Efficient classification for longitudinal data

Xianlong Wang[a,*], Annie Qu[b]

[a] *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA*
[b] *Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, 61820, USA*

Abstract

A new classifier, *QIFC*, is proposed based on the quadratic inference function for longitudinal data. Our approach builds a classifier by taking advantage of modeling information between the longitudinal responses and covariates for each class, and assigns a new subject to the class with the shortest newly defined distance to the subject. For finite sample applications, this enables one to overcome the difficulty in estimating covariance matrices while still incorporating correlation into the classifier. The proposed classifier only requires the first moment condition of the model distribution, and hence is able to handle both continuous and discrete responses. Simulation studies show that *QIFC* outperforms competing classifiers, such as the functional data classifier, support vector machine, logistic regression, linear discriminant analysis, the naive Bayes classifier and the decision tree in various practical settings. Two time-course gene expression data sets are used to assess the performance of *QIFC* in applications.

Keywords: QIFC, Classification, Linear Discriminant Analysis, Longitudinal Data Analysis, Quadratic Inference Function, Quadratic Distance

1. Introduction

In many longitudinal biomedical experiments, such as the gene expression microarray studies on yeast cells [19, 6] and fruit flies [1, 12], the gene expressions of thousands of genes are repeatedly measured over multiple time-points. These genes are assumed to be associated with a set of pre-defined biological functions, and it is of scientific interest to identify which genes are associated with which biological functions. A classifier for longitudinal data is called for to address such a problem. In addition, the sample sizes for most longitudinal studies are small to moderate due to the cost and complexity of the longitudinal design. Hence, a desirable longitudinal classifier should also work effectively for finite

*Corresponding author
Email addresses: xwan2@fhcrc.org (Xianlong Wang), anniequ@illinois.edu (Annie Qu)

sample applications. As high throughput technologies become increasingly cost-effective, longitudinal studies will be conducted in more research fields, and more features or covariates will be collected at each time point. Therefore, there is an emerging demand for longitudinal classification tools to mine such high-dimensional longitudinal data.

Classifications for single point data are well developed, but these methods might not be effective for classifying longitudinal data. For longitudinal data, Choi [4] proposes a mixed model; Bagui and Mehra [2] develop a multi-stage nearest neighbor classification rule; Brown et al. [3] apply support vector machine (SVM); Liang and Kelemen [11] propose regularized neural networks; Lee [8], Rossi and Villa [16], Rossi and Villa [17] and Park et al. [14] apply the functional SVMs; Müller [13] uses functional principal component scores; Leng and Müller [9] use logistic regression; De la Cruz-Mesía et al. [5] apply semi-parametric Bayesian classification based on dependent Dirichlet processes; and Schmah et al. [18] compare several classification methods for longitudinal fMRI studies and identify the adaptive quadratic discriminant function and the support vector machine as the best classifiers. Functional data classifiers [7] are also applicable to most longitudinal data.

We propose a new classification method, *QIFC*, for longitudinal data based on the quadratic inference function (QIF) which builds a semi-parametric model. Our approach builds a classifier by taking advantage of modeling information between responses and covariates of the subjects within each class, and assigns a new subject to the class with the shortest newly defined distance to the subject. Our approach overcomes the difficulty in estimating covariance matrices as in linear discriminant analysis (LDA) while still being able to incorporate into the classifier the correlation among multiple observations on the same subject. We use simulation to compare *QIFC* to commonly used classifiers including the functional data classifier, SVM, logistic regression, linear discriminant analysis, the naive Bayes classifier and the decision tree. The proposed classifier shows advantages for both continuous and discrete response data for various settings. We also provide asymptotic optimality theory for *QIFC*. Applications to time-course gene expression data indicate that the generalization error of *QIFC* is improved compared to other classifiers when the sample sizes are small to moderate.

The paper is organized as follows. We describe *QIFC* in Section 2, and provide the theoretical results in Section 3. Simulation studies and applications follow in Section 4 and Section 5, respectively. Section 6 summarizes our results and provides a brief discussion.

2. *QIFC*

For longitudinal data, let $y_i(t)$ be a response variable and $x_i(t)$ be a $p \times 1$ vector of covariates, measured at time $t, t = t_1, \cdots, t_q$ for subject $i, i = 1, \cdots, N$. We assume that the model satisfies the first moment model assumption

$$\mu_i(t_j) = E\{y_i(t_j)\} = \mu\{x_i(t_j)'\beta\}, \tag{1}$$

where $\mu(\cdot)$ is a known inverse link function and β is a p-dimensional parameter vector. The quasi-likelihood equation [21] for longitudinal data is

$$\sum_{i=1}^{N} \dot{\mu}_i' V_i^{-1}(y_i - \mu_i) = 0,$$

where $V_i = Var(y_i), y_i = (y_i(t_1), \cdots, y_i(t_q))', \mu_i = (\mu_{it_1}, \cdots, \mu_{it_q})'$, and $\dot{\mu}_i = \partial\mu_i/\partial\beta$. In practice, V_i is often unknown, and the empirical estimator of V_i based on sample variance could be unreliable, especially when the sample size is small relative to the number of variance components in V_i. Liang and Zeger [10] introduce generalized estimating equations to substitute V_i by assuming $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i is a diagonal marginal variance matrix and R is a common working correlation matrix, which only involves a small number of nuisance parameters. The advantage of the GEE approach is that the GEE estimator of the regression parameter is consistent, even if the working correlation R is misspecified. However, the GEE estimator is not efficient within the same class of estimating functions when R is misspecified.

Qu et al. [15] introduced the quadratic inference function by assuming that the inverse of the working correlation can be approximated by a linear combination of several basis matrices, that is,

$$R^{-1} \approx a_1 M_1 + \cdots + a_m M_m,$$

where M_i's are symmetric matrices. We observe that the generalized estimating equation is an approximate linear combination of the components in the estimating functions,

$$\bar{g}_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\beta) = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N} (\dot{\mu}_i)' A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{N} (\dot{\mu}_i)' A_i^{-1/2} M_m A_i^{-1/2}(y_i - \mu_i). \end{pmatrix} \quad (2)$$

Hence, the advantage of this approach is that it does not require estimation of linear coefficients a_i's which can be viewed as nuisance parameters.

Since the dimension of (2) is larger than the number of parameters, we cannot set each component in (2) to be zero to solve for β. Instead we estimate β by setting $\bar{g}_N$ as close to zero as possible, in the sense of minimizing the quadratic function,

$$\hat{\beta} = \arg\min_{\beta} \bar{g}_N' \Omega^{-1} \bar{g}_N,$$

where $\Omega = Var(g_i)$. In practice, Ω is often unknown, but can be estimated consistently by $\bar{W}_N = N^{-1}\sum_{i=1}^{N} g_i g_i'$. The quadratic function,

$$Q_N(\beta) = N \bar{g}_N' \bar{W}_N^{-1} \bar{g}_N, \quad (3)$$

is called the quadratic inference function [15] since it provides an inference

function for the regression parameters.

To develop the new classifier, we first define a new distance measure based on the estimates from QIF. For a new subject, y^*, its distance to class c, $c = 1, \ldots, C$, where C is the total number of classes, is defined as

$$QD_c(y) := g_c' W_c^{-1} g_c. \tag{4}$$

In (4), W_c is the estimated covariance matrix from the training data in class c, and g_c is obtained as

$$g_c = \begin{pmatrix} (\hat{\dot{\mu}}_c)' \hat{A}_c^{-1/2} M_1 \hat{A}_c^{-1/2} (y^* - \hat{\mu}_c) \\ \vdots \\ (\hat{\dot{\mu}}_c)' \hat{A}_c^{-1/2} M_m \hat{A}_c^{-1/2} (y^* - \hat{\mu}_c) \end{pmatrix}, \tag{5}$$

where $\hat{\mu}_c$ is the estimated mean of class c, $\hat{A}_c$ is the estimated diagonal marginal variance matrix for class c, and $\hat{\dot{\mu}}_c$ is the estimate of $\dot{\mu}_c$ as in (2). The following algorithm summarizes the classification rule for classifying y^*.

Algorithm

1. For each class, fit a semi-parametric regression model using QIF.
2. Compute $QD_c(y^*)$ for each class $c, c = 1, \cdots, C$.
3. Let $m = \arg\min_c QD_c(y^*)$.
4. Assign the new subject y^* to class m.

For each class, step 1 models the longitudinal data and captures the mean function and correlation information within the same class. Step 2 computes the quadratic distance QD of the new subject to each class. Steps 3 and 4 assign the new subject to the class with the shortest distance. We call the new classifier *QIFC* since it builds on the model information derived from the quadratic inference function.

