Assume that $Y_{D_p} \sim N_k(\mu_D, \Sigma_D)$ and $Y_{H_q} \sim N_k(\mu_H, \Sigma_H)$ are the biomarkers for the disease

and health groups, where N_k is the $k-$dimensional multivariate normal distribution. The vector of the best linear combination coefficients proposed by Su and Liu (1993) is

$$\alpha \propto (\Sigma_D + \Sigma_H)^{-1}(\mu_D - \mu_H),$$

with the corresponding AUC equals to

$$\Phi\Big(\sqrt{(\mu_D - \mu_H)'(\Sigma_D + \Sigma_H)^{-1}(\mu_D - \mu_H)}\Big),$$

where Φ denotes the standard normal distribution. A consistent estimate of α is

$$\Big(\frac{S_D}{n_d - 1} + \frac{S_H}{n_h - 1}\Big)^{-1}(\bar{Y}_D - \bar{Y}_H),$$

where n_1 and n_2 are the sample sizes, $\bar{Y}_D$ and $\bar{Y}_H$ are the sample means of biomarkers, and $S_D = \sum_{p=1}^{n_d}(Y_{D_p} - \bar{Y}_D)'(Y_{D_p} - \bar{Y}_D)$ and $S_H = \sum_{q=1}^{n_h}(Y_{H_q} - \bar{Y}_H)'(Y_{H_q} - \bar{Y}_H)$ are the sample sums of squares of biomarkers for the disease and health groups.

2.2.2 Stepwise method

Kang et al. (2016) proposed a nonparametric stepwise approach to search for the optimal linear coefficient of a single biomarker at each step by maximizing the empirical AUC without distributional assumptions on biomarkers. It is the extension of the method developed by Pepe and Thompson (2000).

Technically, Stepwise approach is started with sorting k ($k \geq 2$) biomarkers from the largest to smallest based on the estimated AUC of each biomarker, then using the method proposed by Pepe and Thompson (2000) to find the vector of linear combination coefficients for the first two biomarkers $\hat{\alpha} = (1, \hat{\alpha}_2)'$ by attaining the largest corresponding AUC, where

$$\widehat{AUC} = \frac{1}{n_d n_h}\sum_{p=1}^{n_d}\sum_{q=1}^{n_h} I(Y_{D_{p1}} + \hat{\alpha}_2 Y_{D_{p2}} > Y_{H_{q1}} + \hat{\alpha}_2 Y_{H_{q2}}).$$

Next, the third largest biomarker is combined in the same manner. The rest biomarkers are incorporated one at each step by repeating the process until the last biomarker is included. The empirical AUC of all biomarkers is estimated as

$$\widehat{AUC} = \frac{1}{n_d n_h} \sum_{p=1}^{n_d} \sum_{q=1}^{n_h} I(Y_{D_{p1}} + \hat{\alpha}_2 Y_{D_{p2}} + ... + \hat{\alpha}_k Y_{D_{pk}} > Y_{H_{q1}} + \hat{\alpha}_2 Y_{H_{q2}} + ... + \hat{\alpha}_k Y_{H_{qk}}).$$

Overall, Stepwise method (Kang et al., 2016) is easy to implement and interpret the linear combination coefficients. It is also robust to the distributional assumptions on multiple biomarkers.

Chapter 3

The proposed covariate-adjusted classification method for cross-sectional biomarkers

In this chapter, we propose a new classification method for cross-sectional biomarkers with covariate adjustment.

3.1 The proposed method for cross-sectional biomarkers

In addition to k biomarkers, suppose that m ($m \geq 1$) covariates are also observed from n_d subjects in the disease group and n_h subjects in the health group. Let $Z_{D_p} = (Z_{D_{p1}}, Z_{D_{p2}}, ..., Z_{D_{pm}})$ and $Z_{H_q} = (Z_{H_{q1}}, Z_{H_{q2}}, ..., Z_{H_{qm}})$ be m covariates values of the p^{th} subject for the disease group and the q^{th} subject for the health group, where $p = 1, ..., n_d$ and $q = 1, ..., n_h$.

The proposed method has two stages. In the first stage, k biomarkers are optimally combined by using Su and Liu (1993) method or Stepwise method (Kang et al., 2016) to obtain the vector of linear combination coefficients α. If the biomarkers follow the multivariate normal distribution, both Su and Liu method and Stepwsie method are used. Stepwise

method is only used when biomarkers are not normally distributed. Let $Y_{D_C} = \alpha' Y_D$ and $Y_{H_C} = \alpha' Y_H$, where $Y_D = (Y_{D_1}, Y_{D_2}, ..., Y_{D_k})$ and $Y_H = (Y_{H_1}, Y_{H_2}, ..., Y_{H_k})$. It is clear that the random variables Y_{D_C} and Y_{H_C} are the optimal combinations of all k biomarkers for the disease and health groups, respectively. The AUC for those combinations of biomarkers is

$$AUC_W = P(Y_{D_C} > Y_{H_C}). \tag{3.1}$$

Next, suppose $\beta = (\beta_1, ..., \beta_m)'$ is an unknown vector of linear combination coefficients for m covariates and γ is an unknown coefficient of the biomarker combination, where $\gamma \in R$ and $\beta \in R^m$. Let $\beta' Z_D$ and $\beta' Z_H$ be the combinations of the entire m covariates for the disease and health groups, where $Z_D = (Z_{D_1}, Z_{D_2}, ..., Z_{D_m})$ and $Z_H = (Z_{H_1}, Z_{H_2}, ..., Z_{H_m})$. Based on the biomarker combinations $\gamma' Y_{D_C}$ and $\gamma' Y_{H_C}$, the classification can be conducted by including the linear combinations of covariates. The proposed covariate-adjusted AUC is defined as

$$AUC^*(\gamma, \beta) = P\Big(Y_{D_C} + \frac{\beta' Z_D}{\gamma} > Y_{H_C} + \frac{\beta' Z_H}{\gamma}\Big), \tag{3.2}$$

which incorporates the covariates naturally in the classification model. The "best" coefficient parameters (γ_0, β_0) can be obtained by maximizing the AUC* in Equation 3.2, that is

$$(\gamma_0, \beta_0) = \underset{\gamma \in R,\, \beta \in R^m}{\operatorname{argmax}} P\Big(Y_{D_C} + \frac{\beta' Z_D}{\gamma} > Y_{H_C} + \frac{\beta' Z_H}{\gamma}\Big). \tag{3.3}$$

The following theorem demonstrates that the proposed method improves the classification accuracy after adjusting for covariates.

Theorem 1. *Assume the AUC with covariate adjustment given in Equation 3.2 is maximized by the "best" coefficient parameters (γ_0, β_0), where $\gamma_0 \in R$ and $\beta_0 \in R^m$, then*

$$AUC^*(\gamma_0, \beta_0) = \max_{\gamma \in R,\, \beta \in R^m} P\Big(Y_{D_C} + \frac{\beta' Z_D}{\gamma} > Y_{H_C} + \frac{\beta' Z_H}{\gamma}\Big) \geq P(Y_{D_C} > Y_{H_C}) = AUC_W.$$

In words, the proposed covariate-adjusted AUC is at least and greater than the AUC without

covariate adjustment.

The proof is straightforward. Note that $P\Big(Y_{D_C} + \frac{\beta' Z_D}{\gamma} > Y_{H_C} + \frac{\beta' Z_H}{\gamma}\Big) = P\Big(Y_{D_C} > Y_{H_C}\Big)$, when $\beta = 0$. If $\beta \neq 0$, by Equation 3.3, there always exists one (γ_0, β_0) such that $\max_{\gamma \in R,\, \beta \in R^m} P\Big(Y_{D_C} + \frac{\beta' Z_D}{\gamma} > Y_{H_C} + \frac{\beta' Z_H}{\gamma}\Big) > P(Y_{D_C} > Y_{H_C})$. In other words, the maximized AUC with covariate adjustment is larger than the one without adjusting for covariates. Thus, Theorem 1 holds.

Empirically, the optimal coefficient parameters can be estimated as follows,

$$(\hat{\gamma}, \hat{\beta}) = \underset{\gamma \in R,\, \beta \in R^m}{\operatorname{argmax}} \frac{1}{n_d n_h} \sum_{p=1}^{n_d} \sum_{q=1}^{n_h} I\Big(Y_{D_{C_p}} + \frac{\beta' Z_{D_p}}{\gamma} > Y_{H_{C_q}} + \frac{\beta' Z_{H_q}}{\gamma}\Big),$$

where $I(\cdot)$ is the indicator function.

The estimated covariate-adjusted AUC of the proposed method is given as

$$\widehat{AUC}^*(\hat{\gamma}, \hat{\beta}) = \frac{1}{n_d n_h} \sum_{p=1}^{n_d} \sum_{q=1}^{n_h} I\Big(Y_{D_{C_p}} + \frac{\hat{\beta}' Z_{D_p}}{\hat{\gamma}} > Y_{H_{C_q}} + \frac{\hat{\beta}' Z_{H_q}}{\hat{\gamma}}\Big).$$

It is worth noting that $(\hat{\gamma}, \hat{\beta})$ is a special case of the maximum rank correlation (MRC) estimator defined in Han (1987). The asymptotic properties of the MRC estimator have been established in the literature.

3.2 Asymptotic properties

In this section, the asymptotic properties of the estimator for the parameter $\theta = (\gamma, \beta)$ in the proposed method are investigated. We note that $\hat{\theta} = (\hat{\gamma}, \hat{\beta})$ is a special case of the MRC estimator. Therefore, we can derive the consistency and asymptotic normality of $\hat{\theta}$ by adopting the proofs for the MRC estimator in Han (1987) and Sherman (1993).

Consider a binary outcome D of a screening or diagnosis study, where $D = 1$ refers to the disease group and $D = 0$ refers to the health group, without loss of generality. Let Y_C be the linear combination of k ($k \geq 2$) biomarkers and $Z = (Z_1, Z_2, ..., Z_m)$ be m ($m \geq 1$) covariates for n subjects. Denote $X = (Y_C, Z)$. For the uniqueness of estimation, the parameters are

normalized as $\theta^* = \theta/\|\theta\|$, where $\|\cdot\|$ is a matrix norm. Let $\hat{\theta}^*$ be the parameter estimator and θ_0^* be the true value of θ^*.

The following assumptions are needed to show the consistency of $\hat{\theta}^*$.

Assumption 1. The true parameter value θ_0^* is an interior point of Θ, which is compact.

Assumption 2. Let S_x be the support of the vector of X.

(i) S_x is not contained in any proper linear subspace of R^{m+1}.

(ii) The m^{th} component of X has everywhere positive density conditional on other components.

Theorem 2. *Under Assumption 1 and Assumption 2,*

$$\hat{\theta}^* \xrightarrow{p} \theta_0^*, \quad as\ n \to \infty.$$

In words, the parameter estimator converges to the value of true parameter in probability asymptotically.

The proof is provided in Appendix A.1. This theorem shows the consistency of the proposed parameter estimators.

Next, the asymptotic normal distribution property of $\hat{\theta}^*$ is studied. Suppose $V = (D, X)$ has a distribution P on a set S such that $P(D|X)$ has a monotone increasing transformation, where $S = \{0, 1\} \otimes R^{m+1}$. For each $v = (d, x) \in S$ and $\theta^* \in \Theta$, define

$$\varphi(v, \theta^*) = E\left\{ I\Big(d > D\Big) I\Big(x'\theta^* > X'\theta^*\Big) \right\} + E\left\{ I\Big(d < D\Big) I\Big(x'\theta^* < X'\theta^*\Big) \right\}.$$

Under the binary outcome $D = \{0, 1\}$, we can show that $\varphi(v, \theta^*)$ is the probability $P\Big(X'_{(D=1)}\theta^* > X'_{(D=0)}\theta^*\Big)$, which is actually the covariate-adjusted AUC* proposed in Section 3.1, where $X_{(D=1)}$ is the vector X given $D = 1$ (disease group) and $X_{(D=0)}$ is the vector X given $D = 0$ (health group). To show the asymptotic normality of $\hat{\theta}^*$, the additional assumption is given as follows.

Assumption 3.

(i) For each $v \in S$, the second partial derivatives of $\varphi(v, \theta^*)$ exist and are bounded on $\varkappa$, where $\varkappa$ is a neighborhood of θ_0^*.

