A Reproducing Kernel Hilbert Space Framework for Functional Classification

Peijun Sang\textsuperscript{a}, Adam B. Kashlak\textsuperscript{b} 	extsuperscript{a}, and Linglong Kong\textsuperscript{b} 	extsuperscript{c}

\textsuperscript{a}Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada; \textsuperscript{b}Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada

ABSTRACT

The intrinsic infinite-dimensional nature of functional data creates a bottleneck in the application of traditional classifiers to functional settings. These classifiers are generally either unable to generalize to infinite dimensions or have poor performance due to the curse of dimensionality. To address this concern, we propose building a distance-weighted discrimination (DWD) classifier on scores obtained by projecting data onto one specific direction. We choose this direction by minimizing, over a reproducing kernel Hilbert space, an empirical risk function containing the DWD classifier loss function. Our proposed classifier avoids overfitting and enjoys the appealing properties of DWD classifiers. We further extend this framework to accommodate functional data classification problems where scalar covariates are involved. In contrast to previous work, we establish a nonasymptotic estimation error bound on the relative misclassification rate. Through simulation studies and a real-world application, we demonstrate that the proposed classifier performs favorably relative to other commonly used functional classifiers in terms of prediction accuracy in finite-sample settings. Supplementary materials for this article are available online.

1. Introduction

In functional data classification, the explanatory variable is usually a random function and the outcome is a categorical random variable with two or more levels, called class labels. Similar to classification problems for scalar covariates, functional classifiers are built upon a collection of observations consisting of a functional covariate and a categorical response for each observation. Functional classifiers can then be used to assign a class label to new observations of the functional covariate. Functional classification has been extensively studied in the literature due to its wide applicability to various fields such as neuroscience, genetics, agriculture and chemometrics.

As pointed out in Fan and Fan (2008) for traditional classifiers, the high dimensionality of scalar predictors negatively impacts classifier predictive accuracy through the curse of dimensionality. This issue is even more serious in functional classification as functional data are intrinsically infinite dimensional. In light of this fact, dimension reduction prior to functional data classification has been suggested. Functional principal component (FPC) analysis is a commonly used technique for dimension reduction, and various FPC-based classifiers have previously been proposed, including discriminant analysis (Hall, Poskitt, and Presnell 2001), the naive Bayes classifier (Dai, Müller, and Yao 2017) and logistic regression (Müller and Stadtmüller 2005). Since FPC analysis is an unsupervised dimension reduction approach, the information contained within the retained FPC scores is not necessarily more predictive of the response than the discarded information. Consequently, treating functional data from the perspective of a Hilbert space, without dimension reduction, has attracted substantial attention in functional classification (Yao, Müller, and Wang 2005) and Cérou and Guyader (2006) considered $k$-nearest neighbor classification.

An optimal separating hyperplane can be constructed to distinguish between two perfectly separated classes. This idea is further extended to accommodate the nonseparable case by support vector machines (SVM). More specifically, with the aid of the so-called kernel trick, a linear boundary that separates the classes very well can be obtained in an expanded feature space. This induces a nonlinear boundary in the original feature space. For a more comprehensive introduction to the SVM, refer to Vapnik (2013) and Cristianini and Shawe-Taylor (2000). Due to SVM’s ability to construct flexible decision boundaries, different extensions for functional data have been proposed in the literature. Rossi and Villa (2006) considered first projecting functional covariates onto a set of fixed basis functions before applying SVM to the projections for classification. Wu and Liu (2013) recovered sparse functional data or longitudinal data trajectories using principal analysis by conditional expectation (PACE) (Yao, Müller, and Wang 2005) and proposed a support vector classifier for the recovered random curves. However, the misclassification error rate was not theoretically investigated in this work.
Marron, Todd, and Ahn (2007) noted that the data-piling problem may cause a deterioration in the performance of SVM classifiers. In response, they proposed the distance-weighted discrimination (DWD) classifier that makes uses of all observations in a training sample, rather than just the support vectors in SVM, to determine the decision boundary. Wang and Zou (2018) proposed an efficient algorithm to solve the DWD problem. In this article, we extend the concept of DWD to functional data to address a binary classification problem. The basic idea is to find a projection direction such that the DWD classifier built upon the corresponding projection scores achieves optimal prediction performance. Additionally, to avoid overfitting on the training sample, we incorporate a roughness penalty term when minimizing the empirical risk function. Penalized approaches have been investigated recently in the context of functional linear regression (Yuan and Cai 2010; Peña and Cai 2011). However, as far as we know, this framework has received little attention in the context of functional classification. The approach to functional classification proposed in this article estimates a slope function by minimizing a regularized empirical risk function over a reproducing kernel Hilbert space (RKHS). This RKHS is closely associated with the penalty term in the regularized empirical risk function. Through the representer theorem, we convert this infinite-dimensional minimization problem into a finite-dimensional one. This result lays the foundation for our numerical implementation of the proposed classifier.

We further extend our proposed framework to accommodate classification problems where the predictor variables include a functional covariate together with one or more scalar covariates. There has been extensive research on partial functional linear regression models in settings with functional and scalar predictors (e.g., Kong et al. 2016; Wong, Li, and Zhu 2019), although little progress has been made for classification. This article thus fills the gap in the literature regarding functional data classification with scalar covariates. More importantly, our proposed framework can accommodate other loss functions such as logistic loss and margin losses like exponential loss and hinge loss to yield various functional classifiers. Besides our novel methodology for regularized functional classification in an RKHS framework, we establish a nonasymptotic “oracle-type inequality” that bounds the convergence rate of the relative loss and relative classification error. These bounds are fundamentally different from those considered in Delaigle and Hall (2012), Dai, Müller, and Yao (2017), and Berrendero, Cuevas, and Torrecilla (2018), all of which focused on asymptotic perfect classification. Furthermore, our theoretical analysis provides a framework on studying a nonasymptotic error bound for functional data classifiers constructed through regularization in an RKHS. In fact, this theoretical framework actually reveals the essential difference between the RKHS-based functional linear regression proposed by Yuan and Cai (2010) and our proposed classifier. Yuan and Cai (2010) is concerned about the minimax convergence rate of the estimated slope function where the true function is assumed to reside in an RKHS. But our framework does not assume that there exists a true slope function. This explains why the convergence rate of prediction risk of an estimated classifier rather than estimated parameters is pursued in the context of classification.