3. Theoretical Properties

We provide a couple of theoretical results for *QIFC*. Specifically, we will provide the optimality property of *QIFC* in Theorem 1, and the upper bound to the generalization error(GE) in Lemma 1.

To develop the statistical theory, we reformulate g_c in (5) as,

$$g_c = \tilde{T}_c' \left(y^* - \hat{\mu}_c \right),$$

where $\tilde{T}_c' \to_p T_c'$ and

$$T_c' = \begin{pmatrix} (\dot{\mu}_c)' A_c^{-1/2} M_1 A_c^{-1/2} \\ \vdots \\ (\dot{\mu}_c)' A_c^{-1/2} M_m A_c^{-1/2} \end{pmatrix}.$$

Note that the above convergence holds since both the GEE and QIF estimators are consistent.

Theorem 1. *Under the first moment assumption in (1) and suitable regularity conditions, QIFC based on $T_c' y_i$ is asymptotically optimal, i.e., the lowest misclassification error rate can be achieved, if*

$$T_c' y_i \ \underline{\underline{d}} \ \mu_i + \Sigma_i^{\frac{1}{2}} u \text{ and } |\Sigma_i| = |\Sigma_j|, \qquad i, j = 1, \ldots, C, \tag{6}$$

where u is a random vector with probability density function $f_0(u'u)$ such that $f_0(\cdot)$ is a strictly decreasing density function on $[0, \infty)$.

The proof of Theorem 1 is provided in the Appendix.

Remark: The requirement of $|\Sigma_i| = |\Sigma_j|$ in Theorem 1 is often approximately satisfied since $log|\Sigma_i|$ tends to be close in a logarithmic scale, even if Σ_i are different [20]. Note that the requirement on the monotone density function in Theorem 1 is satisfied for many statistical distributions commonly used in practice including the Gaussian distribution.

We have shown that *QIFC* is asymptotically optimal under condition (6). In the following discussion, we derive an upper bound for the generalization error, which provides a guideline to assess the error rate in practice.

Lemma 1. *Under the first moment assumption (1), for a two-class classification problem with equal probability priors, the misclassification error rate for QIFC is asymptotically bounded from above as follows,*

$$\begin{aligned} & P(\textit{misclassify a subject}) \\ \leq & \frac{1}{2} \frac{1}{1 + \left(\frac{(\mu_2 - \mu_1)' A_2 (\mu_2 - \mu_1) - tr((A_1 - A_2)\Sigma_1)}{\sqrt{2tr((A_1 - A_2)\Sigma_1 (A_1 - A_2)\Sigma_1) + 4(\mu_2' - \mu_1') A_2 \Sigma_1 A_2 (\mu_2 - \mu_1)}} \right)^2} \\ + & \frac{1}{2} \frac{1}{1 + \left(\frac{(\mu_1 - \mu_2)' A_1 (\mu_1 - \mu_2) - tr((A_2 - A_1)\Sigma_2)}{\sqrt{2tr((A_2 - A_1)\Sigma_1 (A_2 - A_1)\Sigma_2) + 4(\mu_1' - \mu_2') A_1) \Sigma_1 (A_2 (\mu_2 - \mu_1)}} \right)^2}. \end{aligned} \tag{7}$$

The above formula can be further simplified under more specific yet practical circumstances. For example, it is not uncommon that y_i from class c follows a multivariate Gaussian distribution $N(\mu_c, \Sigma)$. Then the upper bound reduces to $\frac{1}{1 + \left(\frac{(\mu_2 - \mu_1)' A (\mu_2 - \mu_1)}{\sqrt{4(\mu_2' - \mu_1') A \Sigma A (\mu_2 - \mu_1)}} \right)^2}$ if we denote $T = T_1 = T_2$ and $A = A_1 = A_2$.

Lemma 1 indicates that if $(\mu_2-\mu_1)'A(\mu_2-\mu_1) > 4$, then the misclassification error rate is strictly less than 0.5. Under such conditions, even though the upper bound itself might be loose, the classifier is guaranteed to be better than random guess.

4. Simulation Studies

We demonstrate the performance of *QIFC* on both continuous and discrete responses through simulation studies.

4.1. Continuous Responses

We first use a simulation setting to assess the performance of our method for continuous responses. We compare the performance of our method with that of the functional data classifier, SVM, logistic regression, and LDA. We also evaluate the performance of *QIFC* when the working correlation structure is misspecified. In addition, we evaluate the classifier when $\hat{\mu}_c$ in (5) is replaced by the GEE estimator. The error rate from Leave-One-Out cross validation will be used to evaluate the classification performance. We use R version 2.5.0 to implement our method. Specifically, we use the package "e1071" for SVM with the choice of the Gaussian kernel for which the C parameter is selected through a grid search over $2^{-5}, 2^{-4}, \ldots, 2^4, 2^5$, and gamma fixed at 1. Package "vgam" is used for the multinomial logistic regression, and "fda.usc" is used for the functional data classifier, and the smoothing parameter is selected by the generalized cross-validation "GCV.S" method.

4.1.1. Exchangeable Correlation Structure

For time-course gene expression data, two functional groups may follow a similar pattern, except that one has a delay in time relative to another due to, e.g., delay in regulation. To evaluate the performance of *QIFC* in this setting, we design the following simulation.

Let $\mathbf{Y} = (Y_{t_1}, \cdots, Y_{t_{20}}), t_1 = -2, t_2 = -1.92, \cdots, t_{20} = 2$ be the repeated measurements from a subject in class 1, and

$$E(Y_{t_i}) = 3t_i + 4\sin(3t_i) - 2\cos(3t_i) + 2\sin(4t_i), \qquad i = 1, \cdots, 20.$$

We assume the correlation structure of the repeated measurements to be *exchangeable*, and

$$Cov(\mathbf{Y}) = \sigma \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

where $\sigma = 100$, and $\rho = 0.85$.

For class 2, let $\mathbf{Y} = (Y_{t1}, \cdots, Y_{t20})$ be the repeated measurements of a subject with the same covariance matrix as class 1, except that

$$E(Y_{t_i}) = 3(t_i-0.5)+4\sin(t_i-0.5)-2\cos(t_i-0.5)+2\sin(2(t_i-0.5)),\ i = 1, \cdots, 20.$$

Clearly, the mean function of class 2 simply shifts to the right from class 1 by 0.5 units in time.

We generate 25 subjects from both class 1 and class 2. Figure 1 shows a realization of the two hypothetical classes. To illustrate the discriminant powers of the classifiers between the two classes, we generate a sequence of new subjects, $\mathbf{Y}(s) = (Y(s)_{t_1}, Y(s)_{t_2}, \cdots, Y(s)_{t_{20}})$, with the same covariance structure as above, and

$$E(Y(s)_{t_i}) = 3(t_i - s) + 4\sin(t_i - s) - 2\cos(t_i - s) + 2\sin(2(t_i - s)),\ i = 1, \cdots, 20,$$

where $s \in [0, 0.5]$. Note that $\mathbf{Y}(s)$ belongs to class 1 when $s = 0$, and belongs to class 2 when $s = 0.5$. For each $s \in [0, 0.5]$, we apply *QIFC*, the functional data classifier, SVM, logistic regression and LDA to predict its class label. For the LDA approach, we can use different non-singular covariance matrices to evaluate its performance due to the small sample size. In this paper, we use a diagonal matrix with the diagonal elements being the marginal variances. To assess the robustness of *QIFC* to misspecification of the working correlation, we also misspecify the correlation structure to be AR-1 and include the performance of *QIFC* in the comparison. When we compare classifiers, a favorable classifier should predict the closer class with high probability.

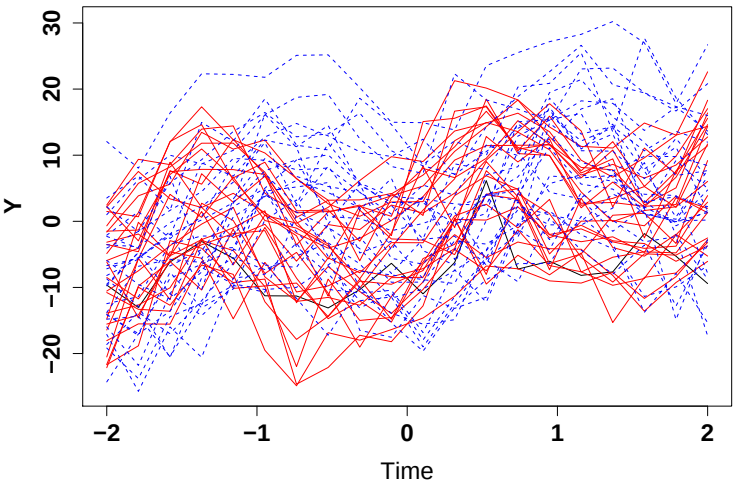

Figure 1: The simulation with 2 hypothetical classes and there is a time shift between them. The solid curves are in class 1, and the dashed curves are in class 2.