(ii) There exists an integrable function $M(x)$ such that $\left|\left|\frac{\partial}{\partial \theta^{*2}}\varphi(v,\theta^*) - \frac{\partial}{\partial \theta_0^{*2}}\varphi(v,\theta_0^*)\right|\right| \leq M(v)|\theta^* - \theta_0^*|$ for all $v \in S$ and $\theta^* \in \varkappa$, where $\left|\left| \cdot \right|\right|$ is a matrix norm.

(iii) $E\left|\frac{\partial}{\partial \theta_0^*}\varphi(v,\theta_0^*)\right|^2 < \infty$ and $E\left\{\frac{\partial}{\partial \theta_0^{*2}}\varphi(v,\theta_0^*)\right\}$ is negative definite for all $v \in S$.

Assumption 3 is the sufficient condition to apply a Taylor expansion of $\varphi(v, \theta^*)$ about θ_0^*.

Theorem 3. *Suppose Assumption 1 - 3 hold, then*

$$\sqrt{n}(\hat{\theta}^* - \theta_0^*) \xrightarrow{D} N(0, \Lambda^{-1}\Sigma\Lambda^{-1}), \text{ as } n \to \infty,$$

where $2\Lambda = E\left\{\frac{\partial}{\partial \theta_0^{*2}}\varphi(v,\theta_0^*)\right\}$ *and* $\Sigma = E\left\{\left[\frac{\partial}{\partial \theta_0^*}\varphi(v,\theta_0^*)\right]\left[\frac{\partial}{\partial \theta_0^*}\varphi(v,\theta_0^*)\right]'\right\}$ *for all* $v \in S$.

The proof is given in Appendix A.2. This result presents the parameter estimators have the asymptotic normal distribution. In practice, it is suggested to use the bootstrap resampling method to estimate the variability of the proposed parameter estimators (see, for example, Zhang and Li, 2011).

3.3 Simulation study

In this section, an intensive simulation study is conducted to evaluate the performance of the proposed method in the finite sample sizes. These performances are compared to those of other classification methods without covariate adjustment, which are Su and Liu (1993) method and Stepwise method (Kang et al., 2016).

The performance of the proposed method is studied under two data generating schemes of covariates: (1) two covariates from normal and binomial distributions with different means for the disease and health groups; (2) multiple covariates from different distributions with equal means and variances for the disease and health groups to investigate the influence of including irrelevant covariates on classification. For each scheme, four biomarkers for

the disease and health groups are generated in two ways: (1) from the multivariate normal distributions with different mean vectors and equal variance matrix for disease and health groups as follows,

$$\mu_D = \begin{pmatrix} 1.2 \\ 1.4 \\ 1.6 \\ 1.8 \end{pmatrix}, \ \mu_H = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \ \Sigma_D = \Sigma_H = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix};$$

(2) from the following four different distributions,

$$Y_{D_1} \sim N(0.20, 1.0),\ Y_{D_2} \sim \text{Pois}(0.30),\ Y_{D_3} \sim \text{Exp}(0.40),\ Y_{D_4} \sim \Gamma(0.50, 1.0),$$
$$Y_{H_1} \sim N(0.10, 1.0),\ Y_{H_2} \sim \text{Pois}(0.15),\ Y_{H_3} \sim \text{Exp}(0.20),\ Y_{H_4} \sim \Gamma(0.25, 1.0).$$

Su and Liu method and Stepwise method are used for normally distributed biomarkers, and Stepwise method is only used for the biomarkers from four non-normal distributions. For all simulations, 1000 replicates are conducted for sample sizes from 20 to 60 for each group.

Under the first data generating scheme, two covariates for the health group Z_{H_1} and Z_{H_2} are generated from $N(\mu_{Z_{H_1}}, 1)$ and $\text{Bin}(p_{Z_{H_2}})$, where $\mu_{Z_{H_1}} = 0$ and $p_{Z_{H_2}} = 0.2$. We generate two covariates for the disease group Z_{D_1} and Z_{D_2} from $N(\mu_{Z_{D_1}}, 1)$ and $\text{Bin}(p_{Z_{D_2}})$, where $\mu_{Z_{D_1}} \in \{0, 1\}$ and $p_{Z_{D_2}} \in \{0.2, 0.5, 0.8\}$. The results for the normally distributed biomarkers are reported in Table 3.1. For equal sample sizes, all estimated AUCs of the proposed method are much larger than those of Su and Liu method and Stepwise method, suggesting the AUCs are increased significantly after adjusting for covariates using the proposed method. The largest improvement on the estimated AUCs achieves with the biggest differences in the means of both covariates for the disease and health groups. Even only the second covariate for the two groups have the different means, there is still an obvious increasing on the estimated AUCs of the proposed method. The same results are obtained for unequal sample sizes of the disease and health groups. Table 3.2 shows similar results

when the four biomarkers are from different distributions.

Table 3.1: The estimated AUCs for two covariates and four biomarkers

Sample size	Covariates		Without covariate adjustment		Covariate-adjusted	
(n_d, n_h)	$\mu_{Z_{D_1}}$	$p_{Z_{D_2}}$	Su & Liu	Stepwise	Su & Liu	Stepwise
(20, 20)	1	0.2	0.784	0.785	0.879	0.871
	1	0.5	0.783	0.783	0.899	0.892
	1	0.8	0.780	0.781	0.939	0.935
	0	0.5	0.784	0.783	0.850	0.840
	0	0.8	0.781	0.782	0.910	0.905
(40, 40)	1	0.2	0.760	0.760	0.855	0.850
	1	0.5	0.758	0.756	0.875	0.870
	1	0.8	0.760	0.759	0.926	0.924
	0	0.5	0.759	0.759	0.817	0.812
	0	0.8	0.760	0.759	0.896	0.893
(60, 60)	1	0.2	0.751	0.750	0.845	0.842
	1	0.5	0.752	0.750	0.868	0.865
	1	0.8	0.751	0.749	0.921	0.920
	0	0.5	0.751	0.749	0.806	0.802
	0	0.8	0.753	0.752	0.889	0.887
(20, 40)	1	0.5	0.770	0.769	0.885	0.879
(20, 60)			0.769	0.767	0.883	0.878
(40, 20)			0.770	0.770	0.888	0.882
(40, 60)			0.757	0.756	0.874	0.870
(60, 20)			0.766	0.765	0.881	0.876
(60, 40)			0.755	0.754	0.874	0.871

Next, under the second data generating scheme, multiple covariates for the disease and health groups are generated from the following different distributions with the same means

Table 3.2: The estimated AUCs for two covariates and four biomarkers based on Stepwise method

Sample size	Covariates		Without covariate	
(n_d, n_h)	$\mu_{Z_{D_1}}$	$p_{Z_{D_2}}$	adjustment	Covariate-adjusted
(20, 20)	1	0.2	0.791	0.867
	1	0.5	0.796	0.897
	1	0.8	0.792	0.934
	0	0.5	0.793	0.841
	0	0.8	0.790	0.910
(40, 40)	1	0.2	0.765	0.842
	1	0.5	0.767	0.868
	1	0.8	0.765	0.922
	0	0.5	0.766	0.811
	0	0.8	0.765	0.894
(60, 60)	1	0.2	0.752	0.834
	1	0.5	0.754	0.859
	1	0.8	0.754	0.915
	0	0.5	0.754	0.801
	0	0.8	0.755	0.887
(20, 40)	1	0.5	0.776	0.876
(20, 60)			0.774	0.875
(40, 20)			0.782	0.880
(40, 60)			0.756	0.863
(60, 20)			0.774	0.874
(60, 40)			0.760	0.862

and variances,

$$Z_{D_1}, Z_{H_1} \sim N(0, 1.5^2), \quad Z_{D_2}, Z_{H_2} \sim \text{Bin}(0.2),$$

$$Z_{D_3}, Z_{H_3} \sim \text{F}(1, 2), \qquad Z_{D_4}, Z_{H_4} \sim \chi^2_2,$$

$$Z_{D_5}, Z_{H_5} \sim \text{Pois}(2), \qquad Z_{D_6}, Z_{H_6} \sim \text{Exp}(0.5).$$

Clearly, the covariates are not relevant to the classification. The sample size for each group

is raised to 200 and 1000. The result for four biomarkers from the multivariate normal distribution is summarized in Table 3.3. It suggests that the estimated AUCs of the proposed method increase with adding more irrelevant covariates for small to moderate sample sizes. However, when the sample sizes are large for both groups, the AUCs produced by the proposed method have almost no change and tend to be same to the estimated AUCs of Su and Liu method and Stepwise method, regardless the number of adjusted covariates. It shows that adding more unrelated covariates has no impact on the classification accuracy of the proposed method when the sample size is large. The similar result for four non-normally distributed biomarkers is summarized in Table 3.4.

Table 3.3: The estimated AUCs for four biomarkers and multiple irrelevant covariates

Sample size		Without covariate adjustment		Covariate-adjusted	
(n_d, n_h)	Covariates	Su & Liu	Stepwise	Su & Liu	Stepwise
(20, 20)	$Z_1 + Z_2$	0.784	0.784	0.819	0.809
	$Z_1 + Z_2 + Z_3 + Z_4$	0.784	0.783	0.834	0.828
	$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$	0.781	0.782	0.845	0.845
(40, 40)	$Z_1 + Z_2$	0.759	0.758	0.780	0.774
	$Z_1 + Z_2 + Z_3 + Z_4$	0.758	0.757	0.787	0.785
	$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$	0.759	0.758	0.799	0.799
(200, 200)	$Z_1 + Z_2$	0.741	0.739	0.745	0.742
	$Z_1 + Z_2 + Z_3 + Z_4$	0.741	0.739	0.748	0.746
	$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$	0.742	0.740	0.751	0.749
(1000, 1000)	$Z_1 + Z_2$	0.737	0.736	0.738	0.736
	$Z_1 + Z_2 + Z_3 + Z_4$	0.738	0.736	0.739	0.737
	$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$	0.737	0.735	0.739	0.737

Finally, we further conduct a simulation to study the asymptotic property of the proposed parameter estimators. Under the second simulation scheme with four non-normally distributed biomarkers, the means and standard errors (SEs) of the proposed parameter estimators for four irrelevant covariates are computed with the sample sizes of 100, 200, 500 and

Table 3.4: The estimated AUCs for four biomarkers and multiple irrelevant covariates based on Stepwise method

Sample size (n_d, n_h)	Covariates	Without covariate adjustment	Covariate-adjusted
(20, 20)	$Z_1 + Z_2$	0.794	0.808
	$Z_1 + Z_2 + Z_3 + Z_4$	0.792	0.825
	$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$	0.794	0.843
(40, 40)	$Z_1 + Z_2$	0.767	0.777
	$Z_1 + Z_2 + Z_3 + Z_4$	0.763	0.786
	$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$	0.765	0.796
(200, 200)	$Z_1 + Z_2$	0.739	0.742
	$Z_1 + Z_2 + Z_3 + Z_4$	0.738	0.744
	$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$	0.738	0.747
(1000, 1000)	$Z_1 + Z_2$	0.733	0.733
	$Z_1 + Z_2 + Z_3 + Z_4$	0.732	0.734
	$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$	0.732	0.734

1000. Only Stepwise method is used since four biomarkers are not normally distributed. It is believed that the true values of β^*s should be 0 because the included covariates are actually useless to the classification. The results are reported in Table 3.5. It seems that the estimates of γ^* converge to 1 and β^*s are approximately unbiased. The associated SEs decrease as sample size increases, which demonstrates the consistency of the proposed parameter estimators.

Overall, the proposed method can greatly improve the AUC with covariate adjustment. The improvement highly depends on the discrepancies of covariates between the disease and health groups. When the sample size is large, adding many irrelevant covariates does not influence the classification performance of the proposed model.

Table 3.5: The estimates (S.E.) that associated with four irrelevant covariates based on Stepwise method

Sample size $n_d = n_h$	$\hat{\gamma}^*$	$\hat{\beta}_1^*$	$\hat{\beta}_2^*$	$\hat{\beta}_3^*$	$\hat{\beta}_4^*$
100	0.857	0.006	0.020	-0.001	0.005
	(0.438)	(0.110)	(0.235)	(0.003)	(0.007)
200	0.938	-0.001	0.030	0	0.003
	(0.281)	(0.079)	(0.179)	(0.013)	(0.04)
500	0.987	0	0.027	0	0.001
	(0.086)	(0.042)	(0.125)	(0.003)	(0.022)
1000	0.996	0.003	0.018	0	0
	(0.008)	(0.029)	(0.080)	(0.002)	(0.014)

3.4 Applications to cancer data

In this section, the proposed model is applied to the data of colorectal cancer and pancreatic cancer for illustration.