Last but not least, when the loss function of the DWD classifier is considered, we find that the regularized empirical risk function to be minimized is not convex. Hence, a direct minimization algorithm is not available. To efficiently address this problem, we apply the majorization-minimization (MM) algorithm to construct a sequence of minimizers, which converges to the desirable minimizer of the regularized empirical risk function.

The rest of this article is organized as follows. In Section 2 we propose our framework of RKHS-based regularized functional classifiers for functional data, both with and without additional scalar covariates. We take the DWD loss function as an example for illustration. Theoretical properties of the proposed classifiers are established in Section 3. Simulation studies in Section 4 and a real-world neuroimaging analysis in Section 5 suggest that the proposed classifier has comparable performance relative to other commonly used classifiers. We summarize this work and discuss directions for future developments in Section 6. The detailed algorithm for fitting the classifier in Section 2.3, some extra results of numerical studies and all technical proofs are delegated to the supplementary materials.

2. Methodology

In this section, we first briefly outline the framework of classifying functional data through regularization in an RKHS. Then we focus on a typical example under this framework: functional DWD classifiers. To account for effects of scalar covariates in classification, we further extend this classifier to incorporate scalar covariates. A detailed algorithm is introduced to implement the classifier with scalar covariates in a finite sample.

Let $X$ denote a random function with a compact domain $I$ and let $Y$ be a binary outcome related to $X$. Without loss of generality, assume that $E \left\{ \int_I X^2(t) dt \right\} < \infty$ and $Y \in \{-1,1\}$. Suppose that the training sample consists of $(x_i, y_i), i = 1, \ldots, n$, that is, $n$ iid copies of $(X, Y)$. Given a sample path of $X$, a functional classifier $f$ assigns either $-1$ or $1$ to it. Let $f$ define a generic functional classifier and $\ell$ a loss function. A functional classifier can be estimated from minimizing empirical risk

$$\hat{f}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$$

with respect to $f$. In this article we focus on functional linear classifiers: that is, $f \in F = \{ f \mid f(x) = \alpha + \int_I x(t) \beta(t) dt \}$ for some $\alpha \in \mathbb{R}$ and $\beta \in L_2(I)$. Namely, $\alpha + \int_I x(t) \beta(t) dt$ determines the linear decision boundary for the function-valued covariate $x$, and $\text{sign}(\alpha)$ determines the label of $x = 0$ if the decision rule is defined by $\text{sign}(f)$ as in margin classifiers. Then the estimated functional classifier is defined as

$$\hat{f} \in \arg \min_{f \in F} \hat{R}(f) = \arg \min_{f \in F} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$$

In fact, the functional classifier defined above is constructed from first projecting an infinite-dimensional random function $x$ onto a specific direction (given by $\beta$), and then using classical empirical risk minimization. In this sense estimating the functional classifier is equivalent to estimating the projection direction. Both pre-determined basis functions like B-splines and data-driven bases such as FPCs have been considered for
this purpose in the literature; see Ratcliffe, Heller, and Leader (2002) and Müller and Stadtmüller (2005) for instance. Since FPCs are orthogonal to each other, using them to represent both the projection direction and the functional covariate can lead to an easy representation of \( \int_x x(t) \beta(t) dt \). However, there is no guarantee that the retained FPC scores are more predictive of the response than the discarded ones. Additionally, this estimation approach cannot control the roughness of the estimated slope function \( \beta \) very well due to the discrete manner in the dimension of the approximation, thus, leading to a poor between bias and variance tradeoff.

To address these issues, we consider the following framework: minimize a regularized empirical risk function

\[
Q(\alpha, \beta) = n^{-1} \sum_{i=1}^{n} \ell(\alpha + \int_I x_i(t) \beta(t) dt, y_i) + \lambda J(\beta),
\]

(1)

where \( J \) is a penalty functional which can be conveniently defined through the slope function \( \beta \) as a squared norm or semi-norm associated with \( H \), the space where the slope function \( \beta \) resides. A canonical choice of \( H \) is a Sobolev space. Without loss of generality, assuming that \( I = [0, 1] \), the Sobolev space of order \( m \) is defined as

\[
\mathcal{W}_2^m([0, 1]) = \{ h : [0, 1] \rightarrow \mathbb{R}, h^{(1)}, \ldots, h^{(m-1)} \}
\]

are absolutely continuous, \( h^{(m)} \in L_2(0, 1) \).

Endowed with the (squared) norm \( \| h \|^2_{\mathcal{W}_2^m} = \sum_{i=0}^{m-1} \left( \int_0^1 h^{(i)}(t) dt \right)^2 + \int_0^1 (h^{(m)}(t))^2 dt, \ \mathcal{W}_2^m([0, 1]) \) is an RKHS. In this case, one possible choice of the penalty functional is given by \( J(\beta) = \int_0^1 (\beta^{(m)}(t))^2 dt \). This idea is in spirits close to that of Du and Wang (2014), which employs RKHS regularization to fit generalized functional linear model. If \( \ell \) in (1) is taken as the loss function of logistic regression, it is a special case of Du and Wang (2014). However, we allow for other loss functions such as margin-based loss functions, which are different than the minus log-likelihood function adopted in Du and Wang (2014). Next we elaborate this framework with a particular loss function.

