*to appear in Biometrika*

# Sparse Functional Linear Discriminant Analysis

Juhyun Park*
Lancaster University, U.K. and ENSIIE, France
Jeongyoun Ahn
University of Georgia, U.S.A.
and Yongho Jeon
Yonsei University, South Korea

December 11, 2020

**Abstract**

Functional linear discriminant analysis offers a simple yet efficient method for classification, with the possibility of achieving a perfect classification. Several methods are proposed in the literature that mostly address the dimensionality of the problem. On the other hand, there is a growing interest in interpretability of the analysis, which favors a simple and sparse solution. In this work, we propose a new approach that incorporates a type of sparsity that identifies non-zero sub-domains in the functional setting, offering a solution that is easier to interpret without compromising performance. With the need to embed additional constraints in the solution, we reformulate the functional linear discriminant analysis as a regularization problem with an appropriate penalty. Inspired by the success of $\ell_1$-type regularization at inducing zero coefficients for scalar variables, we develop a new regularization method for functional linear discriminant analysis that incorporates an $L^1$-type penalty, $\int |f|$, to induce zero regions. We demonstrate that our formulation has a well defined solution that contains zero regions, achieving a functional sparsity in the sense of domain selection. In addition, the misclassification probability of the regularized solution is shown to converge to the Bayes error if the data are Gaussian. Our method does not presume that the underlying function has zero regions in the domain, but produces a sparse estimator that consistently estimates the true function whether or not the latter is sparse. Numerical comparisons with existing methods demonstrate this property in finite samples with both simulated and real data examples.

**Keywords**: Domain selection; Functional classification; Functional sparsity; Interpretability; $L^1$ penalty; Linear discriminant analysis.

*Corresponding author: Juhyun Park, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K. and ENSIIE & LaMME, Université Paris-Saclay, 91025 Évry, France. email: juhyun.park@ensiie.fr

# 1 Introduction

We consider a classification problem when data are curves. Denote the functional predictor by $X \in L^2(\mathcal{I})$, observable on a real interval $\mathcal{I}$, and the class label by $Y$. Assuming that $X$ is a member of two possible groups ($Y = 0$ or $Y = 1$), we seek a classification rule that depends on a linear map, with unknown function $\beta$,

$$F_\beta(X) = \int_{\mathcal{I}} X(t)\beta(t)\,\mathrm{d}t.$$

This defines a functional linear classification problem, where we wish to determine $\beta(\cdot)$ in such a way that the linear map yields a good class separation. As functional data models are inherently infinite-dimensional, dimension reduction techniques are essential in constructing a solution.

The standard logistic regression framework for classification can be extended with functional variables (Müller, 2005) as

$$\log \frac{\mathrm{pr}(Y = 1 \mid X)}{\mathrm{pr}(Y = 0 \mid X)} = a_0 + \int_{\mathcal{I}} X(t)\beta(t)\,\mathrm{d}t\,.$$

Alternatively, one can try to directly extend the Bayes classifier. If $X$ were finite-dimensional, the Bayes rule is defined as maximizing the conditional probability given by

$$\mathrm{pr}(Y = 1 \mid X) = \frac{\pi_1 f_1(X)}{\pi_0 f_0(X) + \pi_1 f_1(X)}\,,$$

where $i = 0, 1, \pi_i$ is the probability of $X$ coming from group $i$ and $f_i(X)$ is the marginal density of $X$ if it belongs to group $i$. An optimal classification can then be achieved when one classifies new variable $X$ to the class 1 if $\pi_1 f_1(X) > \pi_0 f_0(X)$. Consequently, the main effort is devoted to estimating $f_1(X)$ and $f_0(X)$. If $X$ is Gaussian, further simplifications can be made and such an approach is broadly known as linear discriminant analysis. The main difficulty with extending the Bayes rule to functional data is linked to the fact that the marginal densities do not exist for functional data (Delaigle and Hall, 2010). Nevertheless, approaches attempting to directly approximate the marginal densities with dimension reduction techniques have proven useful, even in the absence of well-defined target densities. For example, James and Hastie (2001) use a Gaussian framework to develop regularization methods, while Bongiorno and Goia (2016) and Dai et al. (2017) develop nonparametric approaches without Gaussian assumptions.

It is well known that functional classification can achieve a perfect classification, if the infinite-dimensionality is well exploited. This implies that for the purpose of classification, it is not necessarily advantageous to have a well-defined finite-dimensional representation. Delaigle and Hall (2012) demonstrate such a phenomenon with a simple linear centroid classifier using asymptotic analysis and suggest a practical representation using components obtained from functional principal component analysis and partial least squares. Kraus and Stefanucci (2018) propose $L^2$ regularization methods to obtain the

representation. Berrendero et al. (2018) further clarify this phenomenon under a reproducing kernel Hilbert space framework for Gaussian processes, suggesting an alternative finite dimensional approximation.