3.4.1 Colorectal cancer data

A colorectal cancer observational study was conducted in Taiwan to investigate new biomarkers for colorectal cancer detection. The data of 135 patients diagnosed with colorectal cancer and 78 individuals with no colorectal cancer were collected from Taipei Medical University Hospital and Taipei Veterans General Hospital for the study. *BEND5*, *PPP2R5C* and *EHD3* are three tumor suppressor genes and oncogenes closely related to the formation of colorectal cancer, which are measured as continuous biomarkers in the study. In addition, *BEND5* and *PPP2R5C* can be categorized as binary biomarkers. Age and gender are considered as two covariates in the study, which are summarized in Table 3.6.

Only Stepwise method is capable since the distributions of both continuous and binary biomarkers are not the multivariate normal. The results for the continuous biomarkers are

Table 3.6: The descriptive statistics of covariates for colorectal cancer data

		Age (years)			Gender(%)	
Group	n	Range	Mean	Std. Dev	Male	Female
Disease	135	[37, 92]	64.57	13.63	54.81	45.19
Health	78	[35, 62]	50.82	16.25	63.29	36.71

reported in Table 3.7. The estimated AUC and accuracy of Stepwise method are all around 0.5, which imply the low classification rate of using biomarkers only. However, the AUC and accuracy of the proposed method are close to 0.9, which are increased considerably with covariate adjustment. In Table 3.6, it is easy to see the differences in the descriptive statistics of two covariates between the disease and health groups, which are helpful to the classification performance under the proposed method. The ROC curves given in Figure 3.1 show the improvement on the AUC.

Table 3.7: The classification results by using the continuous biomarkers for colorectal cancer data based on Stepwise method

Continous biomarkers	Covariates	Without covariate adjustment	Covariate-adjusted
		AUC	
BEND5 + *PPP2R5C* + *EHD3*	Age + Gender	0.552	0.886
		Accuracy	
		0.453	0.875

For the binary biomarker data, the results in Table 3.8 show that the remarkable improvements on both estimated AUCs and accuracies are still obtained from comparing the proposed method to Stepwise method without covariate adjustment. Figure 3.2 gives the ROC curves with and without adjustment of covariates.

Figure 3.1: The ROC curve using the continuous biomarkers for colorectal cancer data

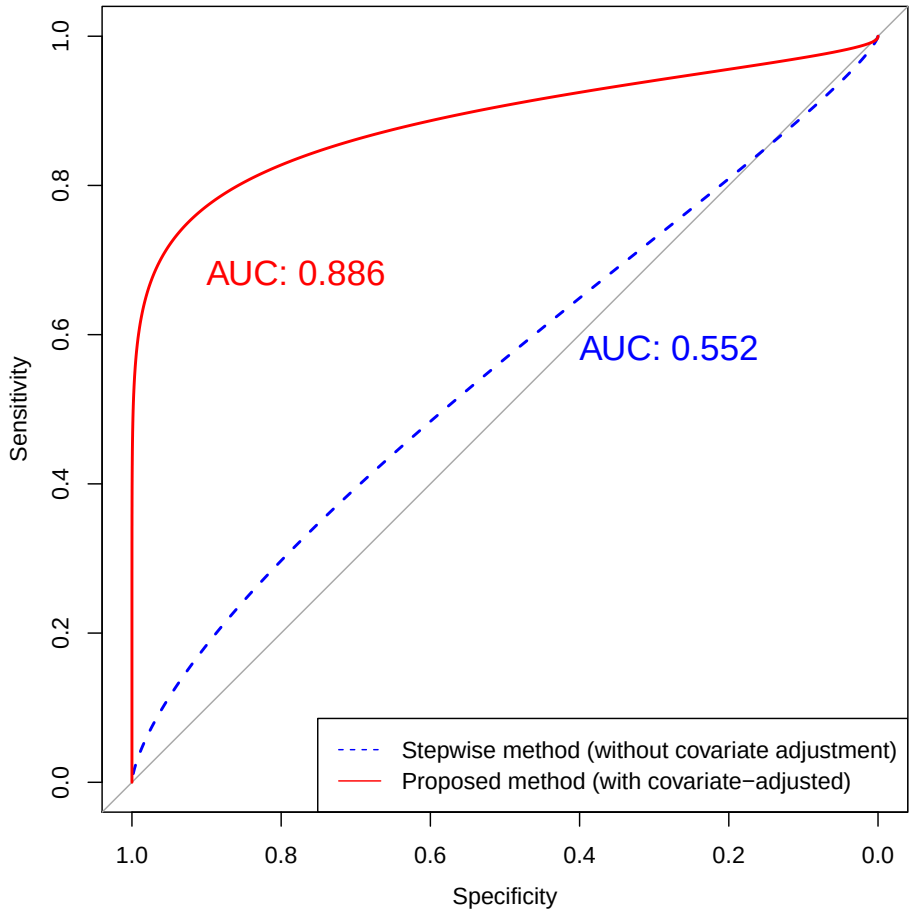
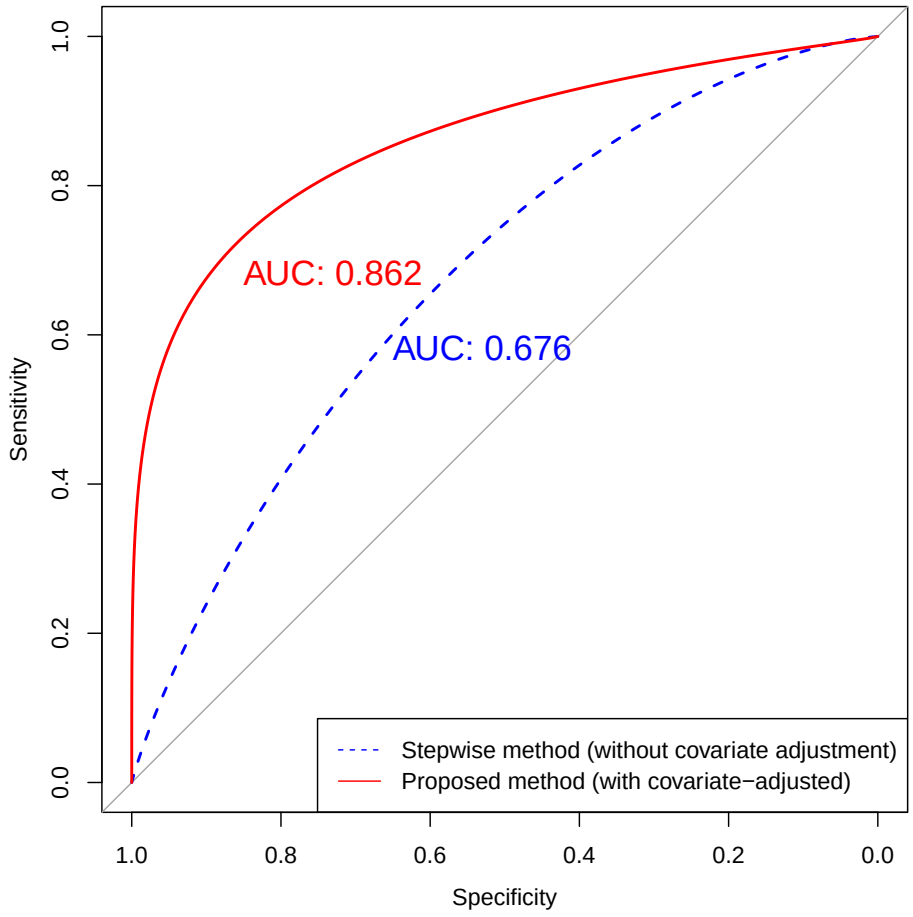


Table 3.8: The classification results by using the binary biomarkers for colorectal cancer data based on Stepwise method

Binary biomarkers	Covariates	Without covariate adjustment	Covariate-adjusted
		AUC	
$BEND5_B$ + $PPP2R5C_B$	Age + Gender	0.676	0.862
		Accuracy	
		0.642	0.827

$BEND5_B$ and $PPP2R5C_B$ are the binary biomarkers

Figure 3.2: The ROC curve using the binary biomarkers for colorectal cancer data



3.4.2 Pancreatic cancer data

In a pancreatic cancer case-control study, the fluorescence-based nanobiosensors for the early detection of pancreatic cancer were developed by a team from Kansas State University (Kalubowilage, 2017). The nanobiosensors are particularly sensitive to detect the activities of protease and arginase in blood serum, which are necessary for tumors development and progression. To demonstrate the potential of the nanobiosensor technology for early detecting pancreatic cancer, the serum samples of 35 patients who have been diagnosed with pancreatic cancer and 48 healthy volunteers were collected from the University of Kansas Cancer Center. Eight biomarkers including arginase, cathepsin B, cathepsin E, UpA, MMPs 1, MMPs 3, MMPs 9 and neutrophil elastase are measured in each serum sample for the study (Udukala et al. 2016; Kalubowilage 2017). The descriptive statistics of two covariates in the study, age and gender, are given in Table 3.9.

Table 3.9: The descriptive statistics of covariates for pancreatic cancer data

		Age (years)			Gender(%)	
Group	n	Range	Mean	Std. Dev	Male	Female
Disease	35	[19, 81]	64.09	13.66	48.57	51.43
Health	48	[19, 81]	63.92	12.43	50.00	50.00

Since the biomarkers are not normally distributed, we only use Stepwise method to combine biomarkers. The results in Table 3.10 present that the AUC and accuracy of the proposed method are almost same as those of Stepwise method. It is noted that the descriptive statistics of two covariates for the disease and health groups in Table 3.9 are almost same because they were controlled by the study. Therefore, little additional information to the classification is delivered by two covariates. However, adjusting for the controlled covariates by the proposed method does not jeopardize the classification performance.

Table 3.10: The classification results for pancreatic cancer data based on Stepwise method

Biomarkers	Covariates	Without covariate adjustment	Covariate-adjusted
		AUC	
Arginase + Cathepsin B + Cathepsin E + UpA + MMPs 1 + MMPs 3 + MMPs 9 + Neutrphil Elastase	Age + Gender	0.8607	0.8613
		Accuracy	
		0.8433	0.8433

3.5 Discussion

In this study, we have developed a new classification method for multiple cross-sectional biomarkers to adjust for covariates, which does not require any assumption on joint distributions of biomarkers and covariates. The approach of the proposed method is using the covariate adjustment over the optimal biomarker combination that maximizes AUC. The proposed method can significantly improve the classification accuracy when covariates for the disease and health groups have discrepancies. In addition, the proposed model works well for discrete or continuous covariates. It is also worth noting that the proposed covariate-adjusted method is capable not only for multiple biomarkers but also works for the case where only one biomarker is involved.

However, the proposed method has some limitations. The considerable improvement on the classification accuracy under the proposed model mainly depends on the discrepancies of covariates between the disease and health groups. When the discrepancy is small, the proposed method may not improve the classification performance considerably by adjusting for such covariates. It is expected to increase the classification rate as including many relevant covariates. However, adjusting for irrelevant covariates may overestimate the accuracy when the sample size is small. For large sample sizes, including more irrelevant covariates does not affect the classification performance under the proposed model.

Overall, the proposed method has provided an easy-implement and distribution-free

covariate-adjusted approach to improve the classification accuracy for multiple cross-sectional biomarkers. It is also robust to include irrelevant covariates when the sample size is large.

Chapter 4

The proposed covariate-adjusted classification method for longitudinal biomarkers

In this chapter, we extend the proposed method to longitudinal biomarkers for the binary and time-independent outcome.

4.1 Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder which will lead to memory loss, cognitive impairment, dementia, and total disability eventually. To date, no intervention or cure can effectively halt, delay or even reverse the progression of AD, thus most of AD patients would require medical assistance and care from professional caregivers, making AD become one of the most costly chronic diseases in the United States. By 2050, an estimated over 14 million Americans will be affected by AD (Mistur et al., 2009; Alzheimer's Association et al., 2012), which will become a great burden of American society.

More than a hundred active AD longitudinal and clinical studies around the world are conducted to study the life course of AD and search for effective treatments. In 2004, the

Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal study was established to track the progression of AD with biomarkers (ADNI). In the ADNI study, the imaging, clinical, genetic and biochemical biomarkers of each participant were measured repeatedly over time. These biomarkers can be particularly used for the early detection of AD. Our study is motivated by using those longitudinal biomarkers in the ADNI study to identify dementia or AD patients. In general, longitudinal biomarkers are considered as a special case of functional biomarkers, which are typically recorded in higher frequencies and dimensions (Rice, 2004; Zhao et al., 2004).