2.1. Functional DWD without Scalar Covariates

We first present an overview of DWD, originally proposed by Marron, Todd, and Ahn (2007). Consider the classification problem where \( z = (z_1, \ldots, z_p)^T \in \mathcal{Z} \) is a vector of \( p \) scalar covariates and \( y \in \{-1, 1\} \) is a binary response. The main task is to build a classifier: \( g : \mathcal{Z} \rightarrow \{-1, 1\} \) based on \( n \) pairs of training observations \((z_i, y_i), i = 1, \ldots, n\). According to Wang and Zou (2018), the decision boundary of the generalized DWD classifier can be obtained by solving \( \min_{\alpha, \beta} n^{-1} \sum_{i=1}^{n} V_q(y_i(\alpha_0 + z_i^T \alpha) + \lambda \beta^T \beta) \), where \( V_q(u) = 1 - u \) if \( u \leq \frac{q}{1+q} \) and \( \frac{q}{1+q} - \frac{q^2}{(q+1)^2} \) otherwise, is the loss function and \( \lambda > 0 \) is a tuning parameter. Note that as \( q \rightarrow \infty \), the generalized DWD loss function converges to the hinge loss function, \( H(u) = \max(0, 1 - u) \), used in SVM. This behavior is illustrated in Figure 1. Denote by \((\hat{\alpha}_0, \hat{\beta})\) the solution to the minimization problem above. Given a new observation \( z \in \mathcal{Z} \), the predicted class label will be 1 if \( \hat{\alpha}_0 + z^T \hat{\beta} > 0 \) and -1 otherwise.

In this article, we incorporate the concept of DWD in our framework for classifying functional data. To do so, the objective function in Equation (1) is replaced by

\[
Q(\alpha, \beta) = n^{-1} \sum_{i=1}^{n} V_q(y_i(\alpha + \int_I x_i(t) \beta(t) dt) + \lambda J(\beta)).
\]

(2)

Let the penalty functional \( J \) be a squared semi-norm on \( \mathcal{H} \). The null space \( \mathcal{H}_0 = \{ \beta \in \mathcal{H} : J(\beta) = 0 \} \) is a finite-dimensional linear subspace of \( \mathcal{H} \). Actually the penalty term \( J(\beta) \) satisfies this requirement if \( \mathcal{H} = \mathcal{W}_2^m([0, 1]) \). Denote by \( \mathcal{H}_1 \) the orthogonal complement of \( \mathcal{H}_0 \) in \( \mathcal{H} \) such that \( \mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1 \). That is, for any \( \beta \in \mathcal{H} \), there exists a unique decomposition \( \beta = \beta_0 + \beta_1 \) such that \( \beta_0 \in \mathcal{H}_0 \) and \( \beta_1 \in \mathcal{H}_1 \). Note that \( \mathcal{H}_1 \) is also an RKHS with an inner product the same as that of \( \mathcal{H} \), but restricted to \( \mathcal{H}_1 \).

Let \( K : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R} \) be the corresponding reproducing kernel of \( \mathcal{H}_1 \) such that \( J(\beta_1) = \| \beta_1 \|_K^2 = \| \beta_1 \|_H^2 \) for any \( \beta_1 \in \mathcal{H}_1 \). Let \( N = \dim(\mathcal{H}_0) \) and \( \psi_1, \ldots, \psi_N \) be basis functions of \( \mathcal{H}_0 \).

We assume that \( K \) is continuous and square integrable. With a slight abuse of notation, write \( (Kf)(s) = \int_I K(s,t) f(t) dt \).

According to Yuan and Cai (2010), \( Kf \in \mathcal{H}_1 \) for any \( f \in L_2(\mathcal{I}) \) and, for any \( \beta \in \mathcal{H}_1 \), \( \int_I \beta_1(t) f(t) dt = (Kf, \beta)_H \). With these observations, we are able to establish the following proposition, which is crucial in this work's subsequent numerical implementations and theoretical analyses.

**Proposition 1.** Let \( \hat{\alpha}_n \) and \( \hat{\beta}_n \) be the minimizer of (2) and \( \hat{\beta}_n \in \mathcal{H}_1 \). Then there exist \( d = (d_1, \ldots, d_N)^T \in \mathbb{R}^N \) and \( c = (c_1, \ldots, c_N)^T \in \mathbb{R}^n \) such that

\[
\hat{\beta}_n(t) = \sum_{i=1}^{N} d_i \psi_i(t) + \sum_{i=1}^{n} c_i (Kx_i)(t).
\]

(3)
For the purpose of illustration, assume $H = W_2^0$ and $f(\beta) = \int_0^1 \beta(t)^2 dt$ in the following numerical implementations. Then $H_0$ is the linear space spanned by $\psi_1(t) = 1$ and $\psi_2(t) = t - 0.5$. One possible choice for the reproducing kernel associated with $H_1$ is $K(s, t) = k_2(s)k_2(t) - k_4(s - t)$, where $k_2(s) = \frac{1}{2} (\psi_2^2(s) - \frac{1}{c})$ and $k_4(s) = \frac{1}{2} (\psi_2^2(s) - \frac{\psi_2(s)}{2} + \frac{7}{240})$. Refer to Chapter 2.3 of Gu (2013) for more details. By Proposition 1, we only need to consider $\beta(t) = d_1 + d_2(t - 0.5) + \sum_{i=1}^n c_i \int_0^1 x_i(t) K(s, t) ds$ for some $d = (d_1, d_2)^T \in \mathbb{R}^2$ and $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ when minimizing the function in (2). As a result, $\int_0^1 x_i(t) \beta(t) dt = d_1 \int_0^1 x_i(t) dt + d_2 \int_0^1 x_i(t)(t - 0.5) dt + \sum_{i=1}^n c_i \int_0^1 \int_0^1 x_i(t) K(s, t) x_i(s) ds dt$.