Based on 500 replicates, Figure 2 shows the probability of predicting class 2 when the time shift moves from 0 to 0.5. In particular, for the two end points $s = 0$ and $s = 0.5$, the probability can be used to calculate the generalization error. Figure 3 illustrate the upper bound as a function of the number of measurements

on a subject and the time shift. Assuming equal class priors, Table 1 provides the generalization error, the standard errors and the 95% confidence interval of the generalization error for each classifier. Note the confidence interval is constructed by assuming the number of errors to be the result from a binomial experiment, the total number of trial being 500.

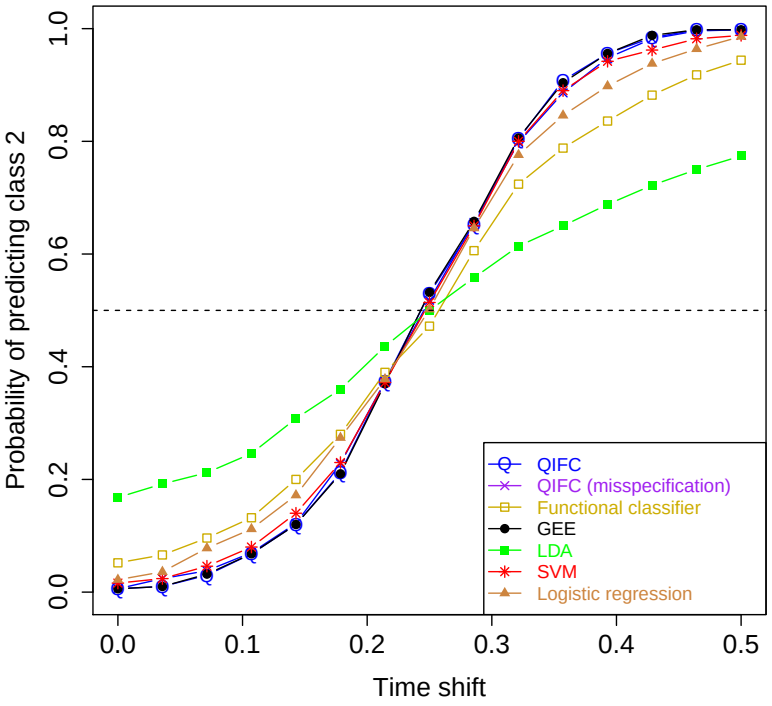

Figure 2: Performance comparison using exchangeable correlation structure on continuous responses.

Figure 2 indicates that *QIFC*, GEE and SVM have the highest sensitivity for the closer class, followed in order by logistic regression, the functional data classifier, and LDA is the least sensitive to the closer class. Note that all the classifiers has little discriminant power when the time shift moves to $s = 0.25$; and when the time shift s moves to either end, all the classifiers reach their highest discriminant power. Overall, the sensitivity of *QIFC* to the closer class is the highest, and misspecification of the working correlation does not appear to significantly affect the performance of *QIFC*.

The upper bound is less than 0.5, which implies that *QIFC* is at least better than random guess. Figure 3 implies that the upper bounds decreases as the number of repeated measurements increases, and decreases rapidly as the time

Table 1: Simulated generalization error, standard error and 95% confidence interval on continuous responses with exchangeable correlation structure.

	Classification error	standard error	confidence interval (95%)
SVM	0.202	0.018	(0.166, 0.237)
Logistic regression	0.018	0.006	(0.006, 0.030)
LDA	0.197	0.018	(0.162, 0.232)
Functional classifier	0.054	0.010	(0.034, 0.074)
GEE	0.004	0.003	(0.000, 0.010)
QIFC(misspecification)	0.004	0.003	(0.000, 0.010)
QIFC	0.004	0.003	(0.000, 0.010)
Upper bound	0.470	-	-

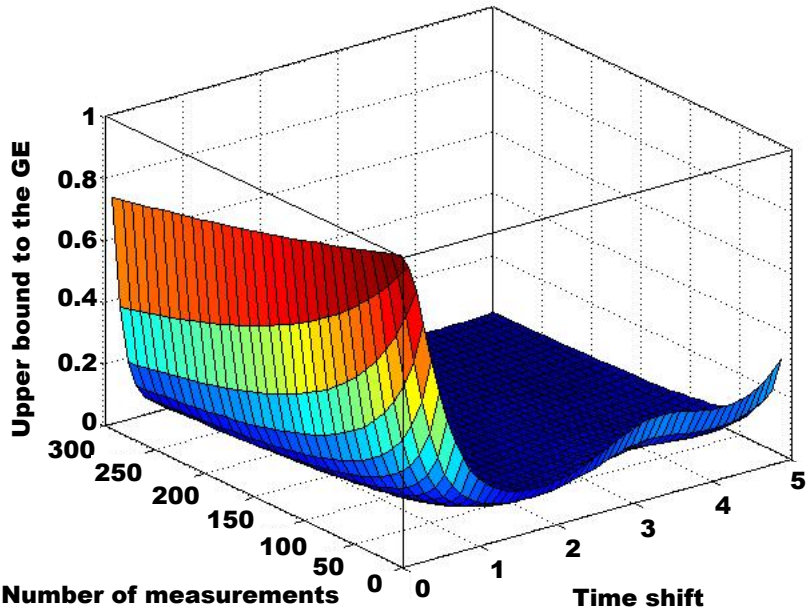

Figure 3: Upper bound of the error rate with respect to time shift and number of repeated measurements under an exchangeable correlation structure.

shift between the two classes increases. We note that the upper bound is derived based on asymptotics, and may not be tight for finite-sample applications such as this simulation study. However, examination of the upper as a function of relevant parameters sheds light on the worst-scenario performance of *QIFC* as the configuration of the classes changes.

The above result is not entirely surprising according to existing learning theory. It is well-known that under normality, LDA asymptotically outperforms other classifiers including SVM and logistic regression. On the other hand, for finite samples, the estimate of the covariance matrix may be singular, and hence the performance of LDA may be unstable. However, *QIFC* does not require direct estimation of the covariance matrix, and its performance is less affected by such an issue.

4.1.2. AR-1 Correlation Structure

We next assess the performance of *QIFC* on continuous responses with the same distribution assumption as above except that the responses assumes an AR-1 correlation structure. To assess the robustness of *QIFC* to misspecification, an exchangeable correlation structure plus an AR-1 component is used. In practice, we commonly apply the QIF with a rich classes of working correlation structures when we are not sure about the true structure. It is worth evaluating this type of misspecification.

Based on 500 replicates, Figure 4 shows the probability of predicting class 2 when the time shift moves from 0 to 0.5. Figure 5 illustrate the upper bound as a function of the number of measurements on a subject and the time shift. Assuming equal class priors, Table 2 sets forth the generalization errors, the standard error and the 95% confidence interval of the generalization error for each classifier.

Table 2: Simulated generalization error, standard error and 95% confidence interval on continuous responses with correlation structure AR-1.

	Classification error	standard error	confidence interval (95%)
SVM	0.202	0.018	(0.167, 0.237)
Logistic regression	0.177	0.017	(0.144, 0.210)
LDA	0.199	0.018	(0.164, 0.234)
Functional classifier	0.174	0.017	(0.141, 0.207)
GEE	0.142	0.016	(0.111, 0.173)
QIFC(misspecification)	0.139	0.015	(0.109, 0.169)
QIFC	0.135	0.015	(0.105, 0.165)
Upper bound	0.909	-	-

Overall, the order of performance does not change when the correlation structure changes from exchangeable to AR-1. That is, *QIFC* performs the best, and GEE does slightly worse, followed in order by logistic regression, the functional data classifier, and SVM and LDA are the least sensitive to the closer class. While the performances of *QIFC*, the functional data classifier, GEE and logistic regression deteriorate, SVM and LDA are robust to such a change. SVM is a model free classifier and expected to perform similarly when we change the correlation structure. The performance of LDA depends on the estimates of