The classification approaches for functional data have been well studied in the past few decades. One major approach is the functional linear discriminant analysis (FLDA), which is an extension of the classical linear discriminant analysis (LDA). This approach classifies a new subject into a group based on its maximum posterior probability (Hall et al., 2001; James and Hastie, 2001). Another popular approach is the functional regression methods to model the relationships between the functional predictors and the binary response variable (James, 2002; Müller and Stadtmüller, 2005; Leng and Müller, 2005). In addition to those two common approaches, the functional support vector machine approach (Rossi and Villa, 2006) and the weighted distance approach (Alonso et al., 2012) are also proposed with the concept of nonparametric.

As a special case of functional data, the classification approaches for longitudinal data are somewhat different. The conventional approaches for longitudinal classification are the generalized estimating equations (Liang and Zeger, 1986) and the generalized linear mixed models (GLMM) (Breslow and Clayton, 1993). The extensions of the generalized linear models (GLM) are also the most common classification methods for longitudinal data. For example, Zeger et al. (1985) extended the logistic regression models to represent the marginal probabilities as the logistic functions of the predictors. Erkanli et al. (2001) proposed the non-homogeneous Markov regression models in which the transition probability is expressed using the logistic regression models for the subject's past outcomes and predictors. Besides, the linear discriminate analysis under the mixed models have been suggested for longitudinal classification (Tomasko et al., 1999; Marshall and Barón, 2000). Bagui and Mehra (1999)

proposed a multi-stage rank nearest-neighbor rule, which extends the k-nearest neighbors to the classification for longitudinal data.

Several methods have been developed to incorporate the covariates into the functional data classification. In general, covariates such as age and gender may deliver the additional information to the outcomes of patients. For example, it is known that aging is significantly related to AD and many cancers. Therefore, incorporating covariates in the classification model may further improve the classification performance (Rowe et al., 1976; Lebowitz, 1996). The most common way to accommodate covariates is using the functional regression models. For example, Chiou et al. (2003) proposed a functional smooth random-effects model to incorporate covariates by the conditional distributions of the functional principal components (FPC) scores. Cardot (2007) suggested the conditional functional principal component analysis (FPCA) to relate covariates to the conditional mean and covariance functions nonparametrically. Recently, Li et al. (2017) presented a subspace projection classifier to adjust for covariates through the mean function of the response variable, and evaluated the results using the area under the receiver operating characteristic curve (AUC). The AUC is a popular measure for the functional or longitudinal classification (Kohlmann et al., 2009; Kim and Kong, 2016)

These methods use covariates in the classification procedure, however, some of them require the assumption of independent and identically distributed, and some of them are unclear about how to deal with longitudinal data measured at unequally spaced time points. In this study, we propose a new classification method for longitudinal biomarkers with covariate adjustment based on the maximum AUC without the distributional assumptions. Technically, the use of the natural cubic spline is suggested for the representation of longitudinal biomarker profile of each subject, and we use the spline coefficients as the predictors in our proposed classification method. Our method is a two-stage method. Obtaining the optimal combination of the spline coefficients using Stepwise method (Kang et al., 2016) is the first stage, then the largest AUC is achieved in the second stage by including covariates based on the combination of spline coefficients. The classification accuracy of longitudinal biomarkers can be improved significantly after adjusting for covariates using the proposed

method. With the help of the natural cubic spline approach, the dimension of longitudinal biomarker data can be greatly reduced. The proposed method can be implemented easily and it works well even when each subject has a different number of measurements or the measurements are observed at different time points for each subject.

The rest of this chapter is organized as follows. In Section 2, we briefly review the natural cubic spline basis. Our new covariate-adjusted classification method for longitudinal biomarkers is proposed in Section 3. In Section 4, the proposed method is illustrated by using the longitudinal data of Alzheimer's disease and primary biliary cirrhosis data. In Section 5, we evaluate the finite-sample performance of the proposed method by a simulation study. Concluding remarks with discussion are provided in Section 6.

4.2 Natural cubic spline basis

In this section, the natural cubic spline basis is briefly introduced. We use a natural cubic spline basis and spline coefficients for each subject's longitudinal biomarker profile representation.

Suppose that $X_{ij}(t) = \Big(X_{ij}(t_{ij1}), X_{ij}(t_{ij2}), ..., X_{ij}(t_{ijn_{ij}})\Big)$ is the j^{th} observed longitudinal biomarker of the i^{th} subject at the time points $t = (t_{ij1}, t_{ij2}, ..., t_{ijn_{ij}})$, where $i = 1, ..., n$ and $j = 1, ..., J$. In this study, we consider a linear combination of basis functions to represent the longitudinal observations of each subject $X_{ij}(t)$. Let $\{S_\kappa(t)\}_{\kappa=1,...,K}$ be a set of K known basis functions of $X_{ij}(t)$, which satisfies

$$X_{ij}(t) = \sum_{\kappa=1}^{K} S_\kappa(t) c_\kappa = \boldsymbol{S}^{'} \boldsymbol{c},$$

where $\boldsymbol{S}$ is the $n_{ij} \times K$ basis matrix and $\boldsymbol{c} = (c_1, c_2, ..., c_K)^{'}$ is the corresponding vector of basis coefficients.

In practice, we pre-select a basis function and predetermine the number of K before using the spline representation. The number of K is the dimension of the expansion, which is decided according to the characteristics of data. The small K is often used to reduce the

dimension of data for the computational purpose. For the choice of a basis function, (Ramsay et al., 2005, Page 45-56) have discussed several widely used bases in practice. For example, a Fourier basis is often used for the data with the periodic feature. For non-periodic data, a spline basis such as a natural cubic spline appears to be appropriate.

A natural cubic spline (NCS) basis is one of the widely used spline bases in the literature (James et al., 2000; James, 2002; James and Sugar, 2003). An NCS is defined as a function g on $[a, b]$ with the knots sequence as $a < \xi_1 < \xi_2 < ... < \xi_\tau < b$, which satisfies that (1) g is a cubic polynomial on each of the $(\tau + 1)$ intervals $(a, \xi_1), (\xi_1, \xi_2), ..., (\xi_{\tau-1}, \xi_\tau), (\xi_\tau, b)$, and $\xi_1, \xi_2, ..., \xi_\tau$ are interior knots; (2) g is continuous up to the second derivative; (3) g has the second and third derivatives at a and b equal to zero. The last condition is also called the natural boundary constraints, ensuring that g is linear beyond the boundary knots a and b (Green and Silverman, 1993, Page 12).

An NCS with K knots can be generally represented by K basis functions as follows:

$$N_1(x) = 1, N_2(x) = x, ..., N_{\kappa+2}(x) = d_\kappa(x) - d_{K-1}(x),$$

where $d_\kappa(x) = \left[(x - \xi_\kappa)_+^3 - (x - \xi_K)_+^3\right] / (\xi_K - \xi_\kappa)$ for $\kappa = 1, ..., K - 2$, and

$$(x - \xi_\kappa)_+ = \begin{cases} x - \xi_\kappa, & \text{if } x - \xi_\kappa > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Comparing with other cubic splines such as a cubic B-spline (see details at Appendix B), an NCS can generate more stable estimates at the boundaries since it has additional boundary constraints (James et al., 2013, Page 274).

For an NCS, the number of K is typically decided by the number of knots. An NCS with more knots placed produces a more flexible curve, and using the fewer knots would result in a less flexible one. The common way to determine K is to try out the different numbers of knots and choose the appropriate smoothed curve. Another approach is to choose K giving the smallest cross-validated residual sum of squares (James et al., 2013, Page 275).

It is easy to generate an NCS basis matrix using the function "ns" in R package with providing the locations of knots or the degrees of freedom (df). The generated NCS matrix is based on the cubic B-spline basis matrix of data by adding the natural boundary constraints. By specifying the degrees of freedom, it automatically places the same number of knots including the boundary knots at evenly spaced quantiles of data. As the default in the "ns" function, the design matrix of an NCS basis does not include the column of the intercepts, that means (df - 1) interior knots are needed. In contrast, (df - 2) interior knots are required to include the intercepts.

4.3 The proposed method for longitudinal biomarkers

In this section, we present a new classification method for longitudinal biomarkers with covariate adjustment, which is an extension of the proposed method in Chapter 3. Technically, a natural cubic spline basis and its coefficients are used to represent the longitudinal biomarker profile of each subject. The information about biomarkers is contained in the spline coefficients, which can be considered as the predictors in our proposed method. We consider using Stepwise method proposed by Kang et al. (2016) to combine the spline coefficients. Based on the combination of the spline coefficients, covariates are incorporated by achieving the maximum AUC.

Suppose $S_{ij}(t) = \Big(S_{ij}(t_{ij1}), S_{ij}(t_{ij2}), ..., S_{ij}(t_{ijn_{ij}})\Big)'$ is an known NCS basis matrix with the dimension of $n_{ij} \times K$ for $X_{ij}(t)$, then

$$X_{ij}(t) = S_{ij}(t)\boldsymbol{c}_{ij}, i = 1, ..., n, j = 1, ..., J$$

where $\boldsymbol{c}_{ij} = (c_{ij1}, c_{ij2}, ..., c_{ijK})'$ is the unknown vector of the NCS spline coefficients. The common approach to obtain $\boldsymbol{c}_{ij}$ is using the ordinary least squares method. It is clear that the profile of the longitudinal biomarkers $X_{ij}(t)$ can be represented by the NCS coefficients $\boldsymbol{c}_{ij}$, which can be further treated as the predictors in the proposed method. Let M_i be the binary outcome of the i^{th} subject, where $M_i = 1$ and $M_i = 0$ are indicative of disease and health,

respectively. For the total number of the spline coefficients $l = J \times K$, $C_p^D = (C_{p1}^D, C_{p2}^D, ..., C_{pl}^D)$ and $C_q^H = (C_{q1}^H, C_{q2}^H, ..., C_{ql}^H)$ are the sets of l spline coefficients for the p^{th} subject in the disease group and the q^{th} subject in the health group, where $p = 1, ..., n_d$ and $q = 1, ..., n_h$ are the sample sizes of those two groups such that $n_d + n_h = n$. For these multiple spline coefficients, we regard each of them as a single biomarker, and can be combined optimally by adopting Stepwise method proposed by Kang et al. (2016).

We denote the l spline coefficients for the disease and health groups as $C^D = (C_1^D, C_2^D, ..., C_l^D)'$ and $C^H = (C_1^H, C_2^H, ..., C_l^H)'$. Let

$$Y_C^D = \alpha' C^D \quad \text{and} \quad Y_C^H = \alpha' C^H,$$

where $\alpha = (1, \alpha_1, \alpha_2, ..., \alpha_{l-1})'$ is the associated vector of parameters and its estimate can be obtained by Stepwise method. Clearly, all spline coefficients for the disease and health groups are optimally combined in the random variables Y_C^D and Y_C^H. By Equation 3.1, the AUC for the linear combinations of the spline coefficients is

$$AUC_W = P(Y_C^D > Y_C^H).$$

In addition to the spline coefficients, let $Z_p^D = (Z_{p1}^D, Z_{p2}^D, .., Z_{pm}^D)$ and $Z_q^H = (Z_{q1}^H, Z_{q2}^H, ..., Z_{qm}^H)$ stand for m univariate covariates of the p^{th} subject for the disease group and the q^{th} subject for the health group, where $m \geq 1$. Assume that $\beta = (\beta_1, ..., \beta_m)'$ is an unknown vector of coefficients for covariates and γ is an unknown coefficient of the combined spline coefficients, where $\gamma \in R$ and $\beta \in R^m$. We denote the combinations of m covariates for the disease and health groups as $\beta' Z^D$ and $\beta' Z^H$, where $Z^D = (Z_1^D, Z_2^D, ..., Z_m^D)'$ and $Z^H = (Z_1^H, Z_2^H, ..., Z_m^H)'$. Then we conduct the classification by including the linear combination of covariates based on $\gamma' Y_C^D$ and $\gamma' Y_C^H$. By Equation 3.2, the proposed covariate-adjusted AUC is given as

$$AUC^*(\gamma, \beta) = P\Big(Y_C^D + \frac{\beta' Z^D}{\gamma} > Y_C^H + \frac{\beta' Z^H}{\gamma}\Big), \tag{4.1}$$

which naturally adjusts for the covariates in the classification procedure. We obtain the

"best" coefficient parameters (γ_0, β_0) by maximizing the proposed AUC* in Equation 4.1, that denotes as

$$(\gamma_0, \beta_0) = \underset{\gamma \in R,\, \beta \in R^m}{\operatorname{argmax}} \; P\Big(Y_C^D + \frac{\beta' Z^D}{\gamma} > Y_C^H + \frac{\beta' Z^H}{\gamma}\Big).$$

By Theorem 1, we can also show that the classification accuracy is improved after adjusting for covariates by the proposed metod.