For the penalty term, we have that $f(\beta) = c^T R c$, where $R$ is an $n \times n$ matrix with $(i, j)$th entry $r_{ij} = \int_0^1 x_i(t) x_j(t) dt$ for $j = 1, 2$. Let $S_j$ and $R_j$ denote the $i$th row of $S$ and $R$, respectively. The infinite-dimensional minimization problem in (2) now reduces to the finite-dimensional problem

$$D(\alpha, d, c) = n^{-1} \sum_{i=1}^n V_q [y_i (\alpha + S'_i d + R'_i c)] + \lambda c^T R c. \quad (4)$$

To find a minimizer of $D$, we apply the MM algorithm. The basic idea is as follows. We first look for a majorization function $M(\theta h')$, where $\theta = (\alpha, d, c)^T$ in this problem, for the target function $D(\theta)$. This majorization function must satisfy $D(\theta) \leq M(\theta h')$ for any $\theta \neq h'$ and $D(h') = M(h')$ for $\theta = h'$. Computationally, it should be easy to find the minimizer of $M(\theta h')$ for any given $h'$. Then given an initial value of $\theta$, say, $\theta^{(0)}$, we can generate a sequence of $\theta$, say, $\theta^{(k)}_{\infty} = \arg \min_{\theta} M(\theta h^{(k)})$ for $k \geq 0$. Provided that this sequence converges, its limit is the minimizer of the objective function $D$.

Given $\theta = (\alpha, d, c)^T$, define $r = (r_1, \ldots, r_n)^T$ with $r_i = y_i V_q [y_i (\alpha + S'_i d + R'_i c)]/n$ and

$$A_{q,l} = \begin{pmatrix} n & 1_{n}^T S & 1_{n}^T R \\ S^T 1_n & S^T S & S^T R \\ R 1_n & R S & R R + \frac{2nq}{(q+1)} R \\ \end{pmatrix},$$

where $1_n$ denotes a vector of length $n$ with each component equal to 1. According to Lemma 2 of Wang and Zou (2018), one majorization function for $D$ is given by

$$M(\theta h') = \frac{1}{n} \sum_{i=1}^n V_q [y_i (\alpha + S'_i d + R'_i c)] + \lambda c^T R c$$

$$+ \begin{pmatrix} 1^T r \\ S^T r \\ R r + 2\lambda R c \\ \end{pmatrix}^T \begin{pmatrix} \alpha - \alpha' \\ d - d' \\ c - c' \\ \end{pmatrix}$$

$$+ \frac{(q + 1)^2}{2nq} \begin{pmatrix} \alpha - \alpha' \\ d - d' \\ c - c' \\ \end{pmatrix}^T A_{q,l} \begin{pmatrix} \alpha - \alpha' \\ d - d' \\ c - c' \\ \end{pmatrix}.$$

It is trivial to show that the minimizing of $M(\theta h')$ is $(\alpha, d, c)^T = -\frac{1}{mq_{(q+1)^2}} A_{q,l}^{-1}(1_n^T r, r^T S, (R r + 2\lambda R c)^T)^T$. The algorithm proceeds until the sequence of minimizers converges. Denoting the limit of this sequence by $(\hat{\alpha}, \hat{d}, \hat{c})^T$, it follows that $\beta(t) = \hat{d}_1 + \hat{d}_2(t - 0.5) + \sum_{i=1}^n \hat{c}_i (K(s_i)(t))$. The functional DWD classifier assigns 1 or -1 to a new functional observation $x$ according to whether the statistic $\alpha + \int_0^1 x(t) \beta(t) dt$ is positive or negative.

2.3. Functional DWD with Scalar Covariates

The previous algorithm addresses binary classification problems for univariate functional data, but can be further extended to accommodate binary classification when both a functional covariate and finitely-many scalar covariates are involved. In particular, the training sample consists of $(x_i, y_i, z_i)$, $i = 1, \ldots, n$, where $z_i = (z_{i1}, \ldots, z_{ip})^T$ denotes the $p$-dimensional vector of scalar covariates for the $i$th observation. With a slight abuse of notation, we consider an extension of (2) given by

$$Q(\alpha, \beta, \gamma) = \frac{1}{n} \sum_{i=1}^n V_q \left[ y_i \left[ \alpha + \int_0^1 x_i(t) \beta(t) dt + z_i^T \gamma \right] \right] + \lambda (\alpha, \beta, \gamma)^T.$$

To minimize (5), we resort to the specific representation of $\hat{\beta}$ given by Proposition 1. It is straightforward to verify that this result still holds in the context of (5). As a result, the infinite-dimensional problem of minimizing (5) reduces to the finite-dimensional problem of minimizing $D(\alpha, d, c, \gamma) = n^{-1} \sum_{i=1}^n V_q [y_i (\alpha + S'_i d + R'_i c + z_i^T \gamma)] + \lambda c^T R c$. More details on solving this minimization problem can be found in the supplementary materials.

2.4. Tuning Parameter Selection

We focus on the binary classification problem where both a functional covariate and several scalar covariates are used as predictors. The prediction performance of the proposed classifier depends on the choice of two tuning parameters, $q$ and $\lambda$. Computationally, it should be easy to find the minimizer of $M(\theta h')$ for any given $h'$. Then given an initial value of $\theta$, say, $\theta^{(0)}$, we can generate a sequence of $\theta$, say, $\theta^{(k)}_{\infty} = \arg \min_{\theta} M(\theta h^{(k)})$ for $k \geq 0$. Provided that this sequence converges, its limit is the minimizer of the objective function $D$.