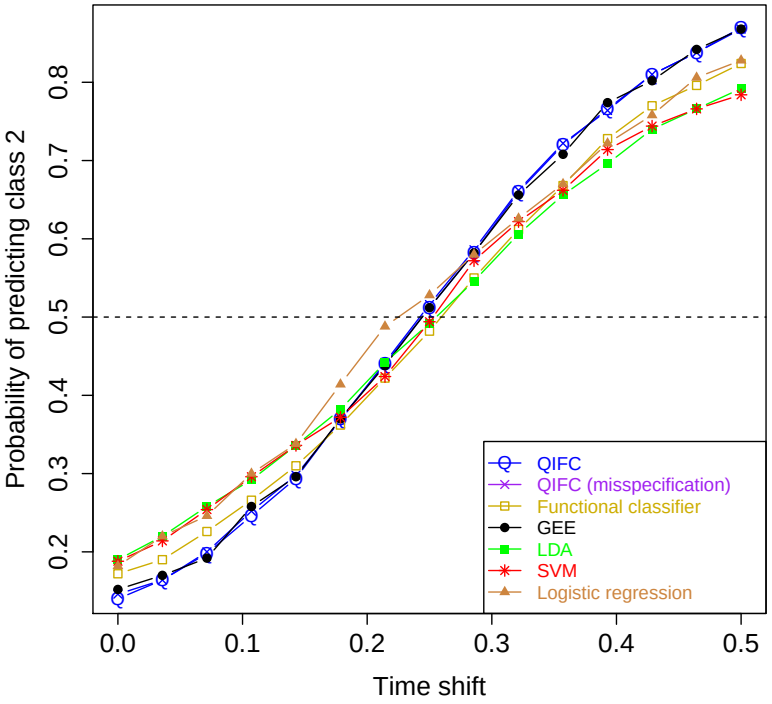

Figure 4: Performance comparison using AR-1 correlation structure on continuous responses.

the mean responses and the diagonal elements of the covariance matrix, and therefore changing the correlation structure should not affect its performance. The other classifiers are complex functions of the covariance matrix, and their performances are expected to be sensitive to the change of the correlation structure.

4.2. Binary Responses

By design, the application of *QIFC* is not restricted to continuous responses, and we next evaluate the performance of *QIFC* on binary responses. We compare the performance of *QIFC* with several commonly used classifiers, namely the functional data classifier, the naive Bayes classifier, the decision tree, and logistic regression.

We apply the naive Bayes classifier in R-package "e1071", and the decision tree in R-package "rpart." R package "bindata" is used to generate the binary data.

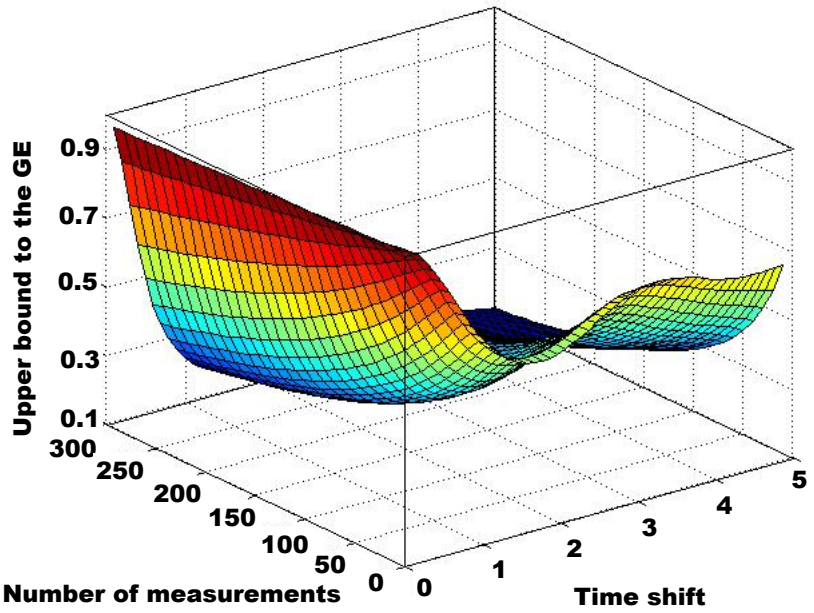

Figure 5: Upper bound of the error rate with respect to time shift and number of repeated measurements under an AR-1 correlation structure.

We generate two classes of subjects, and their mean functions satisfy

$$logit(E(Y_{t_i})) = \begin{cases} t_i - 0.9sin(2t_i) & \text{if } \mathbf{Y} \text{ is from class 1} \\ t_i + 0.3sin(2t_i) & \text{if } \mathbf{Y} \text{ is from class 2,} \end{cases}$$

$i = 1, \cdots, 100, t_1 = -1, t_2 = -0.98, \cdots, t_{100} = 1$, and $Cov(Y)$ has an AR-1 structure with correlation coefficient 0.8. The sample size for each class is 25.

To illustrate the discriminant powers of these classifiers, we generate a sequence of new subjects $\mathbf{Y}(c) = (Y(c)_{t_1}, Y(c)_{t_2}, \cdots, Y(c)_{t_{100}}),\ c \in [-0.9, 0.3]$. The new subjects also assume the same AR-1 correlation structure, and the mean functions satisfy

$$logit(E(Y(c)_{t_i})) = t_i + csin(2t_i),\ i = 1, \cdots, 100.$$

When $c = -0.9$, $\mathbf{Y}(c)$ belongs to class 1; as c increases, $\mathbf{Y}(c)$ moves towards class 2, and belongs to class 2 when $c = 0.3$. For each value c on a chosen grid between -0.9 and 0.3, we classify the new subjects using *QIFC*, the functional data classifier, the naive Bayes classifier, the decision tree and logistic regression. Based on 500 replicates, Figure 6 displays the probability of the response taking value 1 over time for the two classes, and Figure 7 plots the probability of

predicting class 2 as c moves from -0.9 to 0.3. Assuming equal probability priors for the two classes, Table 3 shows the generalization error, the standard errors and the 95% confidence interval of the generalization errors for each classifier.

Figure 7 indicates that *QIFC* picks up discriminant power most rapidly among the 4 classifiers when the new subject gets closer to either class. Logistic regression and GEE have little discriminant power. The decision tree tends to classify the new subject to class 2 more frequently than to class 1 over a wide range of coefficient changes, whereas GEE biases toward class 1. The performances of naive Bayes classifier and the functional data classifier are slightly inferior to *QIFC*.

The theoretical upper bound indicates that *QIFC* is at least guaranteed to be better than random guess. Similar to the previous results on continuous responses, the upper bound decreases as the coefficient of the new subject or the number of repeated measurements increases, as shown in Figure 8.

Table 3: Simulated generalization error, standard error and 95% confidence interval for the decision tree, logistic regression, the naive Bayes classifier, and *QIFC*.

	Classification error	standard error	confidence interval (95%)
Decision Tree	0.424	0.022	(0.381, 0.467)
Logistic regression	0.481	0.022	(0.437, 0.525)
Naive Bayes	0.372	0.022	(0.330, 0.414)
GEE	0.444	0.022	(0.400, 0.488)
Functional classifier	0.354	0.021	(0.312, 0.396)
QIFC	0.330	0.021	(0.289, 0.371)
Upper bound	0.441	-	-

5. Applications

Using two time-course gene expression data sets, we assess the performances of the new classifiers and the functional data classifier, SVM, LDA and logistic regression. Both applications involve multiclass classification. The error rate from Leave-One-Out cross-validation will be used to evaluate their performance.

5.1. Yeast (Saccharomyces cerevisiae) Cell Cycle Data

The yeast cell microarray data contains the gene expression profiles of 2467 budding yeast (*Saccharomyces cerevisiae*) genes over 79 time points [19], including the cell division cycle after synchronization by alpha factor arrest (ALPH, 18 time points), centrifugal elutriation (ELU, 14 time points), a cdc15 mutant (CDC15, 15 time points), sporulation (SPO, 11 time points), shock by high temperature (HT, 6 time points), reducing agents (D, 4 time points) and low temperature (C, 4 time points). Eisen et al. [6] conducted a clustering analysis,

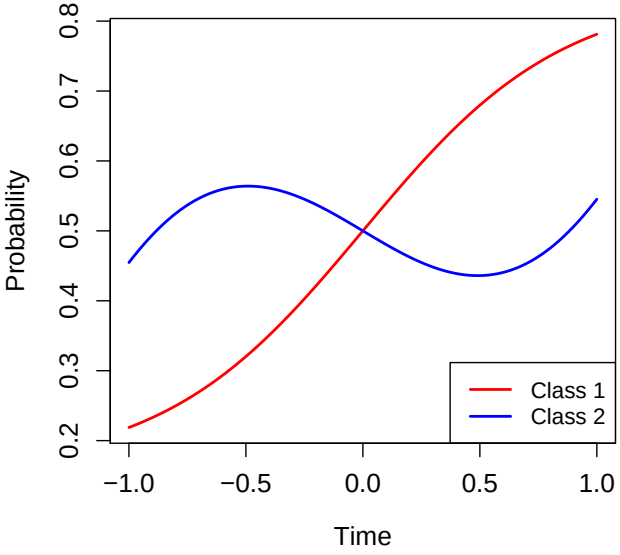

Figure 6: Logit of probabilities over time for the two classes.