The optimal (γ, β) can be estimated empirically as

$$(\hat{\gamma}, \hat{\beta}) = \underset{\gamma \in R,\, \beta \in R^m}{\operatorname{argmax}} \frac{1}{n_d n_h} \sum_{p=1}^{n_d} \sum_{q=1}^{n_h} I\Big(Y_{C_p}^D + \frac{\beta' Z_p^D}{\gamma} > Y_{C_q}^H + \frac{\beta' Z_q^H}{\gamma}\Big),$$

where $I(\cdot)$ is the indicator function.

The proposed covariate-adjusted AUC is estimated as

$$\widehat{AUC}^*(\hat{\gamma}, \hat{\beta}) = \frac{1}{n_d n_h} \sum_{p=1}^{n_d} \sum_{q=1}^{n_h} I\Big(Y_{C_p}^D + \frac{\hat{\beta}' Z_p^D}{\hat{\gamma}} > Y_{C_q}^H + \frac{\hat{\beta}' Z_q^H}{\hat{\gamma}}\Big).$$

It is noteworthy that $(\hat{\gamma}, \hat{\beta})$ is a special case of the maximum rank correlation (MRC) estimator defined in Han (1987), which implies that $(\hat{\gamma}, \hat{\beta})$ is also consistency and asymptotic normality.

4.4 Real data application

In this section, we illustrate the proposed method using the longitudinal data of Alzheimer's disease and primary biliary cirrhosis data.

4.4.1 Alzheimer's disease data

The Alzheimer's disease (AD) data used in this study are from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is a longitudinal study designed to examine multiple types of biomarkers for early detecting and identifying AD. The ADNI study has three phases: ADNI 1, ADNI GO and ADNI 2, in which 1785 participants who were diagnosed repeatedly

every six months for a period up to 144 months from 2004 to 2016. The participants were evaluated with three clinical stages as cognitive normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). In our study, we focus on those participants who were diagnosed as normal (CN) at baseline. The outcome of interest is to identify the patients with MCI or AD. We denote the disease group as $M = 1$ and the health group as $M = 0$. The AD/dementia ($M = 1$) and normal ($M = 0$) groups are defined by whether they stayed at the CN stage until the last visit. For the illustration purpose, the proposed method can be used for the early detection of AD/dementia.

In our study, we consider the Fluorodeoxyglucose positron emission tomography (FDG-PET) and the volume of the hippocampus as the longitudinal biomarkers (i.e., X_1 = FDG-PET, X_2 = hippocampus). In practice, FDG-PET has been used to examine the decreased regional cerebral metabolic rates of glucose, which is an indicator of AD. The regional cerebral hypometabolism often happens at the MCI and AD stages (Biagioni and Galvin, 2011). The hippocampus volume is a primary biomarker with huge potential to detect the presence and progression of AD in many magnetic resonance imaging (MRI) studies. The volume of hippocampus generally declines at the MCI stage and accelerated decreases at the AD stage, which is indicative of AD pathology (Schuff et al., 2009).

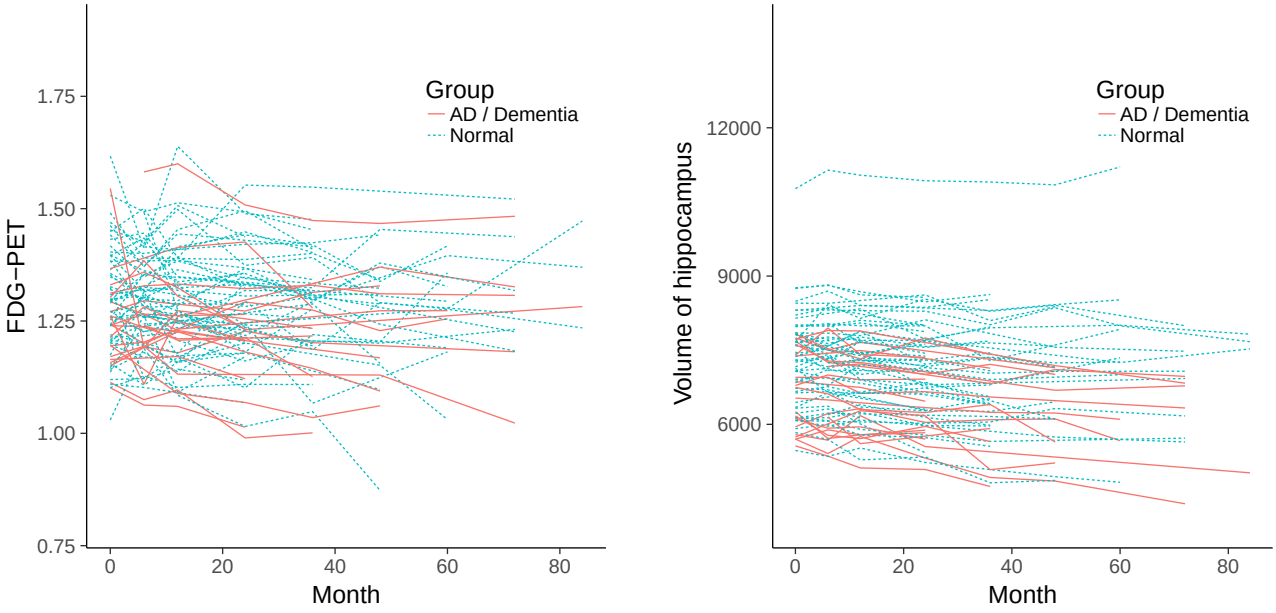
After removing the participants with less than three clinical visits, 23 and 56 participants remain in the AD/dementia and normal groups. The number of total visits for the participants is summarized in Table 4.1. Figure 4.1 (a) and (b) provide the spaghetti plots of FDG-PET and the volume of the hippocampus, suggesting the declining trend of both biomarkers for the AD/dementia participants. Age and gender are regarded as the covariates (i.e., Z_1 = age, Z_2 = gender) in our study, and their descriptive statistics are given in Table 4.2.

A simple classification approach is to apply the classical logistic regression with considering longitudinal biomarkers as multiple predictors. For the raw data of AD, the number of visits for each patient is different from one to the other. Therefore, the classical logistic regression is unable to use for the raw AD data (Zhang et al., 2016). In this study, the proposed method is compared to Stepwise method (Kang et al., 2016) without covariate

Table 4.1: The distribution of visits for Alzheimer's disease data

	Number of participants	
Visits	Normal	AD/Dementia
3	9	2
4	12	7
5	19	8
6	10	4
7	5	2
8	1	0
Total	56	23

Figure 4.1: The spaghetti plots of FDG-PET and the volume of hippocampus



(a) FDG-PET

(b) Volume of hippocampus

adjustment and the functional logistic regression (F-LR) model proposed by James (2002). The functional logistic regression model is the functional generalized linear model (FGLM) using the logit link function. This method smooths each functional profiles by the spline basis functions with the normality assumption on the spline coefficients. The details about

Table 4.2: The descriptive statistics of covariates for Alzheimer's disease data

		Age (years)			Gender(%)	
Group	n	Range	Mean	Std. Dev	Male	Female
AD / Dementia	23	[69.9, 84.8]	77.18	4.12	69.57	30.43
Normal	56	[62.0, 85.8]	74.77	4.88	67.88	32.12

the estimation of the F-LR model can be found in James (2002) (or Appendix C). The overall result of data analysis is reported in Table 4.3. It indicates that the estimated AUCs of the proposed method are slightly larger than those of Stepwise method and the F-LR model after adjusting for age or gender. It is clear that the two covariates for the AD/dementia and normal groups in Table 4.2 have almost no discrepancy. Thus, the two covariates carry little helpful information to the classification of AD, leading to no considerable improvement even after adjusting for both covariates by our method. For adjusting for two covariates, the F-LR model fails to converge.

Table 4.3: The AUCs for Alzheimer's disease data

Without covariate	Covariate-adjusted					
adjustment	Age		Gender		Age + Gender	
Stepwise	Proposed	F-LR	Proposed	F-LR	Proposed	F-LR
0.745	0.761	0.724	0.756	0.717	0.761	–[†]

F-LR: functional logistic regression model;
[†] : the model fails to converge

4.4.2 Primary biliary cirrhosis data

The proposed method is applied to the other longitudinal data, the primary biliary cirrhosis (PBC) data, for illustration. The PBC data came from the clinical trial on the patients with

the chronic liver disease, PBC, conducted by Mayo Clinic between 1974 and 1984 (Murtaugh et al., 1994; Fleming and Harrington, 2011). For each of 312 patients, 20 variables including several biomarkers, age, sex, the number of days between registration and the visit, and the status at the endpoint were contained in the data. In our study, the proposed method is used to detect the survived patients from PBC. The binary outcome is defined based on the patients survived ($M = 0$) or not ($M = 1$) before the end of the study. We select the patients with at least four visits and summarize their visit times in Table 4.5. The data contain 94 and 113 patients in the death and survival groups. The serum bilirubin (bili) and serum albumin (albumin) are considered as the longitudinal biomarkers (i.e., X_1 = bili, X_2 = albumin) because they have been extensively studied (Chan et al., 2015). Their spaghetti plots are presented in Figure 4.2 (a) and (b). The descriptive statistics of two covariates, age and sex (i.e., Z_1 = age, Z_2 = sex), are presented in Table 4.4.

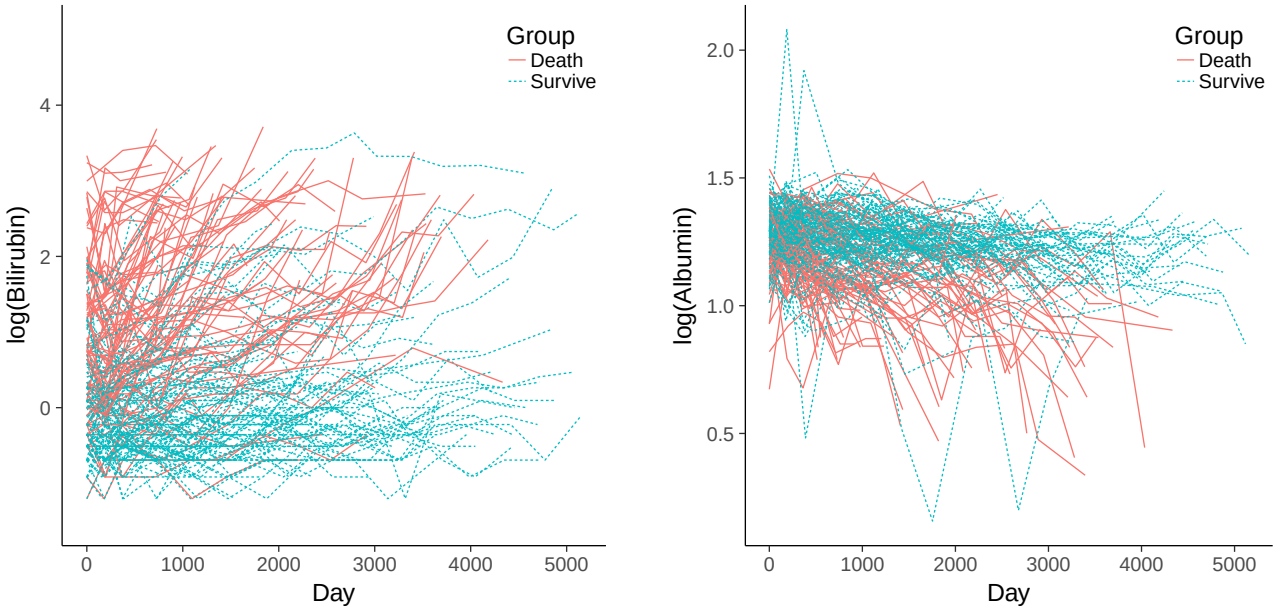
Table 4.4: The descriptive statistics of covariates for primary biliary cirrhosis data

		Age (years)			Sex	
Group	n	Range	Mean	Std. Dev	Male	Female
Death	94	[30.86, 75.01]	52.18	9.97	19 (20.21%)	75 (79.78%)
Survival	113	[28.88, 78.44]	48.67	10.48	7 (6.19%)	106 (93.81%)

Table 4.5: The distribution of visits for primary biliary cirrhosis data

	Number of patients	
Visits	Death	Survival
4	29	9
5	14	12
6	10	10
7	8	11
8	7	13
9	5	10
10	9	13
11	6	9
12	3	9
13	3	3
14	0	5
15	0	6
16	0	3
Total	94	113

Figure 4.2: The spaghetti plots of bilirubin and albumin



(a) bilirubin

(b) albumin

For the comparison purpose, the AUCs estimated by the three methods are shown in Table 4.6. The F-LR model performs almost the same as the proposed method. After adjusting for age, the AUC of the proposed method is slightly larger than that of Stepwise method since the descriptive statistics of age for each group are almost same in Table 4.4. It implies that age provides no additional information for the proposed classification procedure. The percentages of male and female for the dead and survival groups are different, however, the number of survived male is too small. Therefore, the estimated AUCs of the proposed method with the sex adjustment and Stepwise method without covariate adjustment have almost no difference. Even including both covariates by the proposed method has no significant improvement on the AUC, and the F-LR method fails to converge.