Given $\theta = (\alpha, d, c)^T$, define $r = (r_1, \ldots, r_n)^T$ with $r_i = y_i V_q [y_i (\alpha + S'_i d + R'_i c + z_i^T \gamma)]/n$ and

$$A_{q,l} = \begin{pmatrix} n & 1_{n}^T S & 1_{n}^T Z \\ S^T 1_n & S^T S & S^T Z \\ 1_{n}^T R 1_n & 1_{n}^T Z 1_n & Z^T 1_n \\ R 1_n & R S & R Z \end{pmatrix} + \frac{2nq}{(q+1)^2} R R.$$
To compute the matrix, we need to find the inverse of $D$ first. Let $QAQ'$ denote the eigen-decomposition of $R$, which does not depend on $\lambda$. Then we compute the inverse of the diagonal matrix $\Pi^{\lambda}_{q,k}$ as $\Pi^{\lambda}_{q,k} = AA + (2nq^2/(q + 1)^2)A$ for each $q$ and $\lambda$, which can be easily computed for each combination of $q$ and $\lambda$. Furthermore, note that $D^{-1} = Q\Pi^{\lambda}_{q,k}Q'$ and that $B - C^T D^{-1} C$ is a $(3 + p) \times (3 + p)$ matrix, suggesting that the matrix inverse in (7) can be efficiently computed, provided that $p$ is relatively small.

Finally, we employ the form of $A_{q,k}^{-1}$ in (6) to compute

$$A^{-1}_{q,k} \left( \begin{array}{c} 1_n, S, Z \end{array} \right)^T_{\gamma r} \left( \begin{array}{c} 1_n, S, Z \end{array} \right)^T_{\gamma r} = \left( \begin{array}{cc} B^{-1} - B^{-1} C P B^{-1} & -B^{-1} C \end{array} \right) \left( \begin{array}{c} P \end{array} \right)$$

If $R$ is a $(3 + p) \times (3 + p)$ matrix, then we compute the inverse of the diagonal $X \in \mathbb{R}^{(3 + p) \times (3 + p)}$, the Bayes classifier $f^*$ is the loss function for the functional DWD classifier.

For an iid sample of pairs $(x_i(t), y_i)$ and the functional DWD loss function $V_q(\cdot)$, define the regularized estimator $\hat{\beta}$ as

$$\hat{\beta} = \arg \min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^n V_q \left( y_i \int_I x_i(t) \beta(t) dt \right) + \lambda A M J(\beta) \right\},$$

corresponding to the classifier $\hat{f}$. There are two settings to consider, similar to Blanchard, Bousquet, and Massart (2008), which affect how the penalization parameter is controlled. In setting (S1), the risk is considered via the spectral properties of the reproducing kernel. Specifically, the penalization parameter $\lambda_n$ is controlled by the tail sum of the eigenvalues of the reproducing kernel. In setting (S2), the risk is considered via covering numbers under the sup-norm. In contrast, $\lambda_n$ is controlled via $H^*_\infty(\varepsilon)$, the sup-norm $\varepsilon$-entropy. This control is encapsulated in the $\gamma(n)$ term defined in the following theorem.

**Theorem 1.** Under conditions (C1) and (C2), let the penalization parameter $\lambda_n$ be bounded as $\lambda_n \geq C \left\{ \gamma(n) + \frac{\log(\delta^{-1} \log n)^{1/2}}{n} \right\}$ for some universal constant $C$ with $\gamma(n) = \frac{1}{\sqrt{n}} \inf_{\beta \in \mathbb{V}_q^m} \int_0^\infty \sqrt{H^*_\infty(\beta, \varepsilon)} d\varepsilon = \sqrt{n x^2 / AM}$ with $B = \{ \beta \in \mathbb{V}_q^m : \| \beta \| \leq 1 \}$, then, for $f^*$ denotes the Bayes classifier and $\hat{\beta}$ the classifier corresponding to any arbitrary $\beta$,

$$L(\hat{f}, f^*) \leq 2 \inf_{\beta \in \mathbb{V}_q^m} \left\{ L(f^*, f^*) + 2\lambda_n A M J(\beta) + \lambda_n (k_1 + k_0) \right\},$$

holds with probability at least $1 - \delta$, for positive constants $k_0$ and $k_1$.

Conditions (C1) and (C2) and a proof of Theorem 1 can be found in the supplementary materials. We now extend this theorem to the functional DWD estimator with scalar covariates, $z$, in the following corollary. A proof of Corollary 1 follows from that for Theorem 1. Instead of considering superme over the ball $B(R) = \{ \beta \in \mathbb{V}_q^m : \| \beta \| \leq R \}$, we instead consider the product ball $B(R) = B(R) \times \{ y \in \mathbb{R}^d : \| y \| \leq R \}$. For an iid sample of triples $(x_i(t), z_i, y_i)$ and the functional DWD loss function $V_q(\cdot)$, let the regularized estimators $(\hat{\beta}, \hat{\gamma})$ be

$$(\hat{\beta}, \hat{\gamma}) = \arg \min_{\beta, \gamma} \left\{ \frac{1}{n} \sum_{i=1}^n V_q \left( y_i \int_I x_i(t) \beta(t) dt + z_i \gamma \right) \right\} + \lambda A^2 M J(\beta) \bigg\},$$

corresponding to the classifier $\hat{f}$.