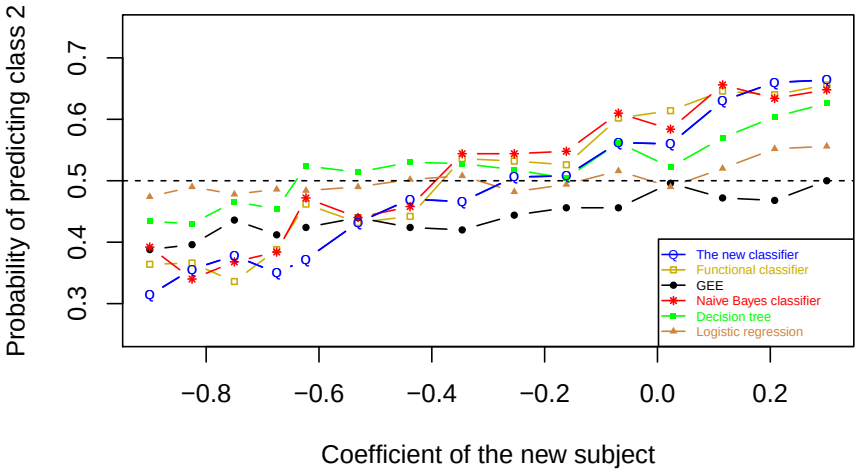

Figure 7: Performance comparison of *QIFC*, logistic regression, the decision tree and the naive Bayes classifier on discrete binary responses.

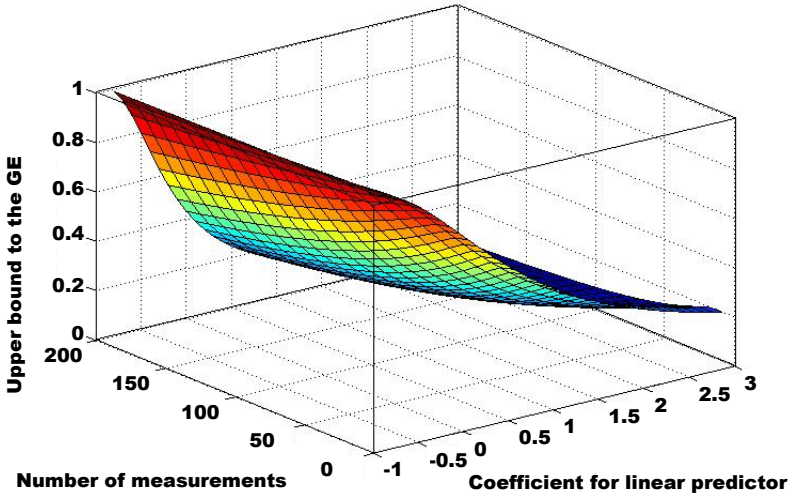

Figure 8: Upper bound of the error rate with respect to the coefficients in the linear predictor and number of repeated measurements.

and assigned these genes to different functional groups. Note these clusters were obtained from the hierarchical clustering method, which does not make use of the longitudinal nature of the data, and there is no apparent reason the results is in the favor of *QIFC*.

To demonstrate the small sample performance of *QIFC*, we analyze the 75 genes in the centrifugal elutriation experiment containing 5 functionally related groups: spindle pole body assembly and function, the proteasome, chromatin structure, the ribosome and translation, and DNA replication. The first five panels in Figure 9 show the individual classes along with the mean functions and the last panel plots the all the classes together. Table 4 summarizes the performance of each classifier. *QIFC* has the lowest generalization error (0.053), followed by GEE, LDA, the functional classifier, logistic regression, and finally SVM.

5.2. Wild-type Fly (Drosophila melanogaster) Temporal Data

The second example is based on Arbeitman et al. [1]'s study of the mRNA levels of 4028 genes in wildtype flies (*Drosophila melanogaster*) with cDNA microarrays over 70 time-points spanning fertilization, embryonic, larval, and pupal stages and the first 30 days of adulthood. Ma et al. [12] developed a data-driven clustering method for this data. We choose 6 clusters consisting of 1120 genes which demonstrate distinct yet similar gene expression patterns in the fertilization stage to evaluate *QIFC*. The first six panels in Figure 10 display the individual classes and the last panel shows the combined data.

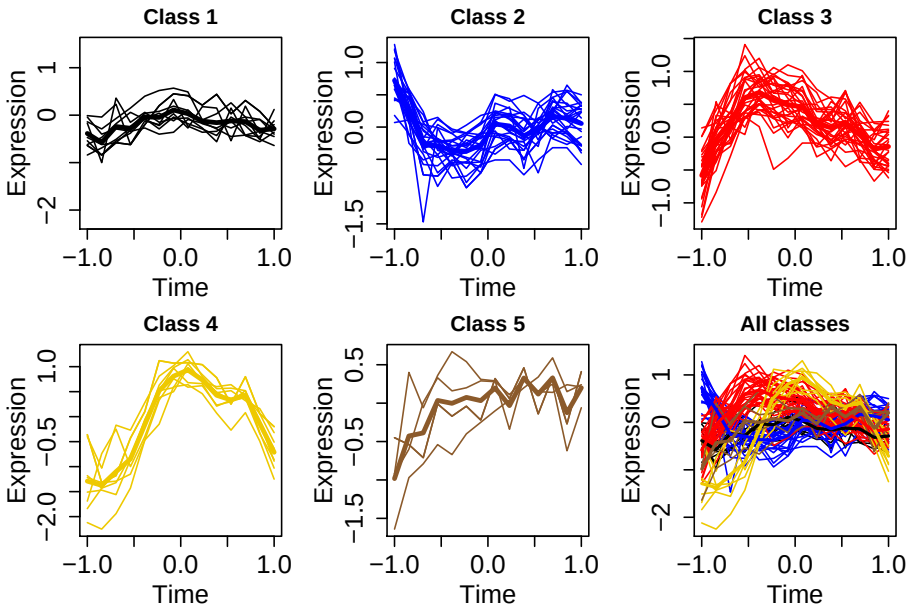

Figure 9: The five classes from the yeast cell cycle data. The thick lines are the fitted means for each class.

Table 4: Performance of SVM, logistic regression, LDA, and *QIFC* on yeast cell data

	number of errors	Classification error	standard error	confidence interval (95%)
SVM	30	0.4	0.057	(0.289, 0.511)
Logistic regression	15	0.200	0.046	(0.109, 0.291)
LDA	7	0.093	0.033	(0.028, 0.158)
GEE	6	0.08	0.031	(0.019, 0.141)
Functional classifier	9	0.12	0.038	(0.045, 0.195)
QIFC	4	0.053	0.026	(0.001, 0.105)

Table 5 summarizes the generalization errors of the four classifiers. *QIFC* demonstrates the best performance with classification error 0.14, followed by the performances of LDA, the functional classifier, GEE, SVM, and finally logistic regression. Compared to the previous data example, the classification errors of all the competing classifiers are higher. We note that classes 2, 4, 5 and 6 show rather similar gene expression patterns, which make the classification difficult. Also, note that the performances of LDA and our classifier are similar, and both achieve approximately the same classification accuracy, which is likely attributed to the large sample size for this data example.

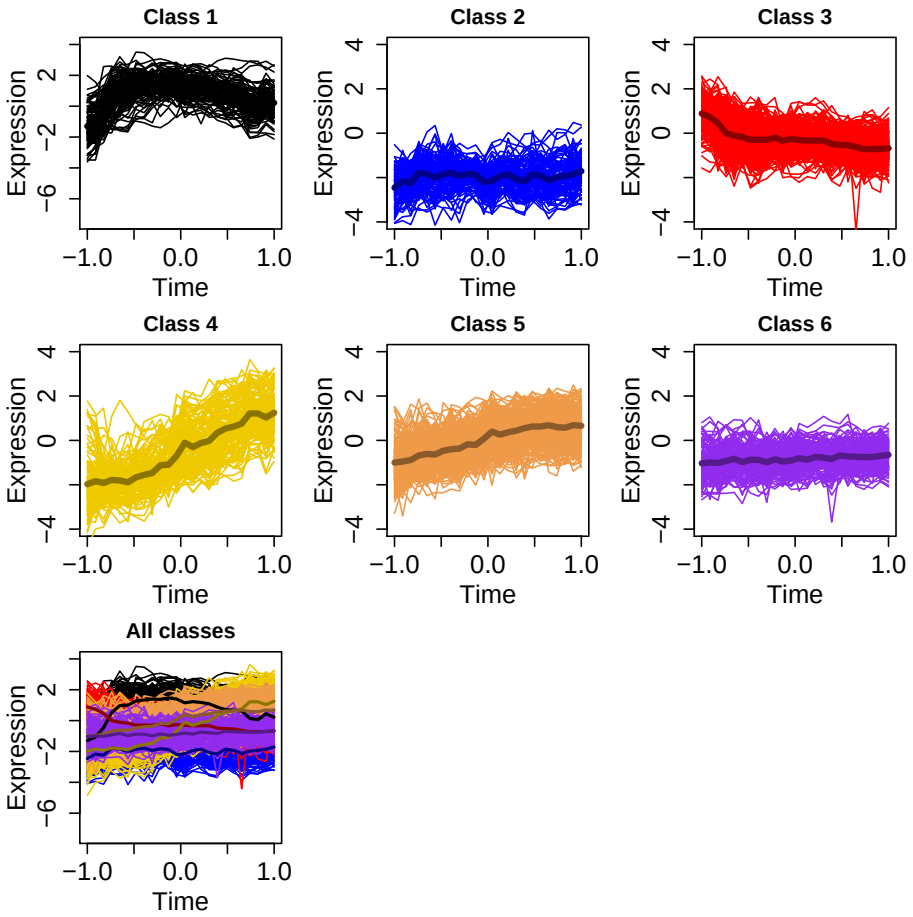

Figure 10: The six classes from the fruit fly gene expression data. The thick lines are the fitted means for each class.