While optimal performance is an important criterion to consider, the increasing impact of statistical analysis on modern scientific investigations has created the need to carefully consider interpretability of the outcomes of the analysis. Some attempts have been made to address interpretability in functional data, based on the idea that a simpler form of function is easier to interpret, and thus more useful in practice. As the function is an infinite-dimensional object, the formulation is often given in terms of a basis function representation. Under this setting, three different approaches have been proposed to construct a simpler form of functions. The first one is to impose a constraint on the coefficients directly with an $\ell_1$ norm (e.g., Zhou et al., 2012). Assuming that $\beta$ can be well approximated by a finite number of basis functions, say $K$, this can be expressed as $\beta(t) = \sum_{k=1}^{K} \alpha_k B_k(t)$ subject to $\sum_{k=1}^{K} |\alpha_j| \leq C$ and thus encourages the coefficients to be zero. The second approach is to limit the class of the functions $\{B_k\}$ in terms of their shape such as constant or strictly linear functions only (e.g., Tian and James, 2013). A difficulty with a standard sparse regularization for a function is that the penalty that encourages a sparse representation of the function does not necessarily have a control over domain selection: i.e., even if $\alpha_k = 0$ for some $k$, $\beta(t) \neq 0$ for $t \in I_k(t) = \{t \in \mathcal{I} : B_k(t) \neq 0\}$. The third one is to limit the support of the function $\beta$ to include zero regions (James et al., 2009; Zhou et al., 2013; Martin-Barragan et al., 2014; Lin et al., 2016; Picheny et al., 2019). We wish to incorporate interpretability in the latter notion of obtaining zero regions in the solution. However, as noted by Kneip et al. (2016) and Roche (2018), the theoretical framework appropriate to deal with a discrete notion of sparsity in high dimensions does not necessarily offer an insight into a problem in an infinite dimensional setting.

A functional formulation on sparsity is relatively scarce. Wang and Kai (2015) introduce the notion of functional sparsity, distinguishing global sparsity, which relates to functional variable selection, from local sparsity, which relates to domain selection with zero regions. Tu et al. (2020) develop a regularization method to achieve simultaneous estimation of both types of sparsity in a time varying functional regression setting and Lin et al. (2017) propose an alternative regularization to achieve local sparsity in functional linear regression. Both approaches rely on a clever grouping of the sparse coefficients in the basis function representation. Hall and Hooker (2016) study the issue of identifiability of domain selection problem in functional linear regression and suggest a domain search strategy. Kraus and Stefanucci (2018) follow a similar line.

In this work, we seek an alternative approach to functional linear classification with a direct estimation method that addresses dimensionality, optimality and interpretability. With the need to embed additional constraints on the form of the solution, we reformulate the functional linear discriminant analysis as a regularization problem with an appropriate choice of penalty functions. So far, the penalty-based regularization methods have been used mostly for either smoothness in regression (e.g., Cardot et al., 2003; Crambes et al., 2009) or invertibility in classification (Kraus and Stefanucci, 2018),

based on an $L^2$-type penalty. The idea of sparsity as variable selection with an $\ell_1$-type penalty is actively developed in the high-dimensional setting with scalar variables, but much less for the infinite-dimensional functional setting. We develop a new regularization method for functional linear discriminant analysis with an $L^1$-type penalty in the functional setting for inducing zero regions, as opposed to zero coefficients. We study the underlying optimization problem in detail, showing that it has a well defined solution that contains zero regions, achieving a functional sparsity in the sense of domain selection. Furthermore, we show that our regularized classifier asymptotically behaves similarly to the optimal classifier. We do not presume, even in the case where the optimal classifier is well defined, that the underlying projection function $\beta$ has a local sparse property with true zero regions in the domain. The role of the $L^1$ penalty is to provide a sparse estimator that consistently estimates the true function whether or not the latter is sparse, enhancing interpretability without compromising prediction performance. Unlike the approaches based on direct dimension reduction techniques such as principal component analysis or partial least squares, our proposed regularization method does not require assumptions on the eigenvalue sequences of the covariance function or any other unknown quantities. Numerical comparisons with existing methods demonstrate the properties of the proposed estimator in finite samples with both simulated and real data examples.

# 2 Methodology

## 2.1 Functional data framework

We use the following standard notation: for $f \in L^p(\mathcal{I})$, $1 \le p < \infty$, the norm is defined by $\|f\|_p = (\int_{\mathcal{I}} |f(t)|^p \, \mathrm{d}t)^{1/p}$. When $p = \infty$, $\|f\|_\infty = \inf\{C : |f(t)| \le C \text{ a.e. on } \mathcal{I}\}$.

Assume that $X \in L^2(\mathcal{I})$ with the mean $\mu(t)$ and covariance function $\mathrm{cov}\{X(s), X(t)\} = \gamma(s,t), s, t, \in \mathcal{I}$. Define the inner product in $L^2(\mathcal{I})$ by $\langle f, g\rangle = \int_{\mathcal{I}} fg$ for $f, g \in L^2(\mathcal{I})$. Assuming $\int_{\mathcal{I}} \int_{\mathcal{I}} \gamma(s,t)^2 \, \mathrm{d}s\mathrm{d}t < \infty$, define the covariance operator $\Gamma : L^2(\mathcal{I}) \to L^2(\mathcal{I})$ as

$$\Gamma(\beta)(t) = \int_{\mathcal{I}} \gamma(s,t)\beta(s) \, \mathrm{d}s \,, \quad t \in \mathcal{I}.$$

It is known that $\Gamma$ is a compact operator, and is Hilbert-Schmidt, that admits a spectral representation given by

$$\Gamma(\beta) = \sum_{j=1}^{\infty} \theta_j \langle \beta, \phi_j \rangle \phi_j \,, \qquad \theta_j \ge 0 \,, \quad \theta_j \to 0 \ \text{ as } j \to \infty \,,$$

where $\theta_j$ and $\phi_j(\cdot)$ correspond to eigenvalues and eigenfunctions of the covariance operator $\Gamma$, respectively. One of the important properties of a compact operator is that it is not invertible unless it has only finitely many distinct eigenvalues. At the same time, since $\theta_j \to 0$ as $j \to \infty$, a finite rank approximation to $\Gamma$ is also well understood, which has been the basis of many regularization and dimension reduction techniques

for functional data, notably functional principal component analysis and its applications (e.g., Hall et al., 2001). More details on theoretical foundations of functional data are found in Hsing and Eubank (2015).

