Continuum centroid classifier for functional data

Zhiyang ZHOU1 and Peijun SANG2*

1Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA
2Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

Key words and phrases: Centroid classifier; continuum regression; functional linear model; functional partial least square; functional principal component.

MSC 2020: Primary 62R10; secondary 62H30.

Abstract: For the binary classification of functional data, we propose the continuum centroid classifier (CCC), which is constructed by projecting the functional data onto one specific direction. This direction is obtained via bridging the regression and classification. Our technique is neither unsupervised nor fully supervised; instead, we control the extent of the supervision. Thanks to the intrinsic infinite dimension of functional data, one of the two subtypes of CCC enjoys an (asymptotic) zero misclassification rate. Our approach includes an effective algorithm that yields a consistent empirical classifier. Simulation studies demonstrate the competitive performance of the CCC in different scenarios. Finally, we apply the CCC to two real examples.

1. INTRODUCTION

With the development of technology, data involving complex structures such as curves and images have become more and more frequent in various fields. Functional data analysis (FDA) handles functional data with measurements that take values over a continuum like space or time; typical examples of functional data include intraday yield curves for high-frequency trading, near-infrared spectra and blood signals measured in functional magnetic resonance imaging. For a comprehensive overview of FDA, see, e.g., Ramsay & Silverman (2005) and Ferraty & Vieu (2006).

An important problem in FDA is the classification for functional data, which is potentially applicable to, for instance, the diagnosis of multiple sclerosis (MS). MS is an immune disorder

* Corresponding author: peijun.sang@uwaterloo.ca

© 2021 Statistical Society of Canada
jeopardizing the central nervous system. Typical symptoms of MS include fatigue, vision problems, and sensory and cognitive impairment. Traditionally the diagnosis of MS relies on signs, symptoms and medical tests; yet, as reported by Sbardella et al. (2013), some important results on MS research have been obtained via applications of diffusion tensor imaging (DTI), since DTI is a powerful tool in quantifying the demyelination that results from MS (Goldsmith et al., 2012). If DTI results are recorded for both healthy controls and MS patients, the corresponding diagnosis can be interpreted as a problem of classifying curves into one of two states, i.e., a binary functional classification problem. More details about this particular application will be provided in Section 3.2.

The literature of FDA includes extensive research on functional data classification. Ferraty & Vieu (2003) proposed a kernel estimator of the posterior probability that a new curve belongs to a given class. Shin (2008) extended linear discriminant analysis (LDA) to functional data. The former technique defines a distance between curves, while the latter approach makes use of the reproducing kernel Hilbert space to define an inner product; both of them are constructed on the infinite dimensional space where functional data lie. As pointed out by Fan, Feng & Tong (2012), a large number of covariates would pose challenges for classification in the multivariate context. Functional data, however, are intrinsically infinite dimensional. In light of this fact, many researchers have suggested first reducing the dimension of the functional data and then implementing classical classification algorithms in the reduced space. Functional principal component (FPC) analysis is one of the most commonly used approaches for dimension reduction. Projections of functional data onto the directions of FPCs are referred to in the literature as FPC scores. Galeano, Joseph & Lillo (2015) proposed both linear and quadratic Bayes classifiers on FPC scores. Dai, Müller & Yao (2017) further extended this idea to density ratios of FPC scores and showed that the resulting classifier is equivalent to quadratic discriminant analysis (QDA) for Gaussian random curves. Additionally, a logistic regression on FPC scores was considered in Leng & Müller (2005) to discriminate temporal gene regulation data. Rossi & Villa (2006) implemented the support vector machine on FPC scores.

Our work is mainly motivated by the centroid classifier proposed by Delaigle & Hall (2012) who suggested projecting functional data to a specific direction first. To identify this direction, these authors converted the classification problem to a functional linear regression problem; the slope function in a functional linear model is taken as the desired projection direction. They proposed two methods of estimating the slope function: one that uses FPC basis functions and the other that uses the functional partial least squares (FPLS) basis functions. The first method accounts for the variability of functional covariates but ignores the information provided by the outcome, whereas the second approach focuses on the covariance between the outcome and the functional covariate. As argued in Jung (2018), the projection direction from FPC basis sometimes fails to capture the difference between mean functions, while the FPLS basis may be sensitive and vulnerable to small signals; a combination of these two ideas may be preferable to either of them. Therefore, we propose a new method for estimating the projection direction, one which captures both the variability of the functional data and the covariance between the functional covariate and the outcome. A distance-based centroid classifier is then employed after projecting functional data onto this specific direction. Under regularity conditions, we verify the asymptotic property of perfect classification. Numerical studies demonstrate that our proposed classifier compares favourably with competitors in terms of finite-sample classification accuracy.

The rest of this article is organized as follows. In Section 2, we introduce the two subtypes of our classifier, including both the population and empirical versions, and then establish the property of (asymptotically) perfect classification. Section 3 investigates the performance of our proposal, highlighting the settings that favour its use. Section 4 makes a discussion on the values of hyperparameters and the functional classification with more than two categories. All technical
details are relegated to the Appendix. Our code trunks are publicly available at https://github .com/ZhiyangGeeZhou/CCC.

2. METHODOLOGY AND THEORY

2.1. Formalization of the Problem

Suppose that \((X_1, Y_1), \ldots, (X_N, Y_N)\) are \(N\) independent and identically distributed (iid) copies of \((X, Y)\), where \(X\) is a random function defined on the interval \(\mathcal{T} = [t_{\text{min}}, t_{\text{max}}]\), and \(Y\) is the label of \(X\) taking values 0 or 1. In other words, each \(X_i\) is sampled from a mixture of two subpopulations, \(\Pi_0\) and \(\Pi_1\), with its origin identified via the indicator \(Y_i = 1(X_i \in \Pi_1)\). Of particular interest is the binary classification for a newly observed \(X^*\) distributed as \(X\) but independent of \(X_1, \ldots, X_N\). Assume that the sub-mean and sub-covariance functions for \(\Pi_k\), \(k = 0, 1\), are

\[
\mu_{[k]}(t) = \mathbb{E}[X(t) \mid Y = k]
\]

respectively, and for all \(s, t \in \mathcal{T}\),

\[
v_X^{[k]}(s, t) = \text{cov}\{X(s), X(t) \mid Y = k\}. \tag{1}
\]

Let \(\pi_0 = \Pr(X_i \in \Pi_0) \in (0, 1)\). We have the following decomposition

\[
\mu(\cdot) = \mathbb{E}[X(\cdot)] = \pi_0 \mu_{[0]}(\cdot) + (1 - \pi_0) \mu_{[1]}(\cdot)
\]

as well as

\[
v_X(s, t) = \text{cov}\{X(s), X(t)\} = v_X^W(s, t) + v_X^B(s, t), \tag{2}
\]

according to the law of total covariance, where

\[
v_X^W(s, t) = \pi_0 v_X^{[0]}(s, t) + (1 - \pi_0) v_X^{[1]}(s, t), \tag{3}
\]

\[
v_X^B(s, t) = \pi_0 (1 - \pi_0) \{\mu_{[1]}(s) - \mu_{[0]}(s)\} \{\mu_{[1]}(t) - \mu_{[0]}(t)\}
\]

are the within- and between-group covariance functions, respectively. Define the corresponding covariance operators \(v_X, v_X^{[k]} : L^2(\mathcal{T}) \to L^2(\mathcal{T})\), \(k = 0, 1\), such that, for \(f \in L^2(\mathcal{T})\),

\[
\mathcal{V}_X(f)(\cdot) = \int_{\mathcal{T}} f(s)v_X(s, \cdot)ds,
\]

\[
\mathcal{V}_X^{[k]}(f)(\cdot) = \int_{\mathcal{T}} f(s)v_X^{[k]}(s, \cdot)ds.
\]

Throughout this article, we abbreviate the Lebesgue integral \(\int_{\mathcal{T}} f(t)dt\) to \(\int_{\mathcal{T}} f\). Functions appearing in this article are limited to \(L^2(\mathcal{T})\) (or \(L^2(\mathcal{T}^2)\)), the collection of square integrable functions defined on \(\mathcal{T}\) (or \(\mathcal{T}^2\)). The square-integrability of \(v_X\) in Equation (2) (or \(v_X^{[k]}\) in Eq. 1) implies that the function possesses only a countable number of nonnegative eigenvalues \(\{\lambda_1, \lambda_2, \ldots\}\) (or \(\{\lambda_{k,1}, \lambda_{k,2}, \ldots\}\)), as appropriate) with corresponding eigenfunctions \(\{\phi_1, \phi_2, \ldots\}\) (or \(\{\phi_{k,1}, \phi_{k,2}, \ldots\}\)). In addition, \(\| \cdot \|\) denotes the \(L^2\)-norm, i.e., \(\|f\|\) equals \((\int_{\mathcal{T}} f^2)^{1/2}\) for \(f \in L^2(\mathcal{T})\) and \((\int_{\mathcal{T}} \int_{\mathcal{T}} f^2)^{1/2}\) if \(f \in L^2(\mathcal{T}^2)\).