Table 4.6: The AUCs for primary biliary cirrhosis data

Without covariate adjustment	Covariate-adjusted					
	Age		Sex		Age + Sex	
Stepwise	Proposed	F-LR	Proposed	F-LR	Proposed	F-LR
0.885	0.903	0.885	0.887	0.875	0.907	$-^{\dagger}$

F-LR: functional logistic regression model
† : the model fails to converge;

4.5 Simulation study

In this section, we conduct an intensive simulation study to assess the finite-sample performance of the proposed method.

We consider two longitudinal biomarkers and one univariate covariate in our simulation study. For both biomarkers, four observations from each subject are measured at the time points $t = (1, 2, 3, 4)$. The performance of the proposed method is evaluated through a comparison of the AUC with the other two approaches: Stepwise method without including covariates and the F-LR model.

4.5.1 Data generated from the functional logistic regression model

In the first data generating scheme, we generate the j^{th} longitudinal biomarkers X_{ij} and the response variable M_i for the i^{th} subject from the F-LR model:

$$X_{ij} = S_{ij}\boldsymbol{c}_{ij} + e_i, \;\; i = 1, ..., n, j = 1, 2,$$

$$M_i = \begin{cases} 1, & \text{if } \left(1 + \exp\{-\omega_0 - \omega_Z{}' Z_i - \sum_{j=1}^{J} \omega_j{}' \boldsymbol{c}_{ij}\}\right)^{-1} > 0.5, \\ 0, & \text{if } \left(1 + \exp\{-\omega_0 - \omega_Z{}' Z_i - \sum_{j=1}^{J} \omega_j{}' \boldsymbol{c}_{ij}\}\right)^{-1} \leq 0.5, \end{cases}$$

where

$$e_i \sim N(0, \sigma_{x_j}^2 I) \text{ and } \boldsymbol{c}_{ij} \sim N(\mu_j, \Gamma_j).$$

Here ωs are the model coefficients and Z_i is a univariate covariate of the i^{th} subject. In our simulation, S_{ij} are the NCS basis matrices of a follow up time $t = (1, 2, 3, 4)$ and $Z_i \sim \text{Exp}(0.5)$. Let $\sigma^2_{x_1} = 1$, $\sigma^2_{x_2} = 2$, $\omega_0 = -1$, $\omega_Z = 0.3$, $\omega_1 = (0.15, 0.25)'$, $\omega_2 = (0.1, 0.2)'$, $\mu_1 = (0.2, 0.4)'$, $\mu_2 = (0.3, 0.5)'$, $\Gamma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ and $\Gamma_2 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}$. We simulate 200 Monte Carlo samples for the sample sizes from 20 to 60 for the health group and 20 for the disease group.

The estimated AUCs and their standard errors (SEs) of the three methods are summarized in Table 4.7. The results suggest that the F-LR model outperforms the other two methods since the data is generated from the framework of the F-LR model. As expected, the estimated AUCs of the proposed method are just slightly smaller than those of the F-LR model. The performance of the proposed method is still better than Stepwise method which is without covariate adjustment, showing that the classification accuracy is significantly improved after adjusting for covariates using the proposed method.

Table 4.7: The estimated AUCs and SEs for data generated from the functional logistic regression model with 200 Monte Carlo samples

Sample size	Without covariate adjustment	Covariate-adjusted	
(n_d, n_h)	Stepwise	F-LR	Proposed
(20, 20)	0.752	0.943	0.914
	(0.073)	(0.048)	(0.046)
(20, 40)	0.724	0.941	0.913
	(0.061)	(0.026)	(0.036)
(20, 60)	0.721	0.938	0.909
	(0.054)	(0.025)	(0.036)

F-LR: functional logistic regression model

4.5.2 Data simulated from different distributions

Under the second data generating scheme, we generate the first longitudinal biomarkers with four-time points for the health and disease groups from the multivariate normal distribution as follows:

$$X_1^H \sim N\left(\begin{bmatrix}2\\2\\2\\2\end{bmatrix}, \begin{bmatrix}1 & 0.5 & 0.5 & 0.5\\0.5 & 1 & 0.5 & 0.5\\0.5 & 0.5 & 1 & 0.5\\0.5 & 0.5 & 0.5 & 1\end{bmatrix}\right), X_1^D \sim N\left(\mu_1^D, \begin{bmatrix}1 & 0.5 & 0.5 & 0.5\\0.5 & 1 & 0.5 & 0.5\\0.5 & 0.5 & 1 & 0.5\\0.5 & 0.5 & 0.5 & 1\end{bmatrix}\right),$$

where $\mu_1^D = \{2 + 0.25 \times t - 0.1 \times t^2\}_{t=1,2,3,4}$. Another biomarker for the health and disease groups are generated from the multivariate exponential distribution (abbreviated MVE) as $X_2^H \sim \text{MVE}(1.5, 1.5, 1.5, 1.5)$ and $X_2^D \sim \text{MVE}(\lambda_2^D)$, where $\lambda_2^D = \{0.4 + 0.25 \times t + 0.1 \times t^2\}_{t=1,2,3,4}$. In addition to the biomarkers, we generate an exponential distributed covariate for the disease and health groups as $Z^D \sim \text{Exp}(0.5)$ and $Z^H \sim \text{Exp}(2)$. Two hundred Monte Carlo samples are simulated for the sample sizes from 20 to 60.

The AUCs with their SEs given in Table 4.8 indicate that the obtained AUCs of the proposed method appear to be the largest, and those of Stepwise method are the smallest. It implies that including the covariates by the proposed model improves the classification performance dramatically.

Table 4.8: The estimated AUCs and SEs for two biomarkers and one covariate with 200 Monte Carlo samples

Sample size	Without covariate adjustment	Covariate-adjusted	
(n_d, n_h)	Stepwise	F-LR	Proposed
(20, 20)	0.723	0.743	0.864
	(0.075)	(0.109)	(0.061)
(40, 40)	0.700	0.738	0.854
	(0.057)	(0.073)	(0.045)
(60, 60)	0.691	0.726	0.850
	(0.049)	(0.059)	(0.038)

F-LR: functional logistic regression model

4.5.3 Simulation for consistency

An additional simulation is conducted to study the consistency of the proposed parameter estimators. With the two longitudinal biomarkers from the second data generating scheme, four covariates are generated from the following different distributions with equal means and variances for the disease and health groups,

$$Z_1^D, Z_1^H \sim N(0, 1.5^2),$$
$$Z_2^D, Z_2^H \sim \text{Exp}(1),$$
$$Z_3^D, Z_3^H \sim \text{Bin}(0.2),$$
$$Z_4^D, Z_4^H \sim \text{Pois}(2).$$

We use 100, 200, 500 and 1000 as the sample sizes of each group with conducting 1000 Monte Carlo simulations. It is clear that those covariates are useless to the classification, implying that the true values of βs in our method should be 0.

The estimates and their SEs of the parameters are provided in Table 4.9. The results indicate that the βs are approximately close to 0 and the estimates of γ are close to 1. The

associated SEs decrease as increasing the sample sizes, showing the evidence of consistency.

Overall, the proposed method can make a massive improvement on the classification accuracy after adjusting for covariates. The functional logistic regression model also performs well, however, the computation time is significantly expensive.

Table 4.9: The estimates (S.E.) for parameters that associated with four irrelevant covariates with 1000 Monte Carlo samples

Sample size $n_d = n_h$	$\hat{\gamma}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
100	0.733	0.013	0.017	0.036	0.017
	(0.584)	(0.088)	(0.146)	(0.287)	(0.094)
200	0.886	0.002	0.001	0.018	0.001
	(0.390)	(0.065)	(0.099)	(0.207)	(0.072)
500	0.981	0.002	-0.001	0.003	0.0009
	(0.091)	(0.041)	(0.068)	(0.148)	(0.044)
1000	0.990	0.001	0.001	0.003	0.0005
	(0.013)	(0.034)	(0.050)	(0.119)	(0.034)

4.6 Discussion

The proposed method is an easy-implemented approach to improve the classification performance of longitudinal biomarkers after adjusting for covariates. In addition, no assumption on the distributions of biomarkers or covariates is required for the proposed method. Our methodology works well particularly when each subject is measured at different time points or has a different number of measurements.

When the discrepancies of the covariates between the disease and health groups are significant, the proposed method with such covariate adjustment can achieve a considerable improvement on the classification performance. The proposed method employs the technique

of the natural cubic spline for reducing the dimension of longitudinal biomarker data.

From a simulation (not shown) using a natural cubic spline basis with different degrees of freedom (i.e., df = 4, 3, 2), the results show that the AUCs are not significantly different. Therefore, we suggest the natural cubic spline basis with two degrees of freedom (df = 2) in our method to achieve the maximum data dimension reduction.

In general, intercepts can be included or excluded in a natural cubic spline basis, and the default is without intercepts in the R package. The result of another simulation (not shown) shows no impact on the classification accuracy with intercepts in a natural cubic spline basis or not. Other basis functions such as a Fourier basis and orthogonal polynomial bases could also work for the proposed method to deal with different types of data, such as periodic data.

Chapter 5

Conclusions and future work

In this study, a new covariate-adjusted classification method for cross-sectional biomarkers has been proposed. It is an easy-implemented approach, which requires no distributional assumption on biomarkers and covariates. The classification performance can be improved remarkably after adjusting for covariates using the proposed method. Our method works well for discrete or continuous covariates. Even including irrelevant covariates does not jeopardize the classification accuracy of biomarkers under the proposed method in large sample sizes. In practice, we suggest conducting an initial screening manually to remove the irrelevant covariates for improving the computational efficiency.

We also extend the proposed method to longitudinal biomarkers. The considerable improvements on classification accuracy of longitudinal biomarkers can be achieved with covariate adjustment as well. Our extended method works well even when each subject has an unequal number of measurements or the measurements are not observed at the same time points for each subject. The proposed method is robust to the assumptions on the biomarkers and covariates distributions, and it is easy to perform. Using the technique of the natural cubic spline, the proposed method can significantly reduce the dimension of longitudinal biomarker data with little loss in the classification performance.

In practice, the issue of dropouts often exists in longitudinal studies. The missing data due to dropouts may affect the classification performance. Computationally, our extended

method can still work for the situation of late dropouts as long as it is non-informative, but a loss of classification accuracy is expected. It is unclear about the performance of the proposed method under other types of missing data mechanism, and a further investigation could be our future work.

Bibliography

Adni study design. http://http://www.adni.loni.usc.edu/study-design/#background-container.

Andrés M Alonso, David Casado, and Juan Romo. Supervised classification for functional data: A weighted distance approach. *Computational Statistics & Data Analysis*, 56(7): 2334–2346, 2012.

Todd A Alonzo and Margaret Sullivan Pepe. Distribution-free roc analysis using binary regression techniques. *Biostatistics*, 3(3):421–432, 2002.

Alzheimer's Association et al. 2012 alzheimers disease facts and figures. *Alzheimer's & Dementia*, 8(2):131–168, 2012.

SC Bagui and KL Mehra. Classification of multiple observations using multi-stage rank nearest neighbor rule. *Journal of statistical planning and inference*, 76(1-2):163–183, 1999.