For our classification problem, we denote the variable in each group by $X_0$ and $X_1$ and we make the following standard assumptions.

**Assumption 1.** *For $k = 0, 1$, $X_k$ is square integrable function on a compact interval $\mathcal{I}$ with $\mu_k = E(X_k)$ and $\mu_0 \neq \mu_1$ with common covariance function $\gamma$. The mean functions and the covariance function are continuous.*

**Assumption 2.** *For $k = 0, 1$, $E(\|X_k\|_2^4) < \infty$.*

Without loss of generality, we assume that $\mu_0 = 0$ and $\mu_1 = \delta$.

## 2.2 An optimal linear classifier

The optimality of the linear methods in the homoskedastic Gaussian scenario is well established by Delaigle and Hall (2012) based on an asymptotic centroid-based classifier. We briefly review their framework to motivate our regularization method, which is introduced in the following section.

For given $X$, the linear classifier with $\beta$ can be defined in multiple ways. We assume that $\beta \in L^2(\mathcal{I})$ is continuous. The class assignment based on $\beta$ is done via

$$T^0(X) = \big(\langle X, \beta\rangle - \langle \delta, \beta\rangle\big)^2 - \big(\langle X, \beta\rangle\big)^2 ,$$

which assigns to $X$ the group label $Y = 1$ if $T^0(X) < 0$ and $Y = 0$ if $T^0(X) > 0$. Then, the misclassification error is $\pi_0 \mathrm{pr}\{T_0(X) < 0 \mid Y = 0\} + \pi_1 \mathrm{pr}\{T_0(X) > 0 \mid Y = 1\}$. If $X$ is Gaussian, the misclassification probability can be expressed as

$$err_X(\beta) = 1 - \Phi\left(\frac{|\langle \delta, \beta\rangle|}{2\langle \beta, \Gamma\beta\rangle^{1/2}}\right) , \tag{1}$$

with its minimal error given by $1 - \Phi(\|\Gamma^{-1/2}\delta\|_2/2)$ (Delaigle and Hall, 2012; Kraus and Stefanucci, 2018). It can be seen that an optimal function $\beta$ that minimizes the misclassification error is

$$\max_{\beta \neq 0} \frac{\langle \delta, \beta\rangle^2}{\langle \Gamma\beta, \beta\rangle} = \max_{\beta \neq 0} \frac{\langle \delta, \beta\rangle^2}{\mathrm{var}(\langle X, \beta\rangle)} .$$

This is equivalent to the extended criterion for Fisher's discrimination analysis for functional data (Shin, 2008) defined by

$$\max_{\beta \neq 0} \frac{\mathrm{var}[E\{F_\beta(X) \mid Y\}]}{E[\mathrm{var}\{F_\beta(X) \mid Y\}]} = \max_{\beta \neq 0} \frac{\pi_0\pi_1\langle \delta, \beta\rangle^2}{\langle \Gamma\beta, \beta\rangle} . \tag{2}$$

Attempting to directly solve (2) would lead to a form of generalized eigenvalue problem. Equivalently, this can be formulated as

$$\max_{\beta} \langle \delta, \beta\rangle^2 \quad \text{subject to} \quad \langle \Gamma\beta,\, \beta\rangle - 1 = 0. \tag{3}$$

The solution can be derived from the Lagrangian formulation of (3)

$$J_\rho(\beta) = \langle \delta,\, \beta\rangle^2 - \rho\{\langle \Gamma\beta,\, \beta\rangle - 1\}\,,$$

for some constant $\rho$, which leads to

$$\Gamma\beta = \delta\,. \tag{4}$$

Hence, we see that the optimal linear classifier can be defined as the solution to (4). However, since a bounded inverse of $\Gamma$ does not exist in the infinite dimensional setting (e.g., Cardot et al., 2007), this equation clearly illustrates that linear discrimination analysis for functional data is an ill-posed inverse problem. When the covariance operator further satisfies $\|\Gamma^{-1}\delta\|_2 < \infty$, a unique classifier exists in $(\mathrm{Ker}(\Gamma))^\perp$ with the optimal error $1 - \Phi(\|\Gamma^{-1/2}\delta\|_2/2)$. When $\|\Gamma^{-1}\delta\|_2 = \infty$, an optimal solution does not exist, but optimal classification can be achieved asymptotically along a non-convergent path and perfect classification may be possible if $\|\Gamma^{-1/2}\delta\|_2 = \infty$. Therefore, unlike regression, it is not necessary to assume the existence of a unique solution to obtain an approximate solution, whose performance can mimic the optimal classifier asymptotically (Delaigle and Hall, 2012; Kraus and Stefanucci, 2018). In the following, we will denote the solution to (4) by $\beta_0$ when it exists and is finite.