2.2. Review of the Centroid Classifier

Before introducing our proposed approach, we first review the centroid classifier proposed by Delaigle & Hall (2012). They projected functional data onto the one-dimensional space spanned
by a given function $\omega \in L^2(\mathcal{T})$, say $\text{span}(\omega)$, and then constructed the classifier based upon the projection. Specifically, they defined a classifier

$$D(X^* \mid \omega) = \left\{ \int T \frac{\omega}{\|\omega\|} (X^* - \mu_{[1]}) \right\}^2 - \left\{ \int T \frac{\omega}{\|\omega\|} (X^* - \mu_{[0]}) \right\}^2 + 2 \ln \frac{\pi_0}{1 - \pi_0}, \ (4)$$

where $\int T \omega(X^* - \mu_{[k]})/\|\omega\|$, the magnitude of the projection of $X^* - \mu_{[k]}$ onto $\text{span}(\omega)$, can be regarded as the distance from $X^*$ to $\mu_{[k]}$. When $D(X^* \mid \omega)$ is positive, $X^*$ is thought to be closer to $\mu_{[0]}$ and hence is assigned to $\Pi_0$, and vice versa. Given $\omega$, this principle is identical to the LDA, assuming that, for each $k$, $\int T X \omega$ conditional on $X \in \Pi_k$ is normally distributed with $\text{var}(\int T X \omega \mid X \in \Pi_k) = \|\omega\|^2$, viz. $\int T X \omega \mid X \in \Pi_k \sim \mathcal{N}\left(\int T \mu_{[k]} \omega, \|\omega\|^2\right)$.

It remains to select a direction $\omega \in L^2(\mathcal{T})$ to minimize the misclassification rate

$$\text{err}\{D(X^* \mid \omega)\} = \pi_0 \text{Pr}\{D(X^* \mid \omega) < 0 \mid X^* \in \Pi_0\} + (1 - \pi_0) \text{Pr}\{D(X^* \mid \omega) > 0 \mid X^* \in \Pi_1\}.$$ 

Delaigle & Hall (2012) proposed taking $\omega = \beta_{p,FPC}$ in Equation (5) (or $\beta_{p,FPLS}$ in Eq. 6), corresponding to FPC basis concentrated on $v^w_X$ in Equation (3) (or FPLS basis), with a positive integer $p$ tuned via cross-validation. The resulting classifier $D(X^* \mid \beta_{p,FPC})$ (or $D(X^* \mid \beta_{p,FPLS})$) is referred to as PCC (or PLCC). More specifically, these two projection directions are defined as

$$\beta_{p,FPC} = \arg \min_{\beta \in \text{span}(\phi^w_1, \ldots, \phi^w_p)} \mathbb{E} \left[ Y - \mathbb{E}(Y) - \int T \beta \{X - \mathbb{E}(X)\} \right]^2 \quad (5)$$

and

$$\beta_{p,FPLS} = \arg \min_{\beta \in \text{span}(w^1_{FPLS}, \ldots, w^p_{FPLS})} \mathbb{E} \left[ Y - \mathbb{E}(Y) - \int T \beta \{X - \mathbb{E}(X)\} \right]^2, \quad (6)$$

where $\phi^w_1, \ldots, \phi^w_p$ are the first $p$ eigenfunctions of $v^w_X$; and $w^1_{FPLS}, \ldots, w^p_{FPLS}$ are solutions to the following sequential optimization problems: Given $w^1_{FPLS}, \ldots, w^j_{FPLS}$, define

$$w^j_{FPLS} = \arg \max_{w: \|w\|=1} \text{cov}^2 \left( Y, \int T Xw \right)$$

subject to $\int T w^i_{FPLS} v^w_X = 0, \ i = 1, \ldots, j - 1$. The minimizers identified in Equations (5) and (6), yet restricted in different linear spaces, are both slope functions of the (constrained) optimal approximations to $Y$ by a linear functional of $X$. Taking either of them as $\omega$ in Equation (4), Delaigle & Hall (2012) succeeded in bridging the binary classification problem to functional linear regression.

2.3. Continuum Centroid Classifier

As mentioned in Section 1, an intermediate state between the FPC and FPLS bases may be preferred. Fixing $\alpha \in [0, 1)$, Zhou (2019, Proposition 4) defined the functional continuum (FC) basis functions as sequential constrained maximizers of

$$T_{j,\alpha}(w) = \text{cov}^2 \left( Y^{l,\alpha}, \int T X^{l,\alpha}w \right), \text{var}^{-1+\frac{\alpha}{2}} \left( \int T X^{l,\alpha}w \right), \quad (7)$$

DOI: 10.1002/cjs.11624 The Canadian Journal of Statistics / La revue canadienne de statistique
where 
\[ X^{[j,\alpha]} = X^{[j-1,\alpha]} - \text{var}^{-\frac{1}{2}} \left( \int_X X w_{j-1,\alpha} \right) \cdot \left( \int_X X^{[1,\alpha]} w_{j-1,\alpha} \right) \cdot \psi_X \left( w_{j-1,\alpha} \right), \]
and, for \( j = 2, 3, \ldots, \)
\[ Y^{[j,\alpha]} = Y^{[j-1,\alpha]} - \mathbb{E}(Y) - \int_X X^{[1,\alpha]} \beta_{j-1,\alpha}, \]
with \( X^{[1,\alpha]} = X - \mathbb{E}(X) \) and \( Y^{[1,\alpha]} = Y - \mathbb{E}(Y) \). Specifically, given \( w_{1,\alpha}, \ldots, w_{j-1,\alpha} \), the \( j \)th FC basis function is
\[ w_{j,\alpha} = \arg \max_{w : \|w\| = 1} T_{j,\alpha}(w). \] (8)

The FC basis captures not only variation in \( X \) but also the covariance between \( X \) and \( Y \); thus, this dimension reduction technique lies midway between (unsupervised) FPC and (fully supervised) FPLS. The FC basis reduces to FPLS when \( \alpha = 1/2 \) and becomes irrelevant to \( Y \) as \( \alpha \) diverges. By tuning \( \alpha \), the FC basis controls the extent of supervision and is expected to be neither unsupervised nor too supervised.

Our continuum centroid classifier (CCC) is defined by substituting
\[ \beta_{p,\alpha} = \arg \min_{\beta \in \text{span}(w_{1,\alpha}, \ldots, w_{p,\alpha})} \mathbb{E} \left[ Y - \mathbb{E}(Y) - \int_T \beta \{ X - \mathbb{E}(X) \} \right]^2 \] (9)
for \( \omega \) in Equation (4) and, simultaneously, dropping the assumption \( \text{var}(\int_T X \omega \mid X \in \Pi_k) = \|\omega\|^2 \); the CCC assigns a label to a random trajectory \( X^* \) by applying LDA and QDA, respectively, to \( \int_T X^* \beta_{p,\alpha} \) and hence includes two subtypes, say CCC-L and CCC-Q. In particular, the CCC-Q is given by
\[ D_Q(X^* \mid \beta_{p,\alpha}) = \sigma_{[1]}^{-2}(\beta_{p,\alpha}) \left\{ \int_T \beta_{p,\alpha}(X^* - \mu_{[1]}) \right\}^2 \]
\[ - \sigma_{[0]}^{-2}(\beta_{p,\alpha}) \left\{ \int_T \beta_{p,\alpha}(X^* - \mu_{[0]}) \right\}^2 + 2 \ln \frac{\pi_0 \sigma_{[1]}(\beta_{p,\alpha})}{(1 - \pi_0) \sigma_{[0]}(\beta_{p,\alpha})} \]
(10)
with
\[ \sigma_{[k]}^2(\omega) = \text{var} \left( \int_T X \omega \mid X \in \Pi_k \right) \]
(11)
for each \( k \). If one believes \( \sigma^2(\omega) = \sigma_{[0]}^2(\omega) = \sigma_{[1]}^2(\omega) \), the CCC-Q reduces to CCC-L which corresponds to
\[ D_L(X^* \mid \beta_{p,\alpha}) = \left\{ \int_T \beta_{p,\alpha}(X^* - \mu_{[1]}) \right\}^2 \]
\[ - \left\{ \int_T \beta_{p,\alpha}(X^* - \mu_{[0]}) \right\}^2 + 2 \sigma^2(\beta_{p,\alpha}) \ln \frac{\pi_0}{1 - \pi_0}. \]
(12)

Analogous to the classifier (4), positive \( D_L(X^* \mid \beta_{p,\alpha}) \) (or \( D_Q(X^* \mid \beta_{p,\alpha}) \)) suggests classifying \( X^* \) to \( \Pi_0 \), and vice versa.