Stuart G Baker. The central role of receiver operating characteristic (roc) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute*, 95(7): 511–515, 2003.

Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415, 1975.

Milton C Biagioni and James E Galvin. Using biomarkers to improve detection of alzheimers disease. 2011.

Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.

Tianxi Cai and Margaret Sullivan Pepe. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American statistical Association*, 97(460):1099–1107, 2002.

Hervé Cardot. Conditional functional principal components analysis. *Scandinavian journal of statistics*, 34(2):317–335, 2007.

William J Catalona, Deborah S Smith, Timothy L Ratliff, Kathy M Dodds, Douglas E Coplen, Jerry JJ Yuan, John A Petros, and Gerald L Andriole. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *New England Journal of Medicine*, 324(17):1156–1161, 1991.

Anthony WH Chan, Ronald CK Chan, Grace LH Wong, Vincent WS Wong, Paul CL Choi, Henry LY Chan, and Ka-Fai To. New simple prognostic score for primary biliary cirrhosis: Albumin-bilirubin score. *Journal of gastroenterology and hepatology*, 30(9):1391–1396, 2015.

Jeng-Min Chiou, Hans-Georg Müller, and Jane-Ling Wang. Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):405–423, 2003.

Alaattin Erkanli, Refik Soyer, and Adrian Angold. Bayesian analyses of longitudinal binary data using markov regression models of unknown order. *Statistics in medicine*, 20(5): 755–770, 2001.

Ruth Etzioni, Charles Kooperberg, Margaret Pepe, Robert Smith, and Peter H Gann. Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics*, 4 (4):523–538, 2003.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

Peter Hall, Donald S Poskitt, and Brett Presnell. A functional dataanalytic approach to signal discrimination. *Technometrics*, 43(1):1–9, 2001.

Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Gareth M James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432, 2002.

Gareth M James and Trevor J Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550, 2001.

Gareth M James and Catherine A Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408, 2003.

Gareth M James, Trevor J Hastie, and Catherine A Sugar. Principal component models for sparse functional data. *Biometrika*, 87(3):587–602, 2000.

Holly Janes and Margaret S Pepe. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *American Journal of Epidemiology*, 168(1):89–97, 2008.

Holly Janes and Margaret S Pepe. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*, 96(2): 371–382, 2009.

Madumali Kalubowilage. *Liquid biopsies of solid tumors: non-small-cell lung and pancreatic cancer*. PhD thesis, Kansas State University, 2017.

Le Kang, Aiyi Liu, and Lili Tian. Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Statistical methods in medical research*, 25 (4):1359–1380, 2016.

Soyoung Kim and Ying Huang. Combining biomarkers for classification with covariate adjustment. *Statistics in Medicine*, 2017.

Yeonhee Kim and Lan Kong. Classification using longitudinal trajectory of biomarker in the presence of detection limits. *Statistical methods in medical research*, 25(1):458–471, 2016.

Mareike Kohlmann, Leonhard Held, and Veit Peter Grunert. Classification of therapy resistance based on longitudinal biomarker profiles. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(4):610–626, 2009.

Michael D Lebowitz. Age, period and cohort effects. *Am J Respir Crit Care Med*, 154(6): S273–S7, 1996.

Xiaoyan Leng and Hans-Georg Müller. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22(1):68–76, 2005.

Pai-Ling Li, Jeng-Min Chiou, and Yu Shyr. Functional data classification using covariate-adjusted subspace projection. *Computational Statistics & Data Analysis*, 2017.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Chunling Liu, Aiyi Liu, and Susan Halabi. A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 30(16):2005–2014, 2011.

Guillermo Marshall and Anna E Barón. Linear discriminant models for unbalanced longitudinal data. *Statistics in medicine*, 19(15):1969–1981, 2000.

Rachel Mistur, Lisa Mosconi, Susan De Santi, Marla Guzman, Yi Li, Wai Tsui, and Mony J de Leon. Current challenges for the early detection of alzheimer's disease: brain imaging and csf studies. *Journal of clinical neurology*, 5(4):153–166, 2009.

Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. *Annals of Statistics*, pages 774–805, 2005.

Paul A Murtaugh, E Rolland Dickson, Gooitzen M Van Dam, Michael Malinchoc, Patricia M Grambsch, Alice L Langworthy, and Chris H Gips. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, 20(1):126–134, 1994.

Deborah Nolan, David Pollard, et al. u-processes: Rates of convergence. *The Annals of Statistics*, 15(2):780–799, 1987.

Margaret Sullivan Pepe. An interpretation for the roc curve and inference using glm procedures. *Biometrics*, 56(2):352–359, 2000.

Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

Margaret Sullivan Pepe and Mary Lou Thompson. Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2):123–140, 2000.

Margaret Sullivan Pepe, Tianxi Cai, and Gary Longton. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1): 221–229, 2006.

Jim Ramsay, James Ramsay, and BW Silverman. *Functional Data Analysis*. Springer Science & Business Media, 2005.

John A Rice. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, pages 631–647, 2004.

Fabrice Rossi and Nathalie Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7):730–742, 2006.

John W Rowe, Reubin Andres, Jordan D Tobin, Arthur H Norris, and Nathan W Shock. The effect of age on creatinine clearance in men: a cross-sectional and longitudinal study. *Journal of gerontology*, 31(2):155–163, 1976.

Enrique F Schisterman, David Faraggi, and Benjamin Reiser. Adjusting the generalized roc curve for covariates. *Statistics in Medicine*, 23(21):3319–3331, 2004.

N Schuff, N Woerner, L Boreta, T Kornfield, LM Shaw, JQ Trojanowski, PM Thompson, CR Jack Jr, MW Weiner, and Alzheimer's; Disease Neuroimaging Initiative. Mri of hippocampal volume loss in early alzheimer's disease in relation to apoe genotype and biomarkers. *Brain*, 132(4):1067–1077, 2009.

Robert P Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, pages 123–137, 1993.

John Q Su and Jun S Liu. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355, 1993.

John A Swets. Indices of discrimination or diagnostic accuracy: Their rocs and implied models. *Psychological bulletin*, 99(1):100, 1986.

Lisa Tomasko, Ronald W Helms, and Steven M Snapinn. A discriminant analysis extension to mixed models. *Statistics in medicine*, 18(10):1249–1260, 1999.

Dinusha N Udukala, Hongwang Wang, Sebastian O Wendel, Aruni P Malalasekera, Thilani N Samarakoon, Asanka S Yapa, Gayani Abayaweera, Matthew T Basel, Pamela Maynez, Raquel Ortega, et al. Early breast cancer screening using iron/iron oxide-based nanoplatforms with sub-femtomolar limits of detection. *Beilstein journal of nanotechnology*, 7:364, 2016.

Scott L Zeger, Kung-Yee Liang, and Steven G Self. The analysis of binary longitudinal data with time independent covariates. *Biometrika*, 72(1):31–38, 1985.

Xin Zhang, Daniel R Jeske, Jun Li, and Vance Wong. A sequential logistic regression classifier based on mixed effects with applications to longitudinal data. *Computational Statistics & Data Analysis*, 94:238–249, 2016.

Yanyu Zhang and Jialiang Li. Combining multiple markers for multi-category classification: An roc surface approach. *Australian & New Zealand Journal of Statistics*, 53(1):63–78, 2011.

Xin Zhao, JS Marron, and Martin T Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, pages 789–808, 2004.

Mark H Zweig and Gregory Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

Appendix A

Proofs of Theorem 2 and 3

A.1 Proof of Theorem 2

The AUC$^*(\theta)$ of the proposed method defined in Equation 3.2 is equivalent to

$$AUC_n^*(\theta) = \frac{1}{n(n-1)} \sum_{p \neq q} \Big\{ I(D_p > D_q) I(X_p{}' \theta > X_q{}' \theta) \Big\},$$

where I is the indicator function. The proposed parameter estimator is given as

$$\hat{\theta} = (\hat{\gamma}, \hat{\beta}) = \underset{\theta}{\operatorname{argmax}}\, AUC_n^*(\theta).$$

The maximum rank correlation (MRC) estimator defined in Han (1987) is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \big(S_n(\theta) \big),$$

where

$$S_n(\theta) = \frac{1}{n(n-1)} \sum_{p \neq q} \Big\{ I(D_p > D_q) I(X_p{}' \theta > X_q{}' \theta) + I(D_p < D_q) I(X_p{}' \theta < X_q{}' \theta) \Big\}.$$

For a binary outcome $D = \{0, 1\}$, if $D_p > D_q$, then $D_p = 1$ indicates that the p^{th} subject belongs to the disease group and $D_q = 0$ indicates that the q^{th} subject belongs to the health group. The inequality $X_p{}'\theta > X_q{}'\theta$ represents the value of biomarkers and covariates combination for the p^{th} subject in the disease group is larger than that for the q^{th} subject in the health group. If $D_p < D_q$, then $D_p = 0$ shows that the p^{th} subject belongs to the health group and $D_q = 1$ shows that the q^{th} subject belongs to the disease group. We use the inequality $X_p{}'\theta < X_q{}'\theta$ to represent the value of biomarkers and covariates combination for the p^{th} subject in the health group is smaller than that for the q^{th} subject in the disease group. Thus, only considering the direction of the inequality $X_p{}'\theta > X_q{}'\theta$,

$$S_n(\theta) = \frac{1}{n(n-1)} \sum_{p \neq q} \Big\{ I(D_p > D_q) I(X_p{}'\theta > X_q{}'\theta) \Big\} = \frac{1}{n(n-1)} \sum_{p \neq q} \Big\{ I(X_p{}'\theta > X_q{}'\theta) \Big\}$$

is actually the Mann-Whitney U statistic of $AUC^*(\theta) = P(X_p{}'\theta > X_q{}'\theta)$, implying the proposed parameter estimator is a special case of the MRC estimator.

Therefore, the consistency of the parameter estimators in the proposed method can be derived base on the same proof of the MRC estimator shown in the following three steps by Han (1987). The parameters are normalized as $\theta^* = \theta/\|\theta\|$, where $\|\cdot\|$ is a matrix norm. The first step is to show the uniqueness that the true value θ_0^* is the unique maximizer of $E\Big(S_n(\theta^*)\Big)$. Next, the uniformly convergence of θ^* is shown by using the Borel-Cantelli lemma. In the final step, we can show that

$$\hat{\theta}^* \xrightarrow{a.s.} \theta_0^*.$$

A.2 Proof of Theorem 3

The proof of Theorem 3 follows the same steps to show the asymptotic normality of the MRC estimator in Sherman (1993). Let $v_1 = (d_1, x_1)$ and $v_2 = (d_2, x_2)$. For each (v_1, v_2), define

$$f(v_1, v_2, \theta^*) = I(d_1 > d_2) \Big\{ I(x_1{}'\theta^* > x_2{}'\theta^*) - I(x_1{}'\theta_0^* > x_2{}'\theta_0^*) \Big\}$$

and a U-statistics

$$\Gamma_n = AUC_n(\theta^*) - AUC_n(\theta_0^*) = \frac{1}{n(n-1)} \sum \Big\{ I(d_1 > d_2) \Big[I(x_1{}'\theta^* > x_2{}'\theta^*) - I(x_1{}'\theta_0^* > x_2{}'\theta_0^*) \Big] \Big\}.$$

For the U-statistics, let the empirical measure be P_n and random measure be U_n. By the decomposition of a U-statistics of order two, it can be obtained that

$$\Gamma_n(\theta^*) = \Gamma(\theta^*) + P_n g(\cdot, \theta^*) + U_n h(\cdot, \cdot, \theta^*), \tag{A.1}$$

where

$$g(\cdot, \theta^*) = Pf(v, \cdot, \theta^*) + Pf(\cdot, v, \theta^*) - 2\Gamma(\theta^*),$$
$$h(v_1, v_2, \theta^*) = f(v_1, v_2, \theta^*) - Pf(v_1, \cdot, \theta^*) - Pf(\cdot, v_2, \theta^*) + \Gamma(\theta^*).$$

Specifically, $Pf(v_1, \cdot, \theta^*)$ denotes the conditional expectation of $f(v_1, v_2, \theta^*)$ given v_1, $Pf(\cdot, v_2, \theta^*)$ denotes the conditional expectation of $f(v_1, v_2, \theta^*)$ given v_2.