### 2.3 A regularized solution to discrimination

The primary purpose of regularization is to solve an ill-posed inverse problem. We introduce our regularization method to solve the ill-posed inverse problem (4) above. A sensible strategy would be to impose some constraints on the solution set or a penalty term to an objective function. In order to impose the functional equation (4), we introduce a corresponding minimization problem. Specifically, viewing the discriminant equation in (4) as a type of score equation leads to defining an objective function as

$$\ell(\beta) = \frac{1}{2}\langle \Gamma\beta, \beta\rangle - \langle \delta, \beta\rangle\,. \tag{5}$$

Hence the functional derivative of (5) with respect to $\beta$ yields (4). In practice, where $\Gamma$ and $\delta$ are not available, replacing $(\Gamma, \delta)$ by their empirical counterpart $(\Gamma_n, \delta_n)$ gives

$$\ell_n(\beta) = \frac{1}{2}\langle \Gamma_n\beta, \beta\rangle - \langle \delta_n, \beta\rangle\,.$$

For example, the standard sample covariance operator and the sample mean could be used for $\Gamma_n$ and $\delta_n$. The standard sample estimators often lack smoothness (e.g., Cardot et al., 1999) so it is often desirable to use their smoothed version by taking into account the underlying design schemes (Zhang and Wang, 2016). Our formulation does not rely on the specific choice of estimators as long as they are consistent. To incorporate additional constraints, we consider the following optimization problem

$$\min_{\beta\in\mathcal{H}}\{\ell_n(\beta) + P(\beta)\}\,, \tag{6}$$

where $\mathcal{H}$ is an appropriate function space and $P(\beta)$ is a penalty corresponding to the constraint. The type of constraint requires some further consideration. As the underlying problem is defined in $L^2(\mathcal{I})$, it seems natural to add a penalty that depends on the $L^2$ norm of $\beta$ such as $\|\beta\|_2^2$ as in Kraus and Stefanucci (2018), while a smoothness constraint on $\beta$ would lead to the $L^2$ penalty on the derivatives. Nevertheless, $L^2$ regularization in general does not produce a sparse solution (e.g., Lin et al., 2017; Tu et al., 2020). This phenomenon is well investigated in multivariate data, and the popularity of the $\ell_1$ or $\ell_0$ penalty demonstrates the effectiveness of a sparse solution in a broader context.

In order to introduce a type of sparsity in the discriminant function, we propose to directly impose the functional norm constraints $\|\beta'\|_2^2 \leq C_1$ and $\|\beta\|_1 \leq C_2$. Although other constraints are possible, it turns out that this combination of constraints allows direct control over the $L^2$ norm. One might wonder whether the derivative penalty is necessary in our sparse regularization. Our experiences suggest that a combined norm give more stable results. We have assumed compact support to simplify the statement in Assumption 1. Under the assumption of compact support for $\beta$, the role of the derivative penalty does not seem so significant with respect to the classification performance. In general, since we only have $\|\beta\|_1 \leq \|\beta\|_2$, the bound on the $L^1$ norm alone is not sufficient to ensure the $L^2$ properties of $\beta$ and the techniques we have developed below do not necessarily work. On the other hand, we have (e.g., Gabushin, 1967; Li and Leoni, 2018) that

$$\|\beta\|_2 \leq \|\beta\|_1 + \|\beta'\|_2 \,,$$

which suggests that the proposed method is able to control the $L^2$ norm of the function more effectively by regularizing both the derivative and the $L^1$ norm of the function.

Taking into account the functional constraints in (6) leads to the following objective function

$$J_n(\beta) = \frac{1}{2}\langle \Gamma_n \beta, \beta \rangle - \langle \delta_n, \beta \rangle + P(\beta) \,, \tag{7}$$

where $P(\beta) = \lambda\|\beta\|_1 + (\eta/2)\|\beta'\|_2^2$ and $\lambda$ and $\eta$ are tuning parameters, chosen from data.

In this formulation, the derivative penalty is a standard smoothness penalty. The non-standard part is the $L^1$ penalty, which is an infinite-dimensional counterpart of the $\ell_1$ sparsity penalty. As is demonstrated below, the $L^1$ norm can be linked to a sparse solution in function space in terms of domain selection (Li and Leoni, 2018), which is a special case of functional sparsity (Wang and Kai, 2015). Arguably, the result would be easier to interpret as the localized effects are automatically identified.

We note that it is possible to use the ridge penalty ($\|\beta\|_2^2$) instead of the derivative norm. In fact, this choice makes the analysis of the optimization problem much easier as it directly controls the $L^2$ norm. In this case, the fact that $\langle \Gamma_n \beta, \beta \rangle + \tau\|\beta\|_2^2 \geq \tau\|\beta\|_2^2$ for all $\tau > 0$ is sufficient to guarantee the existence of a solution. Nevertheless, we have chosen the derivative norm that has the added consideration of smoothness of the solution, since a smoother solution is easier to interpret.

## 2.4 Properties of the regularized classifier in the population model

In this section we study the characteristics of the proposed regularized solution to the infinite-dimensional convex optimization problem. In order to simplify the discussion, we first consider the following optimization problem that replaces the sample quantities by the population counterparts. Let

$$J(\beta) = \frac{1}{2}\langle \Gamma\beta, \beta\rangle - \langle \delta, \beta\rangle + \lambda\|\beta\|_1 + \frac{\eta}{2}\|\beta'\|_2^2 \,. \tag{8}$$

The smoothness constraint with $\beta'$ reduces the feasible set of $\beta$ from $L^2(\mathcal{I})$ to a differentiable subspace $H^1(\mathcal{I}) = \{\beta \in L^2(\mathcal{I}) : \beta \text{ is absolutely continuous}, \beta' \in L^2(\mathcal{I})\}$, which is the Sobolev space $H^1 = W^{1,2}$ in standard notation (Brezis, 2011). Since $L^2(\mathcal{I}) \subset L^1(\mathcal{I})$, we consider $\mathcal{H} = \{\beta \in H^1(\mathcal{I}) : \|\beta\|_1 + \|\beta'\|_2 < \infty\}$ as the space of feasible solutions, equipped with the norm given by $\|\beta\| = \|\beta\|_1 + \|\beta'\|_2$. Then it is clear that $\mathcal{H}$ is a convex set.