As long as condition (C1) in the Appendix is satisfied, in theory one can expect that the CCC-L will provide an asymptotically perfect classification (Proposition 1). It is worth emphasizing that Proposition 1 does not require normality or any specific variance structure for the two subpopulations.
**Proposition 1.** If condition (C1) holds, the CCC-L asymptotically leads to no misclassification as \( p \) diverges.

### 2.3.1. Empirical implementation

In general, it is impossible to observe entire trajectories. In this sense, the procedure for estimating \( \beta_p, \alpha \) (at Eq. 9) was not sufficiently detailed by Zhou (2019). We depict this procedure with more details in what follows.

For brevity, we assume the \( X_i \)'s are densely digitized on \( M + 1 \) equispaced time points \( t_m = t_{\min} + (m - 1)\Delta t, m = 1, \ldots, M + 1 \), with \( \Delta t = (t_{\max} - t_{\min})/M \). Reformulating the infinite-dimensional optimization problem (8) to a finite-dimensional one, we apply penalized cubic B-spline smoothing (Ramsay & Silverman, 2005, Sections 5.2.4 and 5.2.5) to each trajectory, i.e., seeking for a surrogate of \( X_i \) in the \( L \)-dimensional linear space (where \( L = M + 3 \) as recommended by Ramsay & Silverman, 2005, p. 86)

\[
BS_L = \text{span}(\psi_1, \ldots, \psi_L) \tag{13}
\]

which is spanned by the cubic B-splines \( \psi_1, \ldots, \psi_L \); see de Boor (2001, Chapter 4) for more details on B-splines. Specifically, the (fully) recovered estimator for the \( i \)th trajectory, \( i = 1, \ldots, N \), is given by

\[
\hat{X}_i = \hat{c}_i^T \Psi, \tag{14}
\]

where

\[
\Psi = \Psi(\cdot) = [\psi_1(\cdot), \ldots, \psi_L(\cdot)]^T, \tag{15}
\]

\[
\hat{c}_i = (\Psi^T \Psi + \theta_0 \text{Pen})^{-1} \Psi^T X_i, \tag{16}
\]

with

\[
\Psi = [\psi_k(t_m)]_{(M+1)\times L} = [\psi(t_1), \ldots, \psi(t_{M+1})]^T, \tag{17}
\]

\[
\text{Pen} = \left[ \int_I \psi_{i_1}'' \psi_{i_2}'' \right]_{1 \leq i_1, i_2 \leq L}, \tag{18}
\]

\[
X_i = [X_i(t_1), \ldots, X_i(t_{M+1})]^T.
\]

The positive smoothing parameter \( \theta_0 \) is common for all \( i \) and is picked up automatically through generalized cross-validation (GCV, Craven & Wahba, 1979), i.e.,

\[
\theta_0 = \arg \min_{\theta} \frac{\sum_{i=1}^N \sum_{m=1}^{M+1} (X_i(t_m) - \hat{X}_i(t_m))^2}{[M + 1 - \text{trace}(\Psi^T \Psi + \theta \text{Pen})^{-1} \Psi^T)]^2}.
\]

Thanks to dense observations, the presmoothing is able to recover the underlying curves accurately (in the \( L^2 \) sense) under some regularity conditions. Also, presmoothed curves ought to enjoy higher classification accuracy (Dai, Müller & Yao, 2017).

**Proposition 2.** If one has conditions (C2) and (C3) in the Appendix, for each \( i \), \( \|\hat{X}_i - X_i\| \) is zero-convergent in probability as the number of observation time points \( M + 1 \) goes to infinity.

The empirical version of \( T_{j,\alpha} \) (at Eq. 7), say \( \hat{T}_{j,\alpha} \), is then constructed by replacing \( X_i \), \( \text{var}(\cdot) \), and \( \text{cov}(\cdot) \) with their respective empirical counterparts or, equivalently, by substituting \( \hat{X}_i \) from Equation (14) for occurrences of \( X_i \) in Zhou (2019, Proposition 5).
Proposition 3. Suppose \( \hat{w}_{j,a} = \max_{w: \|w\| = 1} \hat{T}_{j,a}(w) \) is an estimator of \( w_{j,a} \) specified at Equation (8). Then \( \hat{w}_{j,a} \) must lie in the linear space \( BS_r \) defined at (13).

We begin by optimizing \( \hat{T}_{1,a}(w) \). Proposition 3 shrinks our search scope to \( \{w : w = b^T W^{-1/2} \psi, b^T b = 1, b \in \mathbb{R}^{L \times 1}\} \) with \( \psi \) defined at (15) and invertible and symmetric

\[
W = \left[ \int_{\mathbb{T}} \psi_{i_1}\psi_{i_2} \right]_{1 \leq i_1 i_2 \leq L}.
\]

(19)

The maximization of \( \hat{T}_{1,a}(w) \) (subject to \( \|w\| = 1 \)) is reformulated as an \( L \)-dimensional optimization problem

\[
\max_{b \in \mathbb{R}^{L \times 1}} (b^T W^{1/2} \hat{\mathcal{C}}_c Y_c)^2 (b^T W^{1/2} \hat{\mathcal{C}}_c W^{1/2} b)^{a/(1-a)-1}
\]

subject to \( b^T b = 1 \), where \( \hat{\mathcal{C}}_c \) is defined in Equation (16),

\[
\hat{\mathcal{C}}_c = \left[ \hat{\mathcal{C}}_c^1, \ldots, \hat{\mathcal{C}}_c^N \right]^T,
\]

\[
Y_c = \left[ Y_1^T - \frac{1}{N} \sum_{i=1}^N Y_i, \ldots, Y_N - \frac{1}{N} \sum_{i=1}^N Y_i \right]^T.
\]

Note that the solution to the optimization problem (20) is necessarily located in the row space of \( \hat{\mathcal{C}}_c W^{1/2} \), i.e., the search region is further restricted to \( \{w : w = b^T V^T W^{-1/2} \psi, b^T b = 1, b \in \mathbb{R}^{r \times 1}\} \), where \( r = \text{rank}(\hat{\mathcal{C}}_c W^{1/2}) \leq \min\{L, N\} \) and the \( L \times r \) matrix \( V \) comes from the thin singular value decomposition of \( \hat{\mathcal{C}}_c W^{1/2} \); that is, \( \hat{\mathcal{C}}_c W^{1/2} \) is decomposed into \( URV^T \) with an invertible diagonal matrix \( R \) and semi-orthogonal matrices \( U \) and \( V \) such that \( U^T U = V^T V = I_r \).

In this way, the dimension of maximization problem (20) further reduces to \( r \).

Write \( G_1 = UR \). The estimator of the first FC basis function then takes the form \( \hat{w}_{1,a} = b_{1,a}^T V^T W^{-1/2} \psi \) in which

\[
b_{1,a} = \arg \max_{b \in \mathbb{R}^{r \times 1} : b^T b = 1} (b^T G_1 Y_c)^2 (b^T G_1^T G_1 b)^{a/(1-a)-1}.
\]

(21)

Subsequently and successively, for \( j \geq 2 \), given \( j - 1 \) vectors \( b_{1,a}, \ldots, b_{j-1,a} \), we just need to replace the previous \( G_1 \) with the deflated \( G_j = P_{j-1} G_1 \), where \( P_0 = I_N \) and \( P_{j-1} = I_N - H_{j-1} (H_{j-1}^T H_{j-1})^{-1} H_{j-1}^T \) is the projection matrix associated with the orthogonal complement of column space of \( H_{j-1} = \hat{\mathcal{C}}_c W^{1/2} [V b_{1,a}, \ldots, V b_{j-1,a}] = UR[b_{1,a}, \ldots, b_{j-1,a}] \).