For all $v \in S$, by a Taylor expansion of $\varphi(v, \theta^*)$ about θ_0^*, *Assumption 3* (i)(ii) and $2\Gamma(\theta^*) = \Big(\varphi(\cdot, \theta^*) - \varphi(\cdot, \theta_0^*)\Big)$, we can show that

$$2\Gamma(\theta^*) = (\theta^* - \theta_0^*)' \Lambda (\theta^* - \theta_0^*) + o\Big(|\theta^* - \theta_0^*|^2\Big) \text{ as } \theta^* \to \theta_0^*, \tag{A.2}$$

where $\Lambda = E\Big\{ \frac{\partial}{\partial \theta_0^{*2}} \varphi(v, \theta_0^*) \Big\}$ for all $v \in S$.

By $E\frac{\partial}{\partial \theta_0^*} \varphi(v, \theta_0^*) = 0$ and *Assumption 3* (iii), then $W_n \xrightarrow{D} N(0, \Sigma)$, where $\Sigma = E\Bigg\{ \Big[\frac{\partial}{\partial \theta_0^*} \varphi(v, \theta_0^*) \Big] \Big[\frac{\partial}{\partial \theta_0^*} \varphi(v, \theta_0^*) \Big]' \Bigg\}$ for all $v \in S$. Since $g(v, \theta^*) = \varphi(v, \theta^*) - \varphi(v, \theta_0^*) - 2\Gamma(\theta^*)$, by A.1 and the weak law of large numbers, we can show that

$$P_n g(\cdot, \theta^*) = \frac{1}{\sqrt{n}} (\theta^* - \theta_0^*)' W_n + o\Big(|\theta^* - \theta_0^*|^2\Big). \tag{A.3}$$

Next, by Theorem 3 in Sherman (1993) and the Euclidean properties of Nolan et al. (1987), then

$$U_n h(\cdot, \cdot, \theta^*) = o_p(1/n). \tag{A.4}$$

By A.2, A.3, A.4 and *Assumption 3* (iii), we can show that

$$\Gamma_n(\theta^*) = \frac{1}{2}(\theta^* - \theta_0^*)' \Lambda (\theta^* - \theta_0^*) + \frac{1}{\sqrt{n}}(\theta^* - \theta_0^*)' W_n + o_p(1/n),$$

which implies that

$$\sqrt{n}(\hat{\theta}^* - \theta_0^*) \xrightarrow{D} N(0, \Lambda^{-1} \Sigma \Lambda^{-1}).$$

Appendix B

A brief introduction of B-spline

Suppose a sequence of knots $\{\xi_1 < \xi_2, ..., < \xi_{\tau-1} < \xi_\tau\}$ is on $[a, b]$. Additional R knots are below or above the boundaries such that $\psi_1 \leq \psi_2 \leq ... \leq \psi_R \leq a, b \leq \psi_{t+R+1} \leq \psi_{t+R+2} \leq ... \leq \psi_{t+2R}$, where $R = 1, 2, 3, ...$ It is often to make the extra knots equal to the two boundary respectively (i.e. $\psi_1 = \psi_2 = ... = \psi_R = a, \psi_{\tau+R+1} = \psi_{\tau+R+2} = ... = \psi_{\tau+2R} = b$). Let $\psi_{e+r} = \xi_e$, where $e = 1, ..., \tau$. It is defined that a B-spline with order r is a piecewise polynomial function of degree $(r-1)$ on $[a, b]$, where $r = 1, 2, 3, ... \leq R$. We denote the L^{th} basis function of a r^{th} order B-spline as $B_{L,\,r}(x)$. For the order $r = 1$, the B-spline basis function is given by

$$B_{L,\,1}(x) = \begin{cases} 1, & \text{if } \psi_L \leq x < \psi_{L+1}, \\ 0, & \text{otherwise,} \end{cases}$$

where $L = 1, ..., \tau + 2R - 1$. The higher order basis functions are provided as

$$B_{L,\,r}(x) = \frac{x - \psi_L}{\psi_{L+r-1} - \psi_L} B_{L,\,r-1}(x) + \frac{\psi_{L+r} - x}{\psi_{L+r} - \psi_{L+1}} B_{L+1,\,r-1}(x),$$

for $L = 1, ..., \tau + 2R - r$ (Friedman et al., 2001, page 186-187). A B-spline has the property that it is continuous up to the $(r-1)^{th}$ derivative but discontinuous at the boundaries a and b. Besides, a linear combination of the r^{th} order B-splines can express any spline function of the same order.

Appendix C

James' (2002) functional logistic regression algorithm

A Monte Carlo simulation is used to compute $\hat{c}_{ij}$ and $V_{c_{ij}}$ because of no closed form solution for them. A sample of $c^*_{ij_1}, ..., c^*_{ij_N}$ for the j^{th} biomarker of the i^{th} subject is generated from the distribution

$$c_{ij}|x_{ij} \sim N\Big((\sigma^2_{x_j}\Gamma_j^{-1} + S_{ij}{}'S_{ij})^{-1}(\sigma^2_{x_j}\Gamma_j^{-1}\mu_j + S_{ij}{}'x_{ij}), (\Gamma_j^{-1} + S_{ij}{}'S_{ij}/\sigma^2_{x_j})^{-1}\Big),$$

where $i = 1, ..., n$, $j = 1, ..., J$, and other parameters are given in Section 4.5.1. Then $\hat{c}_{ij}$ and $V_{c_{ij}}$ are unbiased estimated by

$$\hat{c}_{ij} = \frac{\sum_{u=1}^{N} c^*_{iju} P(M = M_i|c^*_{iju})}{\sum_{u=1}^{N} P(M = M_i|c^*_{iju})}, \tag{C.1}$$

$$V_{c_{ij}} = \frac{\sum_{u=1}^{N} c^*_{iju} c^*_{iju}{}' P(M = M_i|c^*_{iju})}{\sum_{u=1}^{N} P(M = M_i|c^*_{iju})} - \hat{c}_{ij}\hat{c}'_{ij}, \tag{C.2}$$

where

$$P(M = M_i|c^*_{iju}) = \begin{cases} \left(1 + \exp(-\omega_0 - \omega_z{}' z_i - \sum_{j=1}^{J} \omega_j{}' c^*_{iju})\right)^{-1}, & \text{for } M_i = 1, \\ \left(1 + \exp(\omega_0 + \omega_z{}' z_i + \sum_{j=1}^{J} \omega_j{}' c^*_{iju})\right)^{-1}, & \text{for } M_i = 0, \end{cases}$$

and M_i are the outcome of the i^{th} subject. Therefore, μ_j, Γ_j and $\sigma^2_{x_j}$ can be estimated as

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{c}_{ij}, \tag{C.3}$$

$$\hat{\Gamma}_j = \frac{1}{n} \sum_{i=1}^{n} \left\{ V_{c_{ij}} + (\hat{c}_{ij} - \hat{\mu}_j)(\hat{c}_{ij} - \hat{\mu}_j)' \right\}, \tag{C.4}$$

$$\hat{\sigma}^2_{x_j} = \frac{1}{\sum n_{ij}} \sum_{i=1}^{n} \left\{ (x_{ij} - S_{ij}\hat{c}_{ij})'(x_{ij} - S_{ij}\hat{c}_{ij}) + tr(S_{ij} V_{c_{ij}} S_{ij}{}') \right\}, \tag{C.5}$$

where n_{ij} is the number of time points for the j^{th} biomarker of the i^{th} subject. ω_0, ω_z and ω_j are estimated by

$$E(A'WA) \begin{pmatrix} \omega_0 \\ \omega_z \\ \omega_j \end{pmatrix} = E(A'WF),$$

where A is an $n \times (K + 2)$ matrix with the i^{th} row of $(1, z_i, c_{ij}{}')$, W is an $n \times n$ diagonal matrix with the i^{th} diagonal of $\pi_i(1 - \pi_i)$, F is a vector with the i^{th} element of $\omega_0 + \omega_z{}' z_i + \sum_{j=1}^{J} \omega_j{}' c_{ij} + \frac{M_i - \pi_i}{\pi_i(1-\pi_i)}$ and $\pi_i = P(M_i = 1|c_{ij})$. Therefore,

$$\begin{pmatrix} \hat{\omega}_0 \\ \hat{\omega}_z \\ \hat{\omega}_j \end{pmatrix} = E(A^{'}WA)^{-1}E(A^{'}WF)$$

$$= \left[\sum_i E \begin{pmatrix} \pi_i(1-\pi_i) & \pi_i(1-\pi_i)z_i^{'} & \pi_i(1-\pi_i)c_{ij}^{'} \\ \pi_i(1-\pi_i)z_i & \pi_i(1-\pi_i)z_i z_i^{'} & \pi_i(1-\pi_i)z_i c_{ij}^{'} \\ \pi_i(1-\pi_i)c_{ij} & \pi_i(1-\pi_i)c_{ij}z_i^{'} & \pi_i(1-\pi_i)c_{ij}c_{ij}^{'} \end{pmatrix} \right]^{-1}$$

$$\times E \begin{bmatrix} \sum_i (M_i - \pi_i) \\ \sum_i (M_i - \pi_i)z_i \\ \sum_i (M_i - \pi_i)c_{ij}. \end{bmatrix} \quad \text{(C.6)}$$

To complete the calculation of $\hat{\omega}_0, \hat{\omega}_z$ and $\hat{\omega}_j$, the estimates for the elements in (C.6) are given by

$$\frac{\sum_{u=1}^{N} \Omega_u P(M = M_i | c^*_{iju})}{\sum_{u=1}^{N} P(M = M_i | c^*_{iju})}, \quad \text{(C.7)}$$

where

$$\Omega_u = \begin{cases} P(M=1|c^*_{iju})\Big\{1-P(M=1|c^*_{iju})\Big\} & \text{for } E\Big\{\pi_i(1-\pi_i)\Big\}, \\ z_i P(M=1|c^*_{iju})\Big\{1-P(M=1|c^*_{iju})\Big\} & \text{for } E\Big\{\pi_i(1-\pi_i)z_i\Big\}, \\ z_i z_i{}' P(M=1|c^*_{iju})\Big\{1-P(M=1|c^*_{iju})\Big\} & \text{for } E\Big\{\pi_i(1-\pi_i)z_i z_i{}'\Big\}, \\ c^*_{iju} P(M=1|c^*_{iju})\Big\{1-P(M=1|c^*_{iju})\Big\} & \text{for } E\Big\{\pi_i(1-\pi_i)c_{ij}\Big\}, \\ z_i c^*_{iju}{}' P(M=1|c^*_{iju})\Big\{1-P(M=1|c^*_{iju})\Big\} & \text{for } E\Big\{\pi_i(1-\pi_i)z_i c_{ij}{}'\Big\}, \\ c^*_{iju} c^*_{iju}{}' P(M=1|c^*_{iju})\Big\{1-P(M=1|c^*_{iju})\Big\} & \text{for } E\Big\{\pi_i(1-\pi_i)c_{ij} c_{ij}{}'\Big\}, \\ M_i - P(M=1|c^*_{iju}) & \text{for } E(M_i-\pi_i), \\ z_i\Big\{M_i - P(M=1|c^*_{iju})\Big\} & \text{for } E\Big\{z_i(M_i-\pi_i)\Big\}, \\ c^*_{iju}\Big\{M_i - P(M=1|c^*_{iju})\Big\} & \text{for } E\Big\{c_{ij}(M_i-\pi_i)\Big\}. \end{cases}$$

The EM algorithm of the functional logistic regression repeats the three following steps until the convergence of the parameters.

Step 1 (E-step) : calculate the expected values and variance of the $c_{ij}s$ using (C.1), (C.2) and the values of $E\Big\{\pi_i(1-\pi_i)\Big\}$, $E\Big\{\pi_i(1-\pi_i)z_i\Big\}$, $E\Big\{\pi_i(1-\pi_i)z_i z_i{}'\Big\}$, $E\Big\{\pi_i(1-\pi_i)c_{ij}\Big\}$, $E\Big\{\pi_i(1-\pi_i)z_i c_{ij}{}'\Big\}$, $E\Big\{\pi_i(1-\pi_i)c_{ij}c_{ij}{}'\Big\}$, $E(M_i-\pi_i)$, $E\Big\{z_i(M_i-\pi_i)\Big\}$, $E\Big\{c_{ij}(M_i-\pi_i)\Big\}$ using (C.7).

Step 2 (M-step) : estimate the parameters $\sigma^2_{x_j}$, μ_j, Γ_j, ω_0, ω_z and ω_j using (C.3), (C.4), (C.5) and (C.6) respectively.

Step 3: repeat step 1 and 2 until the parameters have converged.