Under this setting, it can be shown that $J$ is convex, so that the existence of a solution is guaranteed. Moreover, we can show that $J$ is strictly convex so the solution is unique. We summarize the result in the following proposition.

**Proposition 1.** *The function $J : \mathcal{H} \to \mathbb{R}$ defined in (8) has a global minimizer $\tilde{\beta} \in \mathcal{H}$ such that*

$$J(\tilde{\beta}) \leq J(\beta)$$

*for all $\beta \in \mathcal{H}$. Furthermore, the solution is unique.*

In order to understand the property of the solution $\tilde{\beta}$, additional characterizations of the solution beyond its existence are necessary. Note that if the objective function is differentiable, the first derivative (along with the second derivative) condition sufficiently characterizes the solution and a Newton-type algorithm can be used to find it. Non-differentiable functions require an alternative form to replace the derivative condition. Following similar arguments in Reyes (2015, ch 6) and Glowinski (1984, p. 70-71), we derive a set of conditions that $\tilde{\beta}$ satisfies in the following proposition, which could be used for developing an optimization algorithm.

**Proposition 2.** *If $\tilde{\beta}$ is the minimizer of $J(\beta)$ in (8), there exists a function $\alpha \in L^2(\mathcal{I})$ with $|\alpha(t)| \leq \lambda$ a.e. for $t \in \mathcal{I}$ such that $\tilde{\beta}$ satisfies the following relation:*

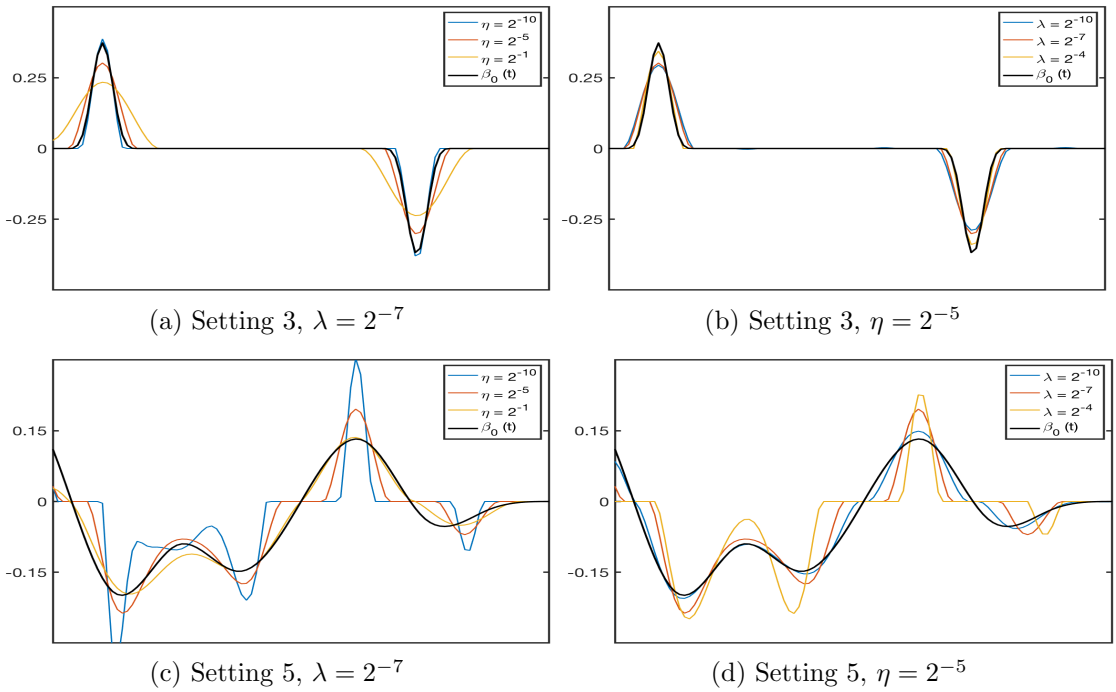
$$\langle \Gamma\tilde{\beta}, \beta\rangle + \eta\langle \tilde{\beta}', \beta'\rangle - \langle \delta, \beta\rangle + \langle \alpha, \beta\rangle = 0 \,, \quad \textit{for all} \quad \beta \in \mathcal{H} \,, \tag{9}$$

*and*

$$\begin{cases} \alpha(t) = \lambda & \textit{on } \{t \in \mathcal{I} : \tilde{\beta}(t) > 0\} \,, \\ |\alpha(t)| \leq \lambda & \textit{on } \{t \in \mathcal{I} : \tilde{\beta}(t) = 0\} \,, \\ \alpha(t) = -\lambda & \textit{on } \{t \in \mathcal{I} : \tilde{\beta}(t) < 0\} \,. \end{cases} \tag{10}$$

The equations (9) and (10) are necessary conditions for the solution to satisfy. Equation (9) means that the subgradient of the objective at the minimizer $\tilde{\beta}$ contains zero.