Thus, for all \( j \),

\[
\hat{w}_{j,a} = b_{j,a}^T V^T W^{-1/2} \psi,
\]

(21)

with

\[
b_{j,a} = \arg \max_{b \in \mathbb{R}^{r \times 1} : b^T b = 1} (b^T G_j Y_c)^2 (b^T G_j^T G_j b)^{a/(1-a)-1}
\]

(22)

\[
= \left\{ Y_c^T G_j (G_j^T G_j + \delta_{j,a}^{-1} \xi_{j,a} I_p)^{-2} G_j Y_c \right\}^{-1/2} (G_j^T G_j + \delta_{j,a}^{-1} \xi_{j,a} I_p)^{-1} G_j Y_c
\]

(23)
in which $\zeta_{j,a}$ is the largest eigenvalue of $G_j^\top G_j$. The ridge-type estimator (23) is deduced from Björkström & Sundberg (1999, Proposition 2.1). The only unknown $\delta_{j,a}$ appearing in (23) is the local maximizer in $(-1, \infty) \setminus \{0\}$ of the univariate function

$$Q_{j,a}(\delta) = \left\{ Y_c^\top G_j(G_j^\top G_j + \delta^{-1} \zeta_{j,a} I_r)^{-1} G_j^\top Y_c \right\}^2 \left\{ Y_c^\top G_j(G_j^\top G_j + \delta^{-1} \zeta_{j,a} I_r)^{-1} G_j^\top Y_c \right\}^{a-1} \times \left\{ Y_c^\top G_j(G_j^\top G_j + \delta^{-1} \zeta_{j,a} I_r)^{-1} G_j^\top G_j(G_j^\top G_j + \delta^{-1} \zeta_{j,a} I_r)^{-1} G_j^\top Y_c \right\}^{a-1},$$

where $Q_{j,a}(\delta)$ is obtained by substituting (23) for $b$ at the right-hand side of Equation (22). Thus, $\delta_{j,a}$ can be determined via an arbitrary computer algebra system. Fixing $p$, we then proceed to an estimator for $\hat{\beta}_{p,a}$ at Equation (9), namely

$$\hat{\beta}_{p,a} = [\hat{w}_{1,a}, \ldots, \hat{w}_{p,a}] (H_p^\top H_p)^{-1} H_p^\top Y_c$$

$$= \psi^\top W^{-\frac{1}{2}} V [b_{1,a}, \ldots, b_{p,a}] (H_p^\top H_p)^{-1} H_p^\top Y_c.$$

**Remark 1.** Despite the possible ambiguity in representing $b_{j,a}$ as the right-hand side of Equation (23), the consistency of $\hat{w}_{j,a}$ is not affected, as long as condition (C5) in the Appendix holds; refer to Zhou (2019, Remark 1).

Analogous to the $X_i$’s, the trajectory to be assigned, $X^*$, is discretely observed and hence is estimated by

$$\hat{X}^* = \hat{e}^* \psi,$$

where $\hat{e}^*$ is available by applying B-spline smoothing to $X^*(t_1), \ldots, X^*(t_M)$. Let $N_0$ (or $N_1$) denote the number of training trajectories belonging to $\Pi_0$ (or $\Pi_1$). If we estimate the mean functions $\mu_{[k]}$ by

$$\hat{\mu}_{[k]} = \frac{1}{N_k} \sum_{i=1}^N \hat{X}_i \mathbb{1}(X_i \in \Pi_k) = \frac{1}{N_k} \sum_{i=1}^N \hat{e}_i^* \psi \mathbb{1}(X_i \in \Pi_k),$$

the empirical CCC-Q and -L are then given by

$$\hat{D}_Q(\hat{X}^* | \hat{\beta}_{p,a}) = \hat{\sigma}_{[1]}^{-2}(\hat{\beta}_{p,a}) \left\{ \int \hat{\beta}_{p,a}(\hat{X}^* - \hat{\mu}_{[1]}) \right\}^2$$

$$- \hat{\sigma}_{[0]}^{-2}(\hat{\beta}_{p,a}) \left\{ \int \hat{\beta}_{p,a}(\hat{X}^* - \hat{\mu}_{[0]}) \right\}^2 + 2 \ln \frac{N_0 \hat{\sigma}_{[1]}(\hat{\beta}_{p,a})}{N_1 \hat{\sigma}_{[0]}(\hat{\beta}_{p,a})}$$

(24)

and

$$\hat{D}_L(\hat{X}^* | \hat{\beta}_{p,a}) = \hat{\sigma}_{\text{pool}}^{-2}(\hat{\beta}_{p,a}) \left\{ \int \hat{\beta}_{p,a}(\hat{X}^* - \hat{\mu}_{[1]}) \right\}^2$$

$$- \hat{\sigma}_{\text{pool}}^{-2}(\hat{\beta}_{p,a}) \left\{ \int \hat{\beta}_{p,a}(\hat{X}^* - \hat{\mu}_{[0]}) \right\}^2 + 2 \ln \frac{N_0}{N_1},$$

(25)

DOI: 10.1002/cjs.11624
respectively, where for $k = 0, 1$,
\[
\hat{\sigma}^2_{\text{pool}}(\omega) = (N - 2)^{-1} \sum_{k=0}^{1} \sum_{i=1}^{N} \left\{ \int_{R} \omega(X_i - \hat{\mu}_{[k]}) \right\}^2 1(X_i \in \Pi_k),
\]
\[
\hat{\sigma}^2_{[k]}(\omega) = (N_k - 1)^{-1} \sum_{i=1}^{N} \left\{ \int_{R} \omega(X_i - \hat{\mu}_{[k]}) \right\}^2 1(X_i \in \Pi_k).
\]

**Proposition 4.** Fix $p \in \mathbb{Z}^+$ and $\alpha \in (0, 1)$. Assume that conditions (C2)–(C5) are all satisfied. The empirical classifier $\hat{D}_Q(\hat{X}^* \mid \hat{\beta}_{p,a})$ at Equation (24) (or $\hat{D}_L(\hat{X}^* \mid \hat{\beta}_{p,a})$ found at Eq. 25) converges to its population version $D_Q(X^* \mid \beta_{p,a})$ at Equation (10) (or $D_L(X^* \mid \beta_{p,a})$ at (12)) in probability as $N$ diverges. Furthermore, if condition (C1) holds as well, a direct corollary is then
\[
\lim_{p \to \infty} \lim_{N \to \infty} \text{err}\{\hat{D}_L(\hat{X}^* \mid \hat{\beta}_{p,a})\} = 0.
\]

2.3.2. Tuning parameter selection

As explained in Section 2.2, this functional classification is connected to a problem of functional linear regression. As an alternative to the cross-validation, GCV is frequently used to choose the tuning parameters in functional linear models (e.g., Cardot, Ferraty & Sarda, 2003; Cardot et al., 2007; Reiss & Ogden, 2010). Computationally, this GCV-based method is considerably more efficient than cross-validation. In this article, we suggest searching for the optimal pair of $(p, \alpha)$ by minimizing
\[
\text{GCV}(p, \alpha) = \frac{\sum_{i=1}^{N} \left\{ Y_i - 1(\hat{D}_Q(\hat{X}_i \mid \hat{\beta}_{p,a}) < 0) \right\}^2}{(N - p - 2)^2}
\]
and
\[
\text{GCV}(p, \alpha) = \frac{\sum_{i=1}^{N} \left\{ Y_i - 1(\hat{D}_L(\hat{X}_i \mid \hat{\beta}_{p,a}) < 0) \right\}^2}{(N - p - 2)^2}
\]
for CCC-Q and -L, respectively, with respect to $(p, \alpha)$, where the digit 2 in the term $(N - p - 2)$ appearing in the denominator corresponds to the number of populations. Algorithm 1 details the implementation of CCC with the GCV-based tuning scheme.