Table 5: Performance of SVM, logistic regression, LDA, and *QIFC* on the fruit fly data

	number of errors	Classification error	standard error	confidence interval (95%)
SVM	442	0.395	0.015	(0.366, 0.423)
Logistic regression	979	0.874	0.010	(0.854, 0.894)
LDA	168	0.150	0.011	(0.128, 0.172)
GEE	289	0.258	0.013	(0.232, 0.284)
Functional classifier	210	0.188	0.012	(0.165, 0.210)
QIFC	157	0.140	0.010	(0.120, 0.160)

6. Summary and Discussion

We propose a new classifier *QIFC* for classifying longitudinal data based on the quadratic inference function. Our approach provides a classifier suitable for both continuous and discrete responses. Simulation studies and real data applications demonstrate the favorable performance of *QIFC*, as compared with other popular classifiers including the functional data classifier, the support vector machine, logistic regression, linear discriminant analysis, the naive Bayes classifier and the decision tree. In addition, *QIFC* is robust to misspecification of the working correlation structure. *QIFC* is the first classifier proposed based on estimation equations for longitudinal data.

QIFC models and incorporates the intrinsic correlations among longitudinal observations within a cluster for finite sample applications. Most classifiers don't specifically model the within-cluster correlation structure, which hinders their performances for longitudinal data. In contrast, LDA estimates the full covariance matrix which involves a large number of parameters. When LDA is used in applications with small to moderate sample size, either the estimate of the covariance matrix is singular or the estimate is unstable so that the classifier is underpowered. *QIFC* treats the covariance matrix as a finite linear combination of basis matrices, which serves most applications in longitudinal studies. Therefore, *QIFC* makes better use of the correlation information and enjoys superior performance. On the other hand, as the sample size increases, traditional classifiers especially LDA will be able to estimate the covariance matrix more precisely and thus picks up more power, and is expected to be comparable to *QIFC* as the sample size is sufficiently large. Since most longitudinal studies are expensive and the number of subjects are usually limited, *QIFC* serves as a favorable classifier.

The power of the *QIFC* draws on the underlying semi-parametric model for longitudinal data. QIF, which improves upon GEE, fits most longitudinal data adequately, and *QIFC* gains its power from the model information. On the other hand, classifiers including SVM, the decision tree and the naive Bayes classifier approach the problem from a somewhat nonparametric perspective, and does not efficiently fit the model inherent in longitudinal data. The functional classifier, which assumes that the measurements on a subject come from a smooth curve, is powerful in its own right for the right types of data, but may not achieve its optimal performance for longitudinal data where the measurements are often not taken at a large number of dense temporal points.

We have assumed balanced longitudinal data so far for developing *QIFC*, i.e., every subject is measured at the same number of time/spatial points. However, unbalanced data are quite common due to measurement constraints, missing data, and quality control. To build a classifier for unbalanced data, we can adopt techniques similar to Zhou and Qu [22] as follows. Suppose subject i has q_i measurements, where not all q_i are equal, and let q be the total number of time points cross all the subjects. To adapt *QIFC* for such unbalanced data, we first define the $q \times q_i$ linear transformation C_i for the ith subject by removing the columns of the identity matrix, where the removed columns

correspond to the missing observations. Now we augment the measurements on each subject to q measurements by introducing $y_i^* = C_i y_i$, $\mu_i^*(\beta) = C_i \mu_i(\beta)$, $\dot{\mu}_i^*(\beta) = C_i \dot{\mu}_i(\beta)$, and $A_i^* = C_i A_i C_i'$. Note the components in y_i^* are the same as in y_i for nonmissing responses but are 0 otherwise. Now the proposed classifier can be used based on the newly defined variables. This works because the 0 values specified in $\dot{\mu}_i^*$ and $y_i^* - \mu_i^* \beta$ corresponding to the missing observations ensure that the missing observations do not contribute to the objective function in (3) and the distance metric in (4).

Since longitudinal studies have been, and will continue to be, an effective statistical design to evaluate long-term covariate effects, profile disease-relevant clinical variables and predict disease outcomes in the biomedical field, our classifier will serve as a powerful alternative to traditional classifiers.

Acknowledgement

Qu's research was supported by the National Science Foundation (DMS-0906660). The authors are grateful to Dr. David Birkes for his insightful discussions.

[1] M. N. Arbeitman, E. E. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White. Gene expression during the life cycle of drosophila melanogaster. *Science*, 297: 2270–2275, 2002.

[2] S. C. Bagui and K. L. Mehra. Classification of multiple observations using multi-stage rank nearest neighbor rule. *J. Statist. Plann. Inf.*, 76(1-2): 163–183, 1999.

[3] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Jr Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.

[4] S. C. Choi. Classification of multiply observed data. *Biometrical J.*, 14: 8–11, 1972.

[5] R. De la Cruz-Mesía, F. A. Quintana, and P. Müller. Semiparametric bayesian classification with longitudinal markers. *Applied Statistics (Journal of the Royal Statistical Society, Series C)*, 56 (2):119–137, 2007.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of National Academy of Sciences*, pages 14863–14868, 1998.

[7] M. Febrero-Bande and M. Oviedo de la Fuente. Statistical computing in functional data analysis: The r package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012.

[8] H. J. Lee. *Functional data analysis: classification and regression.* PhD thesis, Texas A & M University, 2004.

[9] X. Leng and H. G. Müller. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22:68–76, 2006.

[10] K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalised linear models. *Biometrika*, 73:12–22, 1986.

[11] Y. Liang and A. Kelemen. Temporal gene expression classification with regularised neural network. *International Journal of Bioinformatics Research and Applications*, 1(4):399–413, 2005.

[12] P. Ma, C. I. Castillo-Davis, W. Zhong, and J. S. Liu. A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34:1261–1269, 2006.

[13] H. G. Müller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240, 2005.

[14] C. Park, J. Koo, S. Kim, I. Sohn, and J. W. Lee. Classification of gene functions using support vector machine for time-course gene expression data. *Computational Statistics & Data Analysis*, 52:2578–2587, 2008.

[15] A. Qu, B. G. Lindsay, and B. Li. Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4):823–836, 2000.

[16] F. Rossi and N. Villa. Classification in Hilbert spaces with support vector machines. In *Proceedings of XIth International Symposium on Applied Stochastic Models and Data Analysis*, 2005.

[17] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69:730–742, 2006.

[18] T. Schmah, G. Yourganov, R. S. Zemel, G. E. Hinton, Small S. L., and Strother S. C. Comparing of classification methods for longitudinal fmri studies. *Neural Computation*, 2010.

[19] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hibridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[20] S. Velilla and A. Hernández. On the consistency properties of linear and quadratic discriminant analyses. *Journal of Multivariate Analysis*, 96(2): 219–236, 2005.

[21] R. W. M. Wedderburn. Quasilikelihood functions, generalized linear models and the gauss-newton method. *Biometrika*, 61:439–447, 1974.

[22] J. Zhou and A. Qu. Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association*, 107:498:701–710, 2012.

Appendix A.

A.1 Proof of Theorem 1

We prove the theorem for two classes, and the proof can be easily generalized to multiple class settings.

Suppose f_1 is the probability density function corresponding to class 1 with mean μ_1 and variance-covariance matrix V_1.