Figure 1: Illustration of the roles of the tuning parameters $\eta$ and $\lambda$. Each panel plots the true $\beta_0(t)$, as well as $\tilde{\beta}(t)$ with a varying tuning parameter.




(a) Setting 3, $\lambda = 2^{-7}$

(b) Setting 3, $\eta = 2^{-5}$

(c) Setting 5, $\lambda = 2^{-7}$

(d) Setting 5, $\eta = 2^{-5}$

Furthermore, Proposition 2 demonstrates the domain selection property of the solution. In order to make an analogy to the finite high dimensional case, consider a simplistic example where $\Gamma = I$, the identity operator, and $\eta = 0$ in (8). Note that, since the identity operator is not compact, this is not a realistic example from the functional data perspective but serves as a toy example. Then, the optimal solution $\tilde{\beta}$ can be given explicitly as

$$\tilde{\beta}(t) = \begin{cases} \delta(t) - \lambda & \text{if } \delta(t) > \lambda\,, \\ 0 & \text{if } \delta(t) \in [-\lambda, \lambda]\,, \\ \delta(t) + \lambda & \text{if } \delta(t) < -\lambda\,. \end{cases}$$

Since $\delta$ is continuous, the set $\{t \in \mathcal{I} : \tilde{\beta}(t) = 0\}$ will be a union of intervals and then $\tilde{\beta}$ joins the zero intervals continuously at the boundary. Hence the solution exhibits a thresholding behaviour, similar to the sparse estimators in the finite dimensional case, but on the continuous domain, thus justifying the notion of functional sparsity in the sense of domain selection. When $\eta = 0$, the optimization occurs in $L^2(\mathcal{I})$, so the boundary points of the zero intervals are not necessarily differentiable. For the general case, it is difficult to express the solution in an analytical form but the same argument holds and when $\eta > 0$ as in (9) and (10), the function values at zero intervals and non-zero intervals are joined smoothly.

Moreover, suppose that $\beta_c \in H_0^1(\mathcal{I}) = \{u \in H^1(\mathcal{I}) : u \text{ is zero at the boundary of } \mathcal{I}\}$.

Then, (9) leads to

$$\int_{\mathcal{I}}(\Gamma\tilde{\beta}-\delta+\alpha)\beta_c+\int_{\mathcal{I}}(\eta\tilde{\beta}')\beta_c'=0\,,\quad \text{for all}\quad \beta_c\in H_0^1(\mathcal{I})\,. \tag{11}$$

This implies (e.g., Brezis, 2011) that $\tilde{\beta}'$, the derivative of the solution, is also absolutely continuous and satisfies

$$\eta\tilde{\beta}''=\Gamma\tilde{\beta}-\delta+\alpha\,,\quad a.e.\,, \tag{12}$$

which, together with (10), gives an additional necessary condition for the solution.

The following observation helps us to determine the upper bound of the tuning parameter $\lambda$. Suppose that $\delta\in L^\infty(\mathcal{I})$. Then, we have $\langle\delta,u\rangle\le\|\delta\|_\infty\|u\|_1$. In addition, it can be shown that the solution $\tilde{\beta}$ satisfies

$$\langle\Gamma\tilde{\beta},\tilde{\beta}\rangle+\eta\langle\tilde{\beta}',\tilde{\beta}'\rangle+\lambda\|\tilde{\beta}\|_1=\langle\delta,\tilde{\beta}\rangle\le\|\delta\|_\infty\|\tilde{\beta}\|_1\,.$$

It follows that

$$\langle\Gamma\tilde{\beta},\tilde{\beta}\rangle+\eta\langle\tilde{\beta}',\tilde{\beta}'\rangle+(\lambda-\|\delta\|_\infty)\|\tilde{\beta}\|_1\le 0\,.$$

Since the first two terms are non-negative, if $\|\delta\|_\infty\le\lambda$, then $\tilde{\beta}=0$. This gives a range of $\lambda$ values for a non-zero solution and this depends on the magnitude of $\delta$. Figure 1 illustrates the effect of the tuning parameters $\eta$ and $\lambda$ on $\tilde{\beta}$, under two different scenarios for $\beta_0$: sparse and non-sparse. The curves in the top and bottom panels respectively correspond to the simulation settings 3 and 5 in Section 3. We observe that for fixed $\lambda$, larger $\eta$ gives smoother estimates and for fixed $\eta$, larger $\lambda$ gives a more sparse solution.

Suppose that $\mathrm{Ker}(\Gamma)=\{0\}$. If $\|\Gamma^{-1}\delta\|_2<\infty$, then $\beta_0=\Gamma^{-1}\delta\in L_2(\mathcal{I})$ is the unique solution to (4). Denote by $\tilde{\beta}$ the minimizer of $J$ in (8). The following proposition shows that the regularized solution $\tilde{\beta}$ approximates the underlying model solution to (4) whether the latter is finite or not. We write $\tilde{\lambda}=(\lambda,\eta)$, and $\tilde{\lambda}\to 0$ means that $\lambda\to 0$ and $\eta\to 0$.