Rather than the use of the usual rectangular search, the candidate pool for $(p, \alpha)$ here, say $\{(p, \alpha) : p \in \{1, \ldots, p_{\text{max},a}\}, \alpha \in \{\alpha_1, \ldots, \alpha_J\} \subset (0, 1)\}$, is nonrectangular and is set up in a random way: For each $\alpha \in \{\alpha_1, \ldots, \alpha_J\} (= \{0 \times 10^{-1}, \ldots, 9 \times 10^{-1}, 1 - 10^{-2}, \ldots, 1 - 10^{-4}\}$ in Section 3), $p_{\text{max},a}$ is randomly selected from Uniform$\{1, \ldots, p_{\text{upper}}\}$, where $p_{\text{upper}}$ can be determined by
\[
p_{\text{upper}} = \min \left\{ j_0 \in \mathbb{Z}^+ : \sum_{j=1}^{j_0} \hat{\lambda}_j^W / \sum_{j=1}^{\infty} \hat{\lambda}_j^W \geq 99\% \right\}
\]
in which $\hat{\lambda}_j^W$ estimates the $j$th largest eigenvalue of $\lambda_X^W$ at Equation (3). The random grid is typically of smaller cardinality and hence leads to less running time. Zhou (2019, Section 5) illustrated that this strategy incurs little loss in prediction accuracy.

3. NUMERICAL STUDIES

In spite of theoretical arguments illustrating the asymptotically perfect classification of CCC-L in specific cases, we still required further evidence to support our proposed approach, especially

The Canadian Journal of Statistics / La revue canadienne de statistique DOI: 10.1002/cjs.11624
with respect to CCC-Q. Therefore, we resorted to numerical studies to compare the performance of PCC, PLCC, and two CCC classifiers in finite-sample applications. The number of FPC (or FPLS) basis functions for PCC (or PLCC) was selected from \( \{1, \ldots, p_{\text{upper}}\} \) through five-fold

\[
\text{Algorithm 1 CCC tuned via GCV}
\]

1: \( p_{\text{max}} \leftarrow \) upper bound of the number of FC basis functions.
2: \( \text{for } p \text{ from 1 to } p_{\text{max}}, \alpha \text{ do} \)
3: \( \text{if } p = 1 \text{ then} \)
4: \( \text{URV}^T \leftarrow \) thin singular value decomposition of \( \hat{C}_c W^{1/2} \)
5: \( \hat{G}_1 \leftarrow \text{UR} \) and \( \hat{P} \leftarrow I_N \)
6: \( \text{else} \)
7: \( \hat{P} \leftarrow P \{ I_N - \hat{G}_{p-1} \}
\]
8: \( \hat{P}_{p-1} b_{p-1,a} (b_{p-1,a}^T \hat{G}_{p-1} b_{p-1,a})^{-1} b_{p-1,a}^T \hat{G}_{p-1} \}
\]
9: \( \text{end if} \)
10: \( \hat{G}_p \leftarrow \hat{P} G_1 \) and \( \hat{\zeta} \leftarrow \) largest eigenvalue of \( \hat{G}_p G_p \)
11: \( \text{L}(\hat{\delta}) \leftarrow (\hat{G}_p^T G_p + \hat{\delta}^{-1} \hat{\xi} L)^{-1} \)
12: \( \hat{\sigma} \leftarrow \text{GCV} \)
13: \( \hat{\sigma} \leftarrow \text{GCV} \)
14: \( \hat{\sigma} \leftarrow \text{GCV} \)
15: \( \hat{\sigma} \leftarrow \text{GCV} \)
16: \( \hat{\sigma} \leftarrow \text{GCV} \)
17: \( \hat{\sigma} \leftarrow \text{GCV} \)
18: \( \hat{\sigma} \leftarrow \text{GCV} \)
19: \( \hat{\sigma} \leftarrow \text{GCV} \)
20: \( \hat{\sigma} \leftarrow \text{GCV} \)
21: \( \hat{\sigma} \leftarrow \text{GCV} \)
22: \( \hat{\sigma} \leftarrow \text{GCV} \)
23: \( \hat{\sigma} \leftarrow \text{GCV} \)
24: \( \hat{\sigma} \leftarrow \text{GCV} \)
25: \( \hat{\sigma} \leftarrow \text{GCV} \)
26: \( \hat{\sigma} \leftarrow \text{GCV} \)
27: \( \hat{\sigma} \leftarrow \text{GCV} \)
28: \( \hat{\sigma} \leftarrow \text{GCV} \)
29: \( \hat{\sigma} \leftarrow \text{GCV} \)
30: \( \hat{\sigma} \leftarrow \text{GCV} \)
31: \( \hat{\sigma} \leftarrow \text{GCV} \)
32: \( \hat{\sigma} \leftarrow \text{GCV} \)
33: \( \hat{\sigma} \leftarrow \text{GCV} \)
34: \( \hat{\sigma} \leftarrow \text{GCV} \)
35: \( \hat{\sigma} \leftarrow \text{GCV} \)
36: \( \hat{\sigma} \leftarrow \text{GCV} \)
37: \( \hat{\sigma} \leftarrow \text{GCV} \)
38: \( \hat{\sigma} \leftarrow \text{GCV} \)
39: \( \hat{\sigma} \leftarrow \text{GCV} \)

DOI: 10.1002/cjs.11624

The Canadian Journal of Statistics / La revue canadienne de statistique
FIGURE 1: Samples following the simulation designs. Each subfigure displays a sample with red solid trajectories from \( \Pi_0 \) and blue dashed ones from \( \Pi_1 \).

cross-validation. Two more classifiers, functional versions of logit regression and naive Bayes, were involved also in the subsequent comparison. These two were both implemented via R-package \texttt{fda.usc} (Febrero-Bande & Oviedo de la Fuente, 2012).

3.1. Simulation Study
We generated \( R = 200 \) samples, each containing \( N = 200 \) curves \( X_i, i = 1, \ldots, 200 \). In each sample, we randomly chose 160 (80\%) for training and retained 40 (20\%) for testing. Each curve was spotted at 101 equally spaced points in \( T = [0, 1] \), i.e., \( \{0, 1/100, \ldots, 99/100, 1\} \), and iid with

\[
X_i = \sum_{k=0}^{1} \left( \sum_{j=1}^{5} \lambda_{k,j}^{1/2} Z_{i,j} \phi_{k,j} + \mu_{[k]} \right) 1(X_i \in \Pi_k).
\]
Without loss of generality, the difference of two mean functions was set to be exactly \( \mu_{[1]} \), i.e., \( \mu_{[0]}(\cdot) \equiv 0 \). Instead of a mixture of Gaussian processes, a setup without normality was considered, namely \( Z_{ij} \sim \text{Exponential}(1) \), \( i = 1, \ldots, N \), \( j = 1, \ldots, 5 \). Although the sub-covariance functions \( v_{X}^{[0]} \) and \( v_{X}^{[1]} \) shared the identical nonzero eigenvalues 200, 100, 1, 0.2, 0.1, they might have differed with respect to eigenfunctions. Specifically, we used the \( j \)th-order shifted Legendre polynomial—see Hochstrasser (1972, pp. 773–774)—normed to one as the \( j \)th eigenfunction.
TABLE 1: Mean misclassification percentage (%) (with standard deviations in parentheses) for various settings and classifiers. The last six columns, from left to right, correspond to classifiers CCC-L, CCC-Q, PLCC, PCC, (functional) logit regression and (functional) naive Bayes, respectively.