By definition of $QD_1(y)$, we have,

$$\begin{aligned}
&QD_1(y)\\
=&g_1' W_1^{-1} g_1\\
=&[\tilde{T}_1'(y-\hat{\mu}_1)]' W_1^{-1} [\tilde{T}_1'(y-\hat{\mu}_1)]\\
\underrightarrow{p}&[T_1'(y-\mu_1)]'(T_1'V_1T_1)^{-1}[T_1'(y-\mu_1)]\\
=&[T_1'(y-\mu_1)]'\Sigma_1^{-1}[T_1'(y-\mu_1)], \text{ as training sample size } n\to\infty, \text{where } \Sigma_1^{-1}=(T_1'V_1T
\end{aligned} \tag{1}$$

Similarly,

$$QD_2(y) \underrightarrow{p} [T_2'(y-\mu_2)]'\Sigma_2^{-1}[T_2'(y-\mu_2)], \quad \text{where } \Sigma_2^{-1}=(T_2'V_2T_2)^{-1}. \tag{2}$$

Under the inverse location regression model (6),

$$T_c'y_1 \ \underline{\underline{d}} \ \mu_1+\Sigma_1^{\frac{1}{2}}u \Longrightarrow f(T_c'y_1)=|\Sigma_1|^{-\frac{1}{2}} f_0[(T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1))].$$

In the same way, for class 2, we have,

$$f(T_c'y_2)=|\Sigma_2|^{-\frac{1}{2}} f_0[(T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2))].$$

Now the optimal classification boundary is determined by,

$$r(y)=\begin{cases} 1 & \dfrac{|\Sigma_1|^{-\frac{1}{2}} f_0[(T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1))]}{|\Sigma_2|^{-\frac{1}{2}} f_0[(T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2))]}>1 \\ 2 & \dfrac{|\Sigma_1|^{-\frac{1}{2}} f_0[(T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1))]}{|\Sigma_2|^{-\frac{1}{2}} f_0[(T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2))]}<1 \end{cases}. \tag{3}$$

Under the inverse location model, $|\Sigma_i|=|\Sigma_j|$, $i=1,\ldots,C.$, and hence the classification rule (3) is equivalent to,

$$r(y)=\begin{cases} 1 & \dfrac{f_0[(T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1))]}{f_0[(T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2))]}>1 \\ 2 & \dfrac{f_0[(T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1))]}{f_0[(T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2))]}<1 \end{cases}.$$

Further,

$$r(y) = \begin{cases} 1 & f_0[(T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1))] > f_0[(T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2))] \\ 2 & f_0[(T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1))] < f_0[(T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2))] \end{cases}.$$

Using the monotonicity of f_0, we have,

$$r(y) = \begin{cases} 1 & (T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1)) < (T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2)) \\ 2 & (T_1'(y_1-\mu_1))\Sigma_1^{-1}(T_1'(y_1-\mu_1)) > (T_2'(y_2-\mu_2))\Sigma_2^{-1}(T_2'(y_2-\mu_2)) \end{cases}. \tag{4}$$

Combining (1) and (2) with (4), we complete the proof.

A.2 Proof of Lemma 1

Let us first assume the training data are generated from 2 populations, $P1$ with mean μ_1 and variance covariance Σ_1, and $P2$ with mean μ_2 and variance covariance Σ_2. Suppose a new subject, y, is from $P1$. Now we calculate the probability of classifying y to $P2$.

Let $A_1 = T_1W_1^{-1}T_1'$, $A_2 = T_2W_2^{-1}T_2'$, $Q = y'(T_1W_1^{-1}T_1' - T_2W_2^{-1}T_2')y - 2(\mu_1'T_1W_1^{-1}T_1' - \mu_2'T_2W_2^{-1}T_2')y = y'(A_1 - A_2)y - 2(\mu_1'A_1 - \mu_2'A_2)y$, and $A = A_1 - A_2$.

Then, when the training set is infinitely large,

$$\begin{aligned}
\text{classify } y \text{ into } P2 \Leftrightarrow & QD_1(y) > QD_2(y) \\
\Leftrightarrow & (T_1'y - T_1'\mu_1)'W_1^{-1}(T_1'y - T_1'\mu_1) > (T_2'y - T_2'\mu_2)'W_2^{-1}(T_2'y - T_2'\mu_2 \\
\Leftrightarrow & (y-\mu_1)'T_1W_1^{-1}T_1'(y-\mu_1) > (y-\mu_2)'T_2W_2^{-1}T_2'(y-\mu_2) \\
\Leftrightarrow & y'T_1W_1^{-1}T_1'y - 2y'T_1W_1^{-1}T_1'\mu_1 + \mu_1'T_1W_1^{-1}T_1'\mu_1 \\
& > y'T_2W_2^{-1}T_2'y - 2y'T_2W_2^{-1}T_2'\mu_2 + \mu_2'T_2W_2^{-1}T_2'\mu_2 \\
\Leftrightarrow & y'(A_1 - A_2)y - 2(\mu_1'A_1 - \mu_2'A_2) \\
& > \mu_2'A_2\mu_2 - \mu_1'A_1\mu_1 \\
\Leftrightarrow & Q > \mu_2'A_2\mu_2 - \mu_1'A_1\mu_1.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& P\{\text{Classify } y \text{ into } P2 | y \text{ is from } P1\} \\
=& P\{y'(A_1 - A_2)y - 2(\mu_1'A_1 - \mu_2'A_2)y > \mu_2'A_2\mu_2 - \mu_1'A_1\mu_1\}.
\end{aligned}$$

In applications, it is possible to simulate the classification error, but the analytical properties of the error rate remain to be unraveled. Next we investigate

some of its statistical properties.

We have,

$$
\begin{aligned}
&E(y'(A_1 - A_2)y) \\
=&tr((A_1 - A_2)\Sigma_1) + \mu_1'(A_1 - A_2)\mu_1 \\
=&tr((A_1 - A_2)\Sigma_1) + \mu_1'(A_1 - A_2)\mu_1,
\end{aligned}
$$

$$
\begin{aligned}
&E(2(\mu_1'A_1 - \mu_2'A_2)y) \\
=&2(\mu_1'A_1 - \mu_2'A_2)\mu_1 \\
=&2\mu_1'A_1\mu_1 - 2\mu_2'A_2\mu_1,
\end{aligned}
$$

$$
\begin{aligned}
&E(Q) \\
=&E(y'(A_1 - A_2)y) - E(2(\mu_1'A_1 - \mu_2'A_2)y) \\
=&tr((A_1 - A_2)\Sigma_1) + \mu_1'(A_1 - A_2)\mu_1 - 2\mu_1'A_1\mu_1 + 2\mu_2'A_2\mu_1,
\end{aligned}
$$

$$
\begin{aligned}
&Var(y'(A_1 - A_2)y) \\
=&2tr((A_1 - A_2)\Sigma_1(A_1 - A_2)\Sigma_1) + 4\mu_1'(A_1 - A_2)\Sigma_1(A_1 - A_2)\mu_1,
\end{aligned}
$$

$$
\begin{aligned}
&Var(2(\mu_1'A_1 - \mu_2'A_2)y) \\
=&4(\mu_1'A_1 - \mu_2'A_2)\Sigma_1(\mu_1'A_1 - \mu_2'A_2)' \\
=&4(\mu_1'A_1 - \mu_2'A_2)\Sigma_1(\mu_1'A_1 - \mu_2'A_2)',
\end{aligned}
$$

and,

$$
\begin{aligned}
&Cov(y'(A_1 - A_2)y, 2(\mu_1'A_1 - \mu_2'A_2)y) \\
=&Cov(y'(A_1 - A_2)y, 2(\mu_1'A_1 - \mu_2'A_2)y) \\
=&E[y'(A_1 - A_2)y * 2(\mu_1'A_1 - \mu_2'A_2)y] - E[y'(A_1 - A_2)y]E[2(\mu_1'A_1 - \mu_2'A_2)y] \\
=&2\mu_1'(A_1 - A_2)\mu_1(\mu_1'A_1 - \mu_2'A_2)\mu_1 + 2tr((A_1 - A_2)\Sigma_1(\mu_1'A_1 - \mu_2'A_2)\mu_1) \\
&+ 4tr((A_1 - A_2)\mu_1(\mu_1'A_1 - \mu_2'A_2)\Sigma_1) \\
&- [tr((A_1 - A_2)\Sigma_1) + \mu_1'(A_1 - A_2)\mu_1]2(\mu_1'A_1 - \mu_2'A_2)\mu_1.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&Var(Q) \\
=&Var(y'(A_1 - A_2)y) + Var(2(\mu_1'A_1 - \mu_2'A_2)y) \\
&- 2Cov(y'(A_1 - A_2)y, 2(\mu_1'A_1 - \mu_2'A_2)y) \\
=&2tr((A_1 - A_2)\Sigma_1(A_1 - A_2)\Sigma_1) + 4\mu_1'(A_1 - A_2)\Sigma_1(A_1 - A_2)\mu_1 \\
&+ 4(\mu_1'A_1 - \mu_2'A_2)\Sigma_1(\mu_1'A_1 - \mu_2'A_2)' - 2[2\mu_1'(A_1 - A_2)\mu_1(\mu_1'A_1 - \mu_2'A_2)\mu_1 \\
&+ 2tr((A_1 - A_2)\Sigma_1(\mu_1'A_1 - \mu_2'A_2)\mu_1) + 4tr((A_1 - A_2)\mu_1(\mu_1'A_1 - \mu_2'A_2)\Sigma_1) \\
&- [tr((A_1 - A_2)\Sigma_1) + \mu_1'(A_1 - A_2)\mu_1]2(\mu_1'A_1 - \mu_2'A_2)\mu_1] \\
=&2tr((A_1 - A_2)\Sigma_1(A_1 - A_2)\Sigma_1) + 4(\mu_2' - \mu_1')A_2)\Sigma_1(A_2(\mu_2 - \mu_1).
\end{aligned}
$$

To get an upper bound to the error rate, recall that Cantelli's Inequality (Grimmett and Stirzaker, 2001) states the following.