**Proposition 3.** *Fix any $u\in L_2(\mathcal{I})$. If $\|\Gamma^{-1}\delta\|_2<\infty$, then $\langle\tilde{\beta},u\rangle\to\langle\beta_0,u\rangle$ as $\tilde{\lambda}\to 0$. If $\|\Gamma^{-1}\delta\|_2=\infty$, then $\|\tilde{\beta}\|\to\infty$, as $\tilde{\lambda}\to 0$.*

Using Proposition 3 together with (1), we can study the misclassification error of the regularized classifier. We show below that our regularized classifier mimics the behaviour of the optimal classifier in terms of the misclassification rate.

**Proposition 4.** *Assume that $X$ is Gaussian with covariance operator $\Gamma$. The misclassification probability of the regularized classifier $\tilde{\beta}$, denoted by $err_X(\tilde{\beta})$, converges to $1-\Phi(\|\Gamma^{-1/2}\delta\|_2/2)$ as $\tilde{\lambda}\to 0$.*

**Remark 1.** *It may be suspected that the two classes have different covariance functions, say $\Gamma_0$ and $\Gamma_1$, in which case a full generalization of our linear approach is not trivial. For example, a direct generalization of (2) replaces the denominator by $\langle\tilde{\Gamma}\beta,\beta\rangle$ where $\tilde{\Gamma}=\pi_0\Gamma_0+\pi_1\Gamma_1$, the pooled covariance function. Then the same formulation as (8), with $\Gamma$ replaced by $\tilde{\Gamma}$, follows. Alternatively, we can consider the two-dimensional projection*

*approach for quadratic discriminant analysis proposed by Gaynanova and Wang (2019) in the multivariate setting. In our functional framework, we would optimize the following two objective functions*

$$J_i(\beta) = \frac{1}{2}\langle(\Gamma_i\beta, \beta\rangle - \langle\delta, \beta\rangle + \lambda_i\|\beta\|_1 + \frac{\eta_i}{2}\|\beta'\|_2^2,$$

*for $i = 0, 1$. The classification will be then made based on the two linear maps with the respective solutions. An advantage of this method over standard quadratic discriminant analysis is that the computational complexity is essentially the same as the homoscedastic case being considered here. In either case, the theoretical guarantee on the domain selection and classification accuracy need to be carefully investigated in the functional context.*

## 2.5 Properties of the regularized classifier in the sample model

The development in the previous section justifies the use of the $L^1$ penalty in our formulation, by ensuring that we have a well-defined regularization problem (Proposition 1) with the desired sparsity in the solution (Proposition 2). We can show that these properties also hold for the empirical estimator, denoted by $\hat{\beta}$, the minimizer of $J_n$ defined in (7).

In the sample case, our formulation of $J_n$ is based on two estimators $\delta_n$ and $\Gamma_n$. Provided that the sample covariance operator $\Gamma_n$ is non-negative definite and $\delta_n \in L^2(\mathcal{I})$, it can be shown that Propositions 1 and 2 hold for $J_n$ and $\hat{\beta}$. In fact, the proofs are essentially the same, thus we omit the formal statements of the Corollaries here.

In order to study the limit of the empirical estimator $\hat{\beta}$ and the corresponding misclassification probability, we make the following assumptions. Assumption 3 says that some consistent estimators are available for $\delta_n$ and $\Gamma_n$ (or $\gamma_n$) and Assumption 4 assumes that the tuning parameters decrease not too rapidly as the sample size gets large.

**Assumption 3.** *Let $(\delta_n, \gamma_n)$ be consistent estimators such that*

$$\|\gamma_n - \gamma\|_\infty = O_p(a_n), \qquad \|\delta_n - \delta\|_\infty = O_p(b_n)\,,$$

*for some sequences $a_n, b_n$ converging to 0.*

**Assumption 4.** *Let $\lambda = \lambda_n \to 0, \eta = \eta_n \to 0$ with $\lambda_n/\eta_n = O(1), a_n/\eta_n = o(1)$ and $b_n/\eta_n = o(1)$.*

For sufficiently dense data as we consider here, the standard sample mean and sample covariance function can be used with a root-$n$ convergence. Alternatively, any preferred smoothing methods could be employed, then the convergence rates are given as a function of smoothing parameters used, for example, see Yao et al. (2005); Li and Hsing (2010) for local linear smoothers. Normally the error for covariance estimation is larger than that for the mean estimation ($a_n \geq b_n$). Zhang and Wang (2016) generalize those results under a general design scheme and show that a root-$n$ rate can be achieved for dense and ultra-dense data, in which case the uniform convergence rates can be $a_n = b_n = (\log n/n)^{1/2}$.

**Proposition 5.** *Fix any $u \in L_2(\mathcal{I})$. Suppose that Assumptions 1-4 hold. If $\|\Gamma^{-1}\delta\|_2 < \infty$, then $\langle\hat{\beta}, u\rangle \to \langle\beta_0, u\rangle$ in probability as $n \to \infty$. On the other hand, if $\|\Gamma^{-1}\delta\|_2 = \infty$, then $\|\hat{\beta}\| \to \infty$ in probability as $n \to \infty$.*

**Proposition 6.** *Assume that $X$ is Gaussian with covariance operator $\Gamma$. Under Assumptions 1-4, the misclassification probability of the regularized classifier $\hat{\beta}$, denoted by $err_X(\hat{\beta})$, converges to $1 - \Phi(\|\Gamma^{-1}\delta\|_2/2)$ in probability as $n \to \infty$.*

### 2.6 Numerical algorithms

A fundamental strategy to solve infinite-dimensional optimization problems could be said to be either to discretize-and-optimize or to optimize-and-discretize (Reyes, 2015). For example, Glowinski (1984) analysed numerical methods based on the former approaches using piecewise linear and quadratic approximations on $H^1(\mathcal{I})$, while Reyes (2015) includes examples of the latter. Generally, the former is easier, while the latter may give a more elegant solution. According to Proposition 2, the second approach requires solving equations (9), or (12), and (10) simultaneously. Based on these equations, Newton-like update steps for both $\tilde{\beta}$ and $\alpha$ could be derived. However, as this involves further development of the testable conditions at the iteration to be adapted to a proper function space (Reyes, 2015), we leave it for future work.