| Design | \( \rho \) | \( \pi_0 \) | CCC-L | CCC-Q | PLCC | PCC | Logit | Naive Bayes |
|--------|---------|---------|------|-------|------|-----|------|-------------|
| (i) 1  | 50%     | 30 (8.2)| 30 (8.6)| 31 (7.1)| 31 (7.3)| 30 (8.0)| 31 (7.9) |
| (i) 1  | 80%     | 21 (5.9)| 21 (6.0)| 20 (5.9)| 20 (6.0)| 21 (5.9)| 24 (6.6) |
| (i) 10 | 50%     | 0.13 (0.56)| 0.15 (0.60)| 0.19 (1.0)| 0.11 (0.52)| 0.09 (0.46)| 0.15 (0.60) |
| (i) 10 | 80%     | 0.22 (0.70)| 0.22 (0.72)| 0.31 (1.1)| 0.25 (0.75)| 0.11 (0.52)| 0.24 (0.78) |
| (ii) 1 | 50%     | 29 (7.8)| 7.4 (4.1)| 38 (14)  | 37 (14)  | 29 (8.0) | 27 (7.9) |
| (ii) 1 | 80%     | 17 (7.0)| 6.7 (4.0)| 20 (6.1) | 20 (6.2) | 20 (6.8) | 27 (8.3) |
| (ii) 10| 50%     | 0.22 (0.73)| 0.21 (0.62)| 33 (15)  | 35 (15)  | 0.11 (0.52)| 13 (6.9) |
| (ii) 10| 80%     | 0.23 (0.70)| 0.25 (0.79)| 20 (6.3) | 20 (6.7) | 0.14 (0.57)| 5.3 (4.4) |

of \( v_0^{[0]} \), i.e.,

\[
\phi_{0,1}(t) = \sqrt{3}(2t - 1),
\]
\[
\phi_{0,2}(t) = \sqrt{5}(6t^2 - 6t + 1),
\]
\[
\phi_{0,3}(t) = \sqrt{7}(20t^3 - 30t^2 + 12t - 1),
\]
\[
\phi_{0,4}(t) = 3(70t^4 - 140t^3 + 90t^2 - 20t + 1),
\]
\[
\phi_{0,5}(t) = \sqrt{11}(252t^5 - 630t^4 + 560t^3 - 210t^2 + 30t - 1).
\]

As a result, \( \rho_{\text{upper}} \), the upper bound of the number of components, was set to five directly rather than following the strategy at (26). We considered two sorts of combinations of \( \mu_{[1]} \) and \( \phi_{1,j} \):

(i) \( \mu_{[1]} = \rho \lambda^{1/2}_{1,1} \phi_{1,1} \) and \( \phi_{0,j} = \phi_{1,j} \);

(ii) \( \mu_{[1]} = \rho \lambda^{1/2}_{1,3} \phi_{1,3} \) and \( \phi_{0,5-j} = \phi_{1,j} \).

In both scenarios, \( \rho (= 1, 10) \) controlled the magnitude of the difference \( \mu_{[1]} - \mu_{[0]} \) as well as the ratio of between-group variation to within-group one. Although the setting \( \pi_0 = \Pr(X_i \in \Pi_0) = 50\% \) is commonly adopted in the literature, we also checked an unbalanced arrangement \( \pi_0 = 80\% \). Altogether, these eight settings for key study elements would help us clarify the joint impact from \( \rho, \pi_0, \) and \( \mu_{[1]} - \mu_{[0]} \) on misclassification.

Regardless of the values of \( \rho \) and \( \pi_0 \), design (i) favoured all classifiers equally (see Figure 2), since \( \mu_{[1]} - \mu_{[0]} \) was parallel to the top eigenfunction of not only \( v_X \) at (2) but also the function \( v_X^W \) at (3). The value of \( \rho \) had a great impact on classification accuracy for design (i): in Figure 1a (with \( \rho = 1 \)), the two subpopulations overlapped each other. The corresponding average misclassification rates were as high as 20%–30%; see Figure 2a,c and the first two rows of Table 1. However, when \( \rho = 10 \), the two subpopulations became visibly separable, resulting in more reasonable error rates (see Figures 2b,d and the third and fourth rows of Table 1).
FIGURE 3: Boxplots of misclassification percentage for design (ii). In each panel, the six boxes, from left to right, correspond to classifiers CCC-L, CCC-Q, PLCC, PCC, (functional) logit regression, and (functional) naive Bayes, respectively. The four subfigures come with the identical scale.

It was a different story for design (ii) that restricted \( \mu_{[1]} - \mu_{[0]} \) to be parallel to \( \phi_{0,3} \), viz. the least important eigenfunction of

\[
v_X^W(s, t) = 160.02\phi_{0,1}(s)\phi_{0,1}(t) + 80.04\phi_{0,2}(s)\phi_{0,2}(t) \\
+ 40.08\phi_{0,5}(s)\phi_{0,5}(t) + 20.16\phi_{0,4}(s)\phi_{0,4}(t) + \phi_{0,3}(s)\phi_{0,3}(t).
\]

In this case, since PCC was solely focused on decomposing in the function \( v_X^W \) at (3), it probably failed to extract the correct direction of \( \mu_{[1]} - \mu_{[0]} \) and hence generated more misclassification regardless of the values of \( \rho \) or \( \pi_0 \). Moreover, \( v_{[1]}^X \) shared the same eigenfunctions with \( v_X^0 \) but in a reversed order, violating the assumption concerning CCC-L and PLCC. Due to the magnitude of \( \lambda_{1,3} \), the two subgroups did not appear to be separable, even for the case when \( \rho = 10 \) (see Figure 1). Actually, trajectories from \( \Pi_1 \) were more bumpy; this feature would become more obvious for larger \( \rho \), e.g., \( \rho = 100 \) (not illustrated here). When \( \rho = 1 \), CCC-Q significantly
TABLE 2: Average misclassification percentage for real datasets (with standard deviations in parentheses) corresponding to different classifiers. The last six columns correspond to classifiers CCC-L, CCC-Q, PLCC, PCC, (functional) logit regression, and (functional) naive Bayes, respectively. Ties at the last row are caused by rounding.

| Dataset Description            | CCC-L (Mean) | CCC-Q (Mean) | PLCC (Mean) | PCC (Mean) | Logit (Mean) | Naive Bayes (Mean) |
|--------------------------------|--------------|--------------|-------------|------------|--------------|--------------------|
| Tecator™ (original)            | 5.5 (3.3)    | 4.6 (2.8)    | 30 (6.5)    | 29 (6.8)   | 8.8 (4.0)    | 39 (6.6)           |
| Tecator™ (second-order derivative) | 7.3 (4.4)   | 6.0 (3.4)    | 30 (6.5)    | 30 (6.5)   | 7.0 (3.6)    | 24 (5.9)           |
| DTI                            | 11 (3.5)     | 11 (3.5)     | 11 (3.6)    | 11 (3.6)   | 11 (3.7)     | 26 (4.5)           |

outperformed its five other competitors; see Figure 3a,c. As $\rho$ grew up to 10, the performance of all the classifiers was generally improved, though PLCC and PCC still gave rise to rather high misclassification rates (see the last two rows of Table 2 and Figure 3b,d). If we further enlarged $\rho$ to 100, say, the classification problem became no longer challenging, even for PLCC and PCC.

3.2. Real Data Application

We also considered two practical applications. For each dataset, we repeated a random split of the sample data a total of 200 times; each time we trained classifiers with 80% of the sample data and then tested the resulting classifiers on the remaining 20%. Table 2 summarizes the means and standard errors of the misclassification rates, expressed as percentages.

The first example involved the Tecator™ data (accessible at http://lib.stat.cmu.edu/datasets/tecator on July 1, 2021). This dataset was collected by Tecator™ Infratec Food and Feed Analyzer. It consisted of near-infrared absorbance spectra (i.e., the logarithm to base 10 of the transmittance at each wavelength) of 240 finely chopped pure meat samples. Each spectrum ranged from 850 to 1,050 nm and was spotted at 100 “time points” (viz. channels). Additionally, three sample content measurements, namely water, protein and fat, were recorded, as a percentage for each piece of meat. In our study, the meat samples were categorized into two groups: $\Pi_1$ was comprised of meat samples with a protein content less than 16%, and the remaining samples constituted $\Pi_0$. The 240 spectrum curves (or their second-order derivative curves as recommended by Ferraty & Vieu, 2006, Section 7.2.2) were used as functional covariates in the study. For both sorts of trajectories, the classification output was analogous to that for simulated design (ii) with $\rho = 10$: the CCC subtypes and (functional) logit regression yielded considerably fewer errors than the other three classifiers; compare Figures 3b,d and 4a,b.