**Corollary 1.** Under conditions (C1), (C2), and (C3), let $A^* = A + \Pi / M$ be a positive constant, and let the penalization parameter $\lambda_n$ be bounded as $\lambda_n \geq C \left\{ \gamma(n) + \frac{\log(\delta^{-1} \log n)^{1/2}}{n} \right\}$ for some universal constant $C$ with $\gamma(n) = \frac{1}{\sqrt{n}} \inf_{\beta \in \mathbb{V}_q^m} \int_0^\infty \sqrt{H^*_\infty(\beta, \varepsilon)} d\varepsilon = \sqrt{n x^2 / AM}$ with

$$42x^2$$
\( B = \{ \beta \in \mathcal{W}_T^p : \| \beta \| \leq 1 \} \times \{ \gamma \in \mathbb{R}^p : \| \gamma \| \leq 1 \}. \) Then, denoting \( f^* \) as the Bayes classifier and \( f_{\beta,\gamma} \) the classifier corresponding to any arbitrary \( \beta \) and \( \gamma \),

\[
L(\hat{f}, f^*) \leq 2 \inf_{\beta \in \mathcal{W}_T^p, \gamma \in \mathbb{R}^p} \left\{ L(f_{\beta,\gamma}, f^*) + 2\lambda_n A^* M f(\beta) + \lambda_n (k_1 + k_0) \right\},
\]

holds with probability at least \( 1 - \delta \), for positive constants \( k_0 \) and \( k_1 \).

### 4. Simulation Studies

In this section, we consider two different simulation settings to investigate the finite-sample performance of the proposed classifiers. In addition to the proposed functional DWD classifiers (to be abbreviated by fdwd without scalar covariates and plfdwd with scalar covariates), we also consider several other commonly used functional data classifiers for comparison. The centroid classifier in Delaigle and Hall (2012) (to be abbreviated by centroid) first projects the functional covariate onto one specific direction and then performs classification using distance to the centroid in the projected space. The functional quadratic discriminant in Galeano, Joseph, and Lillo (2015) (to be abbreviated by qda) conducts quadratic discriminant analysis on FPC scores while a functional logistic classifier (to be abbreviated by logit) fits a logistic regression model on these scores. The functional k-nearest neighbor (to be abbreviated by knn) leverages the \( L_2 \)-metric to define nearest neighbors. Note that the aforementioned classifiers, excepting our proposed functional DWD classifier with scalar covariates, only account for functional covariates. To study the effect of scalar covariates on classification, we also fit an SVM classifier with only these two scalar covariates when they are involved in the discriminant function (to be abbreviated by S-SVM).

In both settings, the functional covariate is generated as \( X_i(t) = \sum_{j=1}^{50} \xi_j \phi_j(t) \), where the \( \xi_j \)'s are independently drawn from a uniform distribution on \((-\sqrt{3}, \sqrt{3})\), \( \xi_j = (-1)^{j+1}j^{-1} \) for \( j = 1, \ldots, 50 \), \( \phi_1(t) = 1 \) and \( \phi_j(t) = \sqrt{2} \cos((j-1)\pi t), \) with \( j \geq 2 \) and \( t \in [0,1] \) are made available for each observation of \( X(t) \). Two scalar covariates, \( z = (z_1, z_2)^T \), are independently generated from a standard normal distribution truncated to the interval \((-2,2)\) with mean 0 and variance 1. The binary response variable \( y \) with values 1 or \(-1\) is generated from the logistic model

\[
f(X_i, z_i) = a_0 + \int_0^1 X_i(t)\beta(t)dt + z_i^T \gamma,
\]

\[
p(Y_i = 1) = \frac{\exp(f(X_i, z_i))}{1 + \exp(f(X_i, z_i))},
\]

where \( a_0 = 0.1 \) and \( f(X, z) \) is referred to as the discriminant function in this article.

In the first scenario, the slope function \( \beta(t) = e^{-t} \). The coefficient vector for the scalar effects is either \( \gamma = (-0.5, 1)^T \) or \( \gamma = (0, 0)^T \) to simulate settings where the discriminant function \( f \) either depends or does not depend, respectively, on the scalar covariates. In the second scenario, the slope function is represented in terms of the FPCs of \( X \). Particularly, \( \beta(t) = \sum_{j=1}^{50} (-1)^{j+1}j^{-2} \phi_j(t) \) and the coefficient vector for the scalar covariates is either \( \gamma = (-2, 3)^T \) or \( (0, 0)^T \). In each simulation scenario, \( n = 100 \) or 200 curves are generated for training. A test set of 500 samples are generated to assess prediction accuracy.

In each simulation trial, we randomly generate a training set of size \( n = 100 \) or 200 to fit all classifiers and then evaluate the predictive accuracy of each on a test set of size 500. To assess the uncertainty in estimating the prediction accuracy of each classifier, 50 independent simulation trials are conducted in each scenario.

Figure 2 displays boxplots of the misclassification error rate of each classifier in the first scenario. Obviously when \( \gamma = (-0.5, 1)^T \) in model (8), the proposed functional DWD classifier with scalar covariates is considerably more accurate than any other classifier in terms of prediction. This is not surprising since even the SVM classifier with only scalar covariates outperforms the functional classifiers that do not take scalar covariates into consideration. This fact implies the importance of accounting for scalar covariates when the true discriminant function indeed depends on them. Additionally, whether or not the true discriminant function depends on the scalar covariates in these settings, the medians of misclassification error rates of our proposed functional DWD classifiers are very close to the

![Figure 2](image-url)
Bayes errors, which are 0.283 and 0.376, respectively. As the projection function in the centroid classifier and the slope function in the functional logistic approach in Müller and Stadtmüller (2005) are represented in terms of FPCs, these two classifiers should be favored in the second scenario. This is justified by comparing the prediction accuracy of these classifiers, as shown in Figure S1. Nonetheless, our proposed classifier still dominates all competitors regardless of whether the true discriminant function depends on the scalar covariates. A plausible reason why the proposed classifier is superior to the centroid and logistic classifiers is that the roughness of the projection direction is appropriately controlled in our method. Once again, the misclassification rates of our proposed classifiers are very close to the Bayes errors of 0.086 and 0.099 with and without scalar covariates, respectively. Another interesting finding is that in contrast to relatively large training samples, our proposed classifier has a distinct advantage over these competitors in small samples.