Let X be a random variable with expected value μ and finite variance σ^2. Then for any real number $k > 0$,

$$Pr\left(X - \mu \geq k\sigma\right) \leq \frac{1}{1+k^2}.$$

Applying the inequality gives us the following upper bound,

$$\begin{aligned}
&P\{\text{Classify } y \text{ into } P_2 | y \text{ is from } P_1\} \\
=&P\{Q > \mu_2' A_2 \mu_2 - \mu_1' A_1 \mu_1\} \\
=&P\{Q - E(Q) > \mu_2' A_2 \mu_2 - \mu_1' A_1 \mu_1 - E(Q)\} \\
=&P\{Q - E(Q) > \frac{\mu_2' A_2 \mu_2 - \mu_1' A_1 \mu_1 - E(Q)}{\sqrt{Var(Q)}} \sqrt{Var(Q)}\} \\
\leq&\frac{1}{1+\left(\frac{\mu_2' A_2 \mu_2 - \mu_1' A_1 \mu_1 - E(Q)}{\sqrt{Var(Q)}}\right)^2} \\
=&\frac{1}{1+\left(\frac{(\mu_2-\mu_1)' A_2 (\mu_2-\mu_1) - tr((A_1-A_2)\Sigma_1)}{\sqrt{2tr((A_1-A_2)\Sigma_1(A_1-A_2)\Sigma_1) + 4(\mu_2'-\mu_1')A_2\Sigma_1 A_2(\mu_2-\mu_1)}}\right)^2}.
\end{aligned}$$

Similarly, we get,

$$\begin{aligned}
&P\{\text{Classify } y \text{ into } P_1 | y \text{ is from } P_2\} \\
\leq&\frac{1}{1+\left(\frac{(\mu_1-\mu_2)' A_1 (\mu_1-\mu_2) - tr((A_2-A_1)\Sigma_2)}{\sqrt{2tr((A_2-A_1)\Sigma_1(A_2-A_1)\Sigma_2) + 4(\mu_1'-\mu_2')A_1\Sigma_2 A_1(\mu_2-\mu_1)}}\right)^2}.
\end{aligned}$$

Assuming equal frequencies for both populations, the upper bound to the error rate is given by

$$\begin{aligned}
&\frac{1}{2}\frac{1}{1+\left(\frac{(\mu_2-\mu_1)' A_2 (\mu_2-\mu_1) - tr((A_1-A_2)\Sigma_1)}{\sqrt{2tr((A_1-A_2)\Sigma_1(A_1-A_2)\Sigma_1) + 4(\mu_2'-\mu_1')A_2\Sigma_1 A_2(\mu_2-\mu_1)}}\right)^2} \\
+&\frac{1}{2}\frac{1}{1+\left(\frac{(\mu_1-\mu_2)' A_1 (\mu_1-\mu_2) - tr((A_2-A_1)\Sigma_2)}{\sqrt{2tr((A_2-A_1)\Sigma_1(A_2-A_1)\Sigma_2) + 4(\mu_1'-\mu_2')A_1\Sigma_1 A_2(\mu_2-\mu_1)}}\right)^2}.
\end{aligned}$$

A.3 Simulation Result on Continuous Responses with Block Diagonal Correlation

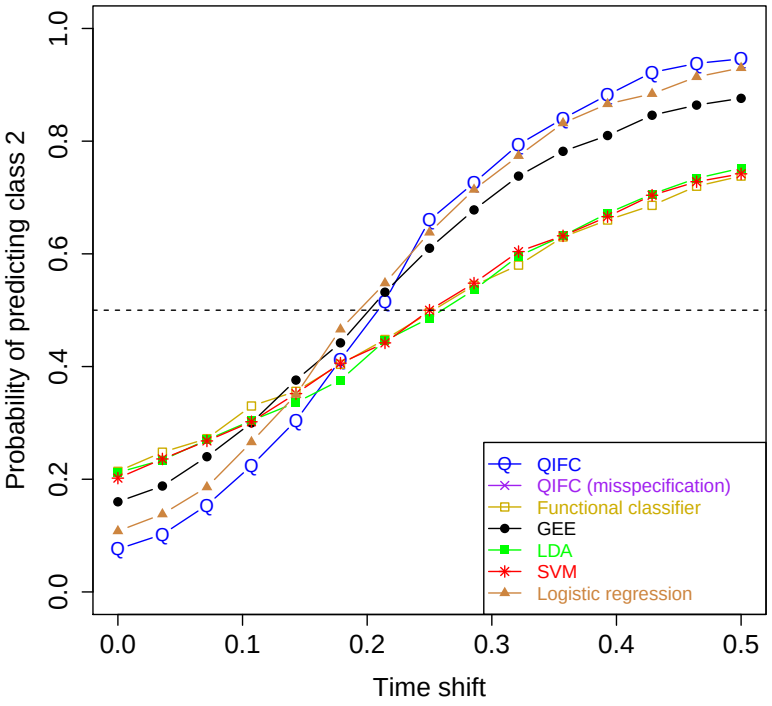

Figure S-1: Performance comparison based on the same setting as in Section 4.1 except that the responses have a block diagonal correlation structure with parameter 0.85 and block size 5, and the misspecification of *QIFC* sets the correlation to be exchangeable plus AR-1.

Table S-1: Simulated generalization error, standard error and 95% confidence interval on continuous responses with block diagonal correlation structure with parameter 0.85 and block size 5.

	Classification error	standard error	confidence interval (95%)
SVM	0.23	0.019	(0.193, 0.267)
Logistic regression	0.089	0.013	(0.064, 0.114)
LDA	0.230	0.019	(0.193, 0.267)
Functional classifier	0.238	0.019	(0.201, 0.275)
GEE	0.142	0.016	(0.111, 0.173)
New classifier(misspecification)	0.064	0.011	(0.043, 0.085)
New classifier	0.065	0.011	(0.043, 0.087)
Upper bound	0.783	-	-

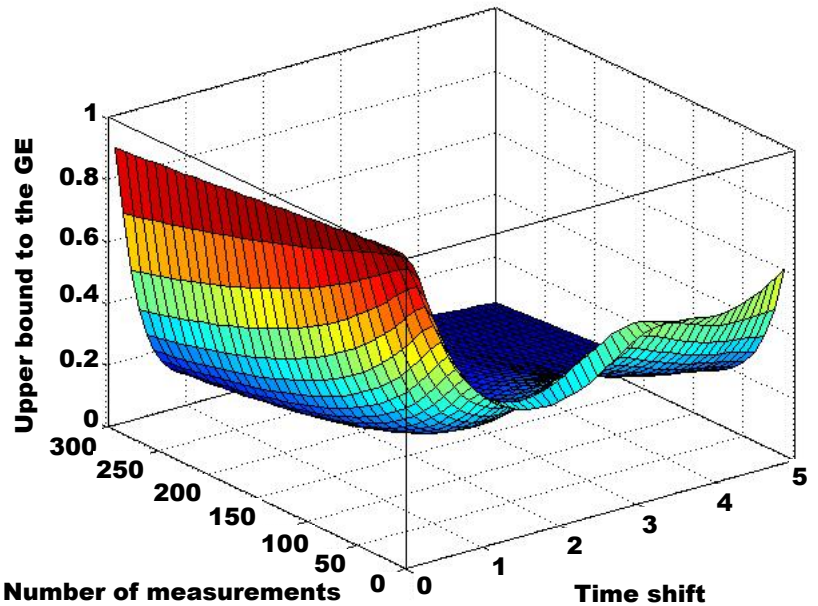

Figure S-2: Upper bound of the error rate with respect to time shift and number of repeated measurements. Simulation setting is the same as that in Section 4.1 except that the correlation has a block diagonal structure with parameter 0.85 and block size 5.