We also investigated a dataset concerning the use of DTI, which is mentioned in Section 1. The fractional anisotropy, which is measured via DTI, is a scalar ranging from 0 to 1 and is used to reflect the fibre density, axonal diameter and myelination in white matter. Along a white matter tract (e.g., the corpus callosum), fractional anisotropy values form a curve. The dataset DTI in the R-package refund (Goldsmith et al., 2019) was initially collected at the Johns Hopkins University and the Kennedy Krieger Institute; it contained fractional anisotropy curves (measured at 93 locations) for the corpus callosum of healthy people ($\Pi_0$) and MS patients ($\Pi_1$), 382 subjects altogether. By classifying these curves (with missing values imputed through local polynomial regression), we tried to identify the status of each subject: healthy or suffering from MS. Except for the (functional) naive Bayes, the classifiers reached a tie after rounding (see the last row of Table 2), i.e., the observed error rates for identifying MS patients were nearly identical.
FIGURE 4: Boxplots of misclassification percentages for Tecator™ and DTI data. The top two panels reflect respective results with original and second-order derivative curves. In each panel, the four boxes, from left to right, correspond to classifiers CCC-L, CCC-Q, PLCC, PCC, (functional) logit regression, and (functional) naive Bayes, respectively. The three subfigures are displayed with the identical scale.

4. CONCLUSION AND DISCUSSION

We have investigated two subtypes of CCC classifiers, viz. CCC-L and -Q, for the binary classification of curves. Theoretically, under certain circumstances, CCC-L enjoys an (asymptotic) zero misclassification rate regardless of the distribution assumption, while, in certain empirical studies, CCC-Q seems superior to CCC-L and some other competitors. Provided certain regularity conditions hold, our proposed approach results in empirical classifiers that are consistent in the case of “fixed \( p \) and infinite \( N \).”

In the numerical studies reported in Section 3, we do see benefits from the introduction of the supervision controller \( \alpha \). Nevertheless, one cannot be too optimistic; actually, tuning one more
parameter may lead to greater variation and even more bias. In addition, the projection direction $\beta_{p,a}$ at (9) potentially becomes unstable for large values of $p$, since $w_{p,a}$ is constructed iteratively without penalty. Nevertheless, adding a penalty term may introduce one more hyperparameter that forces our implementation to be more computationally involved.

One may be concerned in application about the “optimal” values of the hyperparameters for our proposed classifiers. In most of the numerical studies, the hyperparameter $\alpha$ (or $p$) ended with similar values; exceptions to this phenomenon arose for design (ii) with $\rho = 1$ (where CCC-Q outperformed CCC-L significantly). In that case, CCC-Q picked up $\alpha$ with a mean value around 0.7, whereas CCC-L chose a value near 0.5 and hence yielded a classification accuracy comparable with PLCC. As for the size of the basis used, CCC subtypes seemed even more “parsimonious” than PLCC, mostly utilizing less than three basis functions.

With respect to the functional classification with $K \geq 3$ classes, one naive strategy might be to carry out binary classifiers repeatedly. To be explicit, for a newcomer $X^*$, each time we only consider two distinct labels and then assign either label to it, or equivalently, throw a vote for either of the two labels. After all $\binom{K}{2}$ binary classifications, the label that wins the most votes is eventually assigned to $X^*$. Another route might be to embed the FC basis into the iteratively reweighted least squares (Green, 1984). In that way, the resulting modification could be utilized for fitting generalized linear models with functional covariates.

REFERENCES

Björkström, A. & Sundberg, R. (1999). A generalized view on continuum regression. Scandinavian Journal of Statistics, 26, 17–30.

de Boor, C. (2001). A Practical Guide to Splines, revised ed., Springer, New York.

Cardot, H., Crambes, C., Kneip, A., & Sarda, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. Computational Statistics & Data Analysis, 51, 4832–4848.

Cardot, H., Ferraty, F., & Sarda, P. (2003). Spline estimators for the functional linear model. Statistica Sinica, 13, 571–591.

Chui, C. K. (1971). Concerning rates of convergence of Riemann sums. Journal of Approximation Theory, 4, 279–287.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. Numerische Mathematik, 31, 377–403.

Dai, X., Müller, H.-G., & Yao, F. (2017). Optimal Bayes classifiers for functional data and density ratios. Biometrika, 104, 545–560.

Delaigle, A. & Hall, P. (2012). Achieving near perfect classification for functional data. Journal of the Royal Statistical Society Series B, 74, 267–286.

Fan, J., Feng, Y., & Tong, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. Journal of the Royal Statistical Society Series B, 74, 745–771.

Febrero-Bande, M. & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software, 51, 1–28.

Ferraty, F. & Vieu, P. (2003). Curves discrimination: A nonparametric functional approach. Computational Statistics & Data Analysis, 44, 161–173.

Ferraty, F. & Vieu, P. (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York.

Galeano, P., Joseph, E., & Lillo, R. E. (2015). The Mahalanobis distance for functional data with applications to classification. Technometrics, 57, 281–291.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., & Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. Journal of the Royal Statistical Society Series C, 61, 453–469.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., et al. (2019). refund: Regression with Functional Data. R package version 0.1-21.
APPENDIX

Our theoretical perspectives involve the following conditions.

(C1) \[ \left\{ \frac{\int \beta_{p,a}(\mu_{[1]} - \mu_{[0]})}{\int \beta_{p,a}(X - \mu_{[k]}) | X \in \Pi_k} \right\}^2 / \text{var} \left\{ \int \beta_{p,a}(X - \mu_{[k]}) | X \in \Pi_k \right\} \text{diverges as } p \to \infty \text{ for each } \alpha \text{ and } k. \]

(C2) Realizations of \( X \) are twice continuously differentiable and \( \|X'\| \) is bounded almost surely.

(C3) \( \tau_1, \ldots, \tau_L \) are eigenvalues of \( \text{Pen}^{-1/2} W \text{Pen}^{-1/2} \) such that \( \tau_1 = \tau_2 = 0, \tau_3 \geq \cdots \geq \tau_L, \) and \( C_1 (l - 2)^{-4} \leq \tau_l \leq C_2 (l - 2)^{-4} \) for \( l \geq 3, \) with neither \( C_1 \) nor \( C_2 \) depending on \( l \) or \( L. \)

(C4) \( M \to \infty \) and \( M^{-1} \theta_0 \to 0 \) as \( N \to \infty, \) where \( \theta_0 > 0 \) is the smoothing parameter for all trajectories.

(C5) For all \( j (\leq p \leq \text{rank}(\hat{C}_c W^{1/2})), T_{j,a}(w) \) attains a unique maximizer (up to sign) in \( \left\{ w \in L^2(\mathcal{T}) : \|w\| = 1 \right\}. \)

Condition (C1) implies that, after the projection to the direction of \( \beta_{p,a} \), as \( p \) diverges, the within-group covariance becomes more and more ignorable when compared with the corresponding between-group covariance, i.e., the two groups become more and more separable. It is analogous to assumption (4.4), part (d), of Delaigle & Hall (2012) and assures us of the (asymptotic) perfect classification of CCC-L. Assumptions (C2) and (C3) jointly guarantee that the smoothed curves converge to the true ones as observations become more and more dense; although the sufficiency of condition (C3) for natural splines has been proved by Utreras (1983, Eq. (4)), we have little knowledge on whether a similar result holds for general cubic B-splines. Consequently, we have to follow Craven & Wahba (1979, Eq. (A4.3.1)) and assume both (C2) and (C3). If we have extra regularity conditions (C4) and (C5), the proposed empirical classifiers described in Section 2.3.1 turn out to be convergent in probability.
Proof of Proposition 1. Write \( \gamma_{p,a} = \int \beta_{p,a}(\mu_{[1]} - \mu_{[0]}) \) and \( R_{p,a}^{[k]} = \int \beta_{p,a}(X^* - \mu_{[k]}) \). Recalling (11), \( \sigma_{[k]}^2(\beta_{p,a}) = \text{var}(R_{p,a}^{[k]} | \mathbf{X} \in \Pi_k) \), \( k = 0, 1 \). Thus,

\[
\Pr\{D_L(X^* | \beta_{p,a}) < 0 | X^* \in \Pi_0\} = \Pr\left\{ \left( R_{p,a}^{[0]} - \gamma_{p,a} \right)^2 - (R_{p,a}^{[0]})^2 < 2\sigma_{[0]}^2(\beta_{p,a}) \ln \frac{1 - \pi_0}{\pi_0} \left| X^* \in \Pi_0 \right. \right\}
\]

\[
= \Pr \left[ \frac{R_{p,a}^{[0]}}{\sigma_{[0]}(\beta_{p,a})} > \frac{\gamma_{p,a}^2 + 2\sigma_{[0]}^2(\beta_{p,a}) \ln \{\pi_0/(1 - \pi_0)\}}{2\gamma_{p,a}^2(\beta_{p,a})} \right| X^* \in \Pi_0 \right]
\]

\[
\leq \frac{4\sigma_{[1]}^2(\beta_{p,a})/\gamma_{p,a}^2}{\left[ 1 + 2\gamma_{p,a}^2(\beta_{p,a}) \ln \{ (1 - \pi_0)/\pi_0 \} \right]^2},
\]

where the upper bound is derived from Chebyshev’s inequality as well as the identity that \( R_{p,a}^{[0]}/\sigma_{[0]}(\beta_{p,a}) \) (conditional on the event \( X^* \in \Pi_0 \)) has zero mean and unit variance. Similarly, we deduce that

\[
\Pr\{D_L(X^* | \beta_{p,a}) > 0 | X^* \in \Pi_1\} \leq \frac{4\sigma_{[1]}^2(\beta_{p,a})/\gamma_{p,a}^2}{\left[ 1 + 2\gamma_{p,a}^2(\beta_{p,a}) \ln \{ (1 - \pi_0)/\pi_0 \} \right]^2}.
\]