5. Real Data Examples

In this section, we apply the proposed classifiers as well as several alternative classifiers to one real-world dataset to demonstrate the performance of our proposed approach.

Alzheimer’s disease (AD) is an irreversible and progressive brain disorder that can lead to increasingly serious dementia symptoms over only a few years. Previous studies have shown that increasing age is one of the most important risk factors of AD; most patients with AD are above 65. However, early-onset AD, occurring before the age of 65, is also a substantial concern. Together with the facts that there is currently no cure for AD and that the condition eventually destroys individuals’ abilities to perform even simple tasks, AD has understandably received considerable attention in recent years.

The present study uses data obtained from the ongoing Alzheimer’s Disease Neuroimaging Initiative (ADNI), which aims to unite researchers from around the world to collect, validate and analyze AD-related data. In particular, the ADNI is interested in identifying biomarkers of AD from genetic, structural, functional neuroimaging and clinical data. The data used in this section consist of two main parts. The first part consists of neuroimaging data collected via diffusion tensor imaging (DTI). More specifically, fractional anisotropy (FA) values were measured at 83 locations along the corpus callosum (CC) fiber tract for each subject. The second part is composed of demographic features such as gender (male or female), handedness (left- or right-handed), age, number of years of education, AD status, mini-mental state examination (MMSE) score and genotypes for apolipoprotein E $\epsilon-4$. AD status is a categorical variable with three levels: normal control (NC), mild cognitive impairment (MCI) and Alzheimer’s disease (AD). We focus on classifying subjects with the NC or AD status in the following analysis. According to http://adni.loni.usc.edu, E $\epsilon-4$ allele of APOE is the strongest known genetic risk factor for AD with a 2- to 3-fold increased risk for AD in people with one $\epsilon-4$ allele rising to about 12-fold in those with two alleles. MMSE score is one of the most widely-used tests of cognitive functions such as orientation, attention, memory, language and visual-spatial skills for assessing dementia severity. A more detailed description of the data can be found at http://adni.loni.usc.edu. Previous studies, such as Li, Huang, and Zhu (2017) and Tang et al. (2021), focused on building regression models to investigate the relationship between AD status progression and neuroimaging and demographic data. However, our main objective is to use DTI data and demographic and genetic features to predict AD status.

The brain imaging and demographic and genetic data from the baseline of the study were downloaded from the ADNI publicly available database at the website mentioned above. We include $N = 89$ subjects whose AD status is either NC or AD in our analysis after removing MCI subjects and subjects with missing values. Among the remaining subjects, $N_0 = 48$ subjects are from the first group, that is, their AD status is NC, and $N_1 = 41$ subjects are from the AD group. The functional predictor $X(t)$ is taken as the FA profiles, which are displayed for the entire dataset in Figure 3. The scalar covariates $z$ consist of gender, handedness, age, number of years of education and E $\epsilon-4$ allele. For an honest comparison between different classifiers, we also consider incorporating scalar covariates in functional logistic regression (to be abbreviated by pllogit) and functional K-nearest neighbor classification (to be abbreviated by pkknn). To compare the prediction performance of each classifier, the 89 subjects are randomly divided into a training set with $n$ subjects and a testing set with the other $N - n$ subjects. In the study, we consider two particular choices of $n$: 0.5$N$ and 0.8$N$. Following this rule, we randomly split the whole dataset into training and testing sets $M = 500$ times.

Table 1 summarizes the mean misclassification error rates and the corresponding standard errors across the 500 splits for each classifier. When scalar covariates are not accounted for in functional data classification, fDWD outperforms all competitors, excepting the functional logistic and the K-nearest neighbor classifiers, in terms of prediction accuracy. Even more remarkably, incorporating scalar covariates, even in a linear manner, results in a reduction in the misclassification error for
our proposed classifier. Another interesting find is that in contrast to our proposed method, when incorporating scalar covariates in either functional logistic regression or functional K-nearest neighbor classification, the prediction accuracy becomes worse. These comparisons indicate the superiority of our proposed classifier and, furthermore, suggest that appropriately accounting for scalar covariates can enhance prediction accuracy in functional data discrimination.

6. Conclusion

In this article we proposed a novel methodology that combines the concepts of the canonical DWD classifier and regularized functional linear regression under an RKHS framework to classify functional data. The use of the RKHS permits control over the roughness of the estimated projection direction, thus, enhancing prediction accuracy in comparison to conventional functional logistic regression and the centroid classifier. Additionally, in comparison to the commonly used dimension reduction via FPC analysis, the proposed framework greatly circumvent the issue of retaining less important information for the response. Moreover, we further extend the framework to classification problems involving both a functional covariate and several scalar covariates. Though we focused on a specific loss function to achieve the desirable properties of DWD classifiers in our study, our framework can be extended to other loss functions such as the logistic loss function in functional logistic regression and the hinge loss function used in functional SVM classifiers.