Eventually, as \( p \) diverges, the zero-convergence of

\[
\text{err}\{D_L(X^* | \beta_{p,a})\} = \pi_0 \Pr\{D_L(X^* | \beta_{p,a}) < 0 | X^* \in \Pi_0\} + (1 - \pi_0) \Pr\{D_L(X^* | \beta_{p,a}) > 0 | X^* \in \Pi_1\}
\]

results from condition (C1) (i.e., \( \sigma_{[k]}^2(\beta_{p,a})/\gamma_{p,a}^2 \to 0 \) as \( p \) diverges for each \( \alpha \) and \( k \)).

Proof of Proposition 2. Recall that \( \Delta t = (t_{\text{max}} - t_{\text{min}})/M \) and the definitions of the matrices \( \Psi, \hat{c}, \Psi, \mathbf{W} \) and \( \text{Pen} \), all specified in Section 2.3.1; see Equations (15)–(19). Introduce the operator \( \mathcal{P}_{BS_L} \) such that, for each \( f \in L^2(\mathcal{I}) \), \( \mathcal{P}_{BS_L} f \) is the orthogonal projection of \( f \) onto \( BS_L \) at (13), i.e.,

\[
\mathcal{P}_{BS_L} f = \left[ \int f \psi_1, \ldots, \int f \psi_L \right] \mathbf{W}^{-1} \Psi.
\]

For each \( i \), specifically, \( \mathcal{P}_{BS_L} X_i = c_i^T \Psi \) with

\[
c_i = \mathbf{W}^{-1} \left[ \int f \psi_1, \ldots, \int f \psi_L \right]^T.
\]

We chop \( \| \hat{X} - X_i \| \) into two segments: \( \| \mathcal{P}_{BS_L} X_i - X_i \| \) and \( \| \hat{X} - \mathcal{P}_{BS_L} X_i \| \). Combining Lyche, Manni & Speleers (2018, Theorem 16) with condition (C2) implies

\[
\| \mathcal{P}_{BS_L} X_i - X_i \| = O_p(\Delta t) = O_p(M^{-1}) \quad \text{as} \quad M \to \infty.
\]
Furthermore, condition (C2) allows us to follow Chui (1971, Theorem 5) to verify that, as \( M \to \infty \), \( \| W - \Delta t \Psi^T \Psi \|_F^2 = O(M^{-2}) \), \( \| W - \Delta t \Psi^T \Psi - \Delta t \theta_0 \text{Pen} \|_F = O(1) \), and

\[
\left\| \left[ \int_X \varphi_1, \ldots, \int_X \varphi_L \right]^T - \Delta t X_i \Psi \right\|_F^2 = O_p(M^{-3}),
\]

where \( \| \cdot \|_F \) denotes the Frobenius norm. Let \( \tau_1, \ldots, \tau_L \) be eigenvalues of

\[
Z = \text{Pen}^{-\frac{1}{2}} W \text{Pen}^{-\frac{1}{2}}
\]

with corresponding eigenvectors \( e_1, \ldots, e_L \). Note the finiteness of both \( \lim_{L \to \infty} \max_{l,1} | \int \varphi_l \varphi_{l'}^T | \) and \( \lim_{L \to \infty} \max_{l,1} | \int \varphi_l'' \varphi_{l''}^T | \). Thus the squared second term is

\[
\| \hat{X}_i - P_{BSL} X_i \|^2 = (\hat{e}_i^T - c_i^T) W (\hat{e}_i - c_i)
\]

\[
= \left\{ \Delta t X_i^T \Psi (\Delta t \Psi^T \Psi + \Delta t \theta_0 \text{Pen})^{-1} - c_i \right\} W
\]

\[
\times \left\{ (\Delta t \Psi^T \Psi + \Delta t \theta_0 \text{Pen})^{-1} \Delta t \Psi^T X_i - c_i \right\}
\]

\[
= c_i^T \text{Pen}^{-\frac{1}{2}} \left\{ Z (Z + \Delta t \theta_0 I_L)^{-1} - I_L \right\} Z
\]

\[
\times \left\{ (Z + \Delta t \theta_0 I_L)^{-1} Z - I_L \right\} \text{Pen}^{-\frac{1}{2}} c_i + o_p(1)
\]

\[
= (\Delta t)^2 \theta_0 c_i^T \text{Pen}^{-\frac{1}{2}} (Z + \Delta t \theta_0 I_L)^{-1} Z (Z + \Delta t \theta_0 I_L)^{-1} \text{Pen}^{-\frac{1}{2}} c_i + o_p(1)
\]

\[
= \sum_{l=1}^{L} \tau_l \frac{\theta_0}{\Delta t \theta_0 (\Delta t)^{-1} \tau_l + 1} (c_i^T \text{Pen}^{-\frac{1}{2}} e_i)^2 + o_p(1)
\]

\[
\leq \Delta t \theta_0 \sum_{l=1}^{L} \tau_l (c_i^T \text{Pen}^{-\frac{1}{2}} e_i)^2 + \sum_{l=L+1}^{L} \tau_l (c_i^T \text{Pen}^{-\frac{1}{2}} e_i)^2 + o_p(1)
\]

\[
= o_p(1) \quad \text{as} \quad M \to \infty,
\]

where \( L_1 \in \mathbb{Z}^+ \) is so defined that \( \tau_{L_1} = \max \{ \tau_l, (\Delta t \theta_0)^{1/2} \} \) and diverges as \( M \to \infty \) owing to condition (C3).

**Proof of Proposition 3.** Recall \( P_{BSL} \) defined in the proof of Proposition 2. Writing \( \hat{w}_{j,a} = P_{BSL} \hat{w}_{j,a} + (I - P_{BSL}) \hat{w}_{j,a} \), with identity operator \( I \), one has \( 0 < \| P_{BSL} \hat{w}_{j,a} \| \leq 1 \) and \( \int \hat{X}_i P_{BSL} \hat{w}_{j,a} = \int \hat{X}_i \hat{w}_{j,a} \) since \( \hat{X}_i \in BS_L \) for all \( i \). If \( 0 < \| P_{BSL} \hat{w}_{j,a} \| < 1 \) (i.e., \( (I - P_{BSL}) \hat{w}_{j,a} > 0 \)), then \( P_{BSL} \hat{w}_{j,a} / \| P_{BSL} \hat{w}_{j,a} \| \) satisfies

\[
\hat{T}_{j,a} \left( \frac{P_{BSL} \hat{w}_{j,a}}{\| P_{BSL} \hat{w}_{j,a} \|} \right) = \| P_{BSL} \hat{w}_{j,a} \|^{\frac{2}{\alpha - 1}} \hat{T}_{j,a} (\hat{w}_{j,a}) > \hat{T}_{a}(\hat{w}_{j,a}),
\]

which violates the definition of \( \hat{w}_{j,a} \) at (21). This contradiction implies that \( (I - P_{BSL}) \hat{w}_{j,a} \) must be zero. 

\[
\text{DOI: 10.1002/cjs.11624}
\]

The Canadian Journal of Statistics / La revue canadienne de statistique
Proof of Proposition 4. Under conditions (C2)–(C5), fixing $p$, Proposition 2 and Zhou (2019, Theorem 1) assure us that $\|\hat{\beta}_{p,a} - \beta_{p,a}\|$ converges in probability to zero as $N$ diverges. Since Proposition 2 applies to $X^*$, the convergence of the empirical classifiers follows. □