The proposed theoretical framework concerns the nonasymptotic error bound of the estimated classifier rather than the statistical properties of the slope function. It is worth studying both of these two properties when the Bayes classifier is given by a linear functional of a random function and the coefficient function resides in an RKHS. This problem can be essentially recast as $M$-estimation in functional data analysis. Additionally, the proposed classifiers can be easily extended to multiclass classification of functional data. Every time we pick two different labels in the training set to fit the proposed classifier and apply the majority vote to assign a label for a new observation.

Numerical studies including both simulation studies and one real-world neuroimaging application suggested that the proposed classifiers are superior to many other competitors in terms of predication accuracy. The application of our classifier to an AD dataset provided numerical evidence that both neuroimaging data and demographic features are relevant to AD, and that ignoring either of them negatively impacts prediction accuracy.

**Supplementary Materials**

The supplementary material contains extra results of numerical studies and the proofs of the lemma and theorems in the main manuscript.

**References**

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006), “Convexity, Classification, and Risk Bounds,” *Journal of the American Statistical Association*, 101, 138–156. [1004]

Berrendero, J. R., Cuevas, A., and Torrecilla, J. L. (2018), “On the Use of Reproducing Kernel Hilbert Spaces in Functional Classification,” *Journal of the American Statistical Association*, 113, 1210–1218. [1001]

Biau, G., Bunea, F., and Wegkamp, M. H. (2005), “Functional Classification in Hilbert Spaces,” *IEEE Transactions on Information Theory*, 51, 2163–2172. [1000]

Blanchard, G., Bousquet, O., and Massart, P. (2008), “Statistical Performance of Support Vector Machines,” *The Annals of Statistics*, 36, 489–531. [1004]

Cérrou, F. and Guyader, A. (2006), “Nearest Neighbor Classification in Infinite Dimension,” *ESAIM: Probability and Statistics*, 10, 340–355. [1000]

Cristianini, N., and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge: Cambridge University Press. [1000]

Dai, X., Müller, H.-G., and Yao, F. (2017), “Optimal Bayes Classifiers for Functional Data and Density Ratios,” *Biometrika*, 104, 545–560. [1000,1001]

Delailge, A., and Hall, P. (2012), “Achieving Near Perfect Classification for Functional Data,” *Journal of the Royal Statistical Society, Series B*, 74, 267–286. [1001,1005]

Du, P., and Wang, X. (2014), “Penalized Likelihood Functional Regression,” *Statistica Sinica*, 24, 1017–1041. [1002]

Fan, J., and Fan, Y. (2008), “High Dimensional Classification using Features Annealed Independence Rules,” *The Annals of Statistics*, 36, 2605–2637. [1000]

Galeano, P., Joseph, E., and Lillo, R. E. (2015), “The Mahalanobis Distance for Functional Data with Applications to Classification,” *Technometrics*, 57, 281–291. [1005]

Gu, C. (2013), *Smoothing Spline ANOVA Models* (2nd ed.), New York: Springer. [1003]

Hall, P., Poskitt, D. S., and Presnell, B. (2001), “A Functional Data-analytic Approach to Signal Discrimination,” *Technometrics*, 43, 1–9. [1000]

Kong, D., Xue, K., Yao, F., and Zhang, H. H. (2016), “Partially Functional Linear Regression in High Dimensions,” *Biometrika*, 103, 147–159. [1001]

Li, J., Huang, C., and Zhu, H. (2017), “A Functional Varying-Coefficient Single-Index Model for Functional Response Data,” *Journal of the American Statistical Association*, 112, 1169–1181. [1006]

Marron, J. S., Todd, M. J., and Ahn, J. (2007), “Distance-Weighted Discrimination,” *Journal of the American Statistical Association*, 102, 1267–1271. [1001,1002]

Müller, H.-G., and Stadtmüller, U. (2005), “Generalized Functional Linear Models,” *The Annals of Statistics*, 33, 774–805. [1000,1002,1006]

Ratcliff, S. J., Heller, G. Z., and Leader, L. R. (2002), “Functional Data Analysis with Application to Periodically Stimulated Foetal Heart Rate Data: II: Functional Logistic Regression,” *Statistics in Medicine*, 21, 1115–1127. [1002]

Rossi, F., and Villa, N. (2006), “Support Vector Machine for Functional Data Classification,” *Neurocomputing*, 69, 730–742. [1000]

Sun, X., Du, P., Wang, X., and Ma, P. (2018), “Optimal Penalized Function-on-Function Regression under a Reproducing Kernel Hilbert Space Framework,” *Journal of the American Statistical Association*, 113, 1601–1611. [1001]
Tang, Q., Kong, L., Ruppert, D., and Karunamuni, R. J. (2021), "Partial Functional Partially Linear Single-Index Models," *Statistica Sinica*, 31, 107–133. [1006]

Vapnik, V. (2013), *The Nature of Statistical Learning Theory*, New York: Springer. [1000]

Wang, B., and Zou, H. (2018), "Another Look at Distance-Weighted Discrimination," *Journal of the Royal Statistical Society*, Series B, 80, 177–198. [1000,1002,1003]

Wong, R. K., Li, Y., and Zhu, Z. (2019), “Partially Linear Functional Additive Models for Multivariate Functional Data,” *Journal of the American Statistical Association*, 114, 406–418. [1001]

Wu, Y., and Liu, Y. (2013), "Functional Robust Support Vector Machines for Sparse and Irregular Longitudinal Data," *Journal of Computational and Graphical Statistics*, 22, 379–395. [1000]

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [1000]

Yao, F., Wu, Y., and Zou, J. (2016), "Probability-Enhanced Effective Dimension Reduction for Classifying Sparse Functional Data," *Test*, 25, 1–22. [1000]

Yuan, M., and Cai, T. T. (2010), "A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression," *The Annals of Statistics*, 38, 3412–3444. [1000,1002]