Here we have used piecewise linear approximations in our implementation with equidistant grid points. We use estimates from the method by Yao et al. (2005) for $\Gamma_n$ and $\delta_n$ and evaluate the derivative by the finite-difference approximation. This strategy, combined with the rather simple form of the optimization problem, leads to a lasso-type problem, for which many efficient optimization algorithms are available. We implement the coordinate descent algorithm by Fu (2012) to solve our discretized problem.

### 2.7 Implications in functional linear regression

Although we started with a classification problem, our regularization method with an $L^1$-type penalty can be motivated from a regression point of view. Consider a functional regression problem with scalar response $Y$ and functional predictor $X \in H = L^2(\mathcal{I})$. For simplicity, assume that $E(Y) = 0$ and $E(X) = 0$ and consider a linear regression given by

$$Y = \int_{\mathcal{I}} \beta(t)X(t)\,\mathrm{d}t + \epsilon\,,$$

where $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$. As before, the covariance operator of $X$ is denoted by $\Gamma$. In addition, let $\Lambda : L^2(\mathcal{I}) \to \mathbb{R}$ denote the covariance operator between $X$ and $Y$, defined as

$$\Lambda(u) = \int_{\mathcal{I}} E\{X(s)Y\}u(s)\,\mathrm{d}s\,.$$

Then, the population least squares criterion can be expressed as

$$E\{(Y - \langle\beta, X\rangle)^2\} = \langle\Gamma\beta, \beta\rangle - 2\Lambda(\beta) + const\,,$$

from which it follows that if $\tilde{\beta}$ is the minimizer if and only if it satisfies

$$\langle \Gamma\tilde{\beta}, u\rangle = \Lambda(u) \qquad \text{for all } u \in H\,. \tag{13}$$

To be consistent with the earlier notation, write $\Lambda(u) = \langle \Delta, u\rangle$ where $\Delta(\cdot) = E\{X(\cdot)Y\}$. Then, (13) can be expressed as

$$\langle \Gamma\tilde{\beta}, u\rangle = \langle \Delta, u\rangle \qquad \text{for all } u \in H\,.$$

Note that when $Y$ is binary with values in $\{-1, 1\}$, $\Delta$ is equal to $\delta$ in (4), the difference in the group means in $X$ between two classes.

The existence of the solution requires the so-called Picard condition (Cardot et al., 2003; Hsing and Eubank, 2015)

$$\sum_{j=1}^{\infty} \frac{\langle E\{X(\cdot)Y\}, \phi_j\rangle^2}{\theta_j^2} = \sum_{j=1}^{\infty} \frac{\langle \Delta, \phi_j\rangle^2}{\theta_j^2} < \infty\,,$$

which is equivalent to $\|\Gamma^{-1}\delta\|_2 < \infty$ in the case of classification problems (c.f., Section 2.2).

The above discussion shows that even though the motivation differs, the underlying problem for functional linear regression is equivalent to that of linear classification. Hence, our methodology is equally applicable in a regression setting and offers a regularization method to estimate the coefficient function $\beta$. Given a sample of size $n$ of $(Y_i, X_i), i = 1, \ldots, n$, denote the sample version of the operators of $\Gamma$ and $\Lambda$ by $\Gamma_n$ and $\Lambda_n$, respectively, defined as

$$\Gamma_n(u)(t) = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, u\rangle X_i(t)\,, \quad \Lambda_n(u) = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, u\rangle Y_i = \langle \Delta_n, u\rangle \qquad u \in L^2(\mathcal{I})\,.$$

where $\Delta_n(\cdot) = (1/n)\sum_{i=1}^{n} X_i(\cdot)Y_i$. Assuming that the true coefficient function $\beta$ is a smooth function, Cardot et al. (2003) proposed a generalization of a ridge regression with an $L^2$ derivative penalty as

$$\frac{1}{2}\langle \Gamma_n\beta_K, \beta_K\rangle - \langle \Delta_n, \beta_K\rangle + \frac{\eta}{2}\|\beta_K^{(m)}\|_2^2\,,$$

where $\beta_K = \sum_{j=1}^{K} c_k B_k$ with $B_k$ is a B-spline basis function and $\beta_K^{(m)}, m \geq 1$ is the $m$th derivative of $\beta_K$. Our analysis suggests that adding an $L^1$ penalty would lead to a sparse estimator that consistently estimates the true function, whether or not the latter is sparse.

**Remark 2.** *Although the theoretical framework differs, our main consideration of interpretability is not limited to this type of problem. For example, semi-parametric problems of single-index or multiple-index regression or classification can be expressed as*

$$Y = F\left(\sum_k \langle X, \beta_k\rangle\right) + \epsilon$$

*with a known or unknown link function $F$. Then an $L^1$-type functional regularization for $\beta_k$ can be developed with an appropriate loss function. We leave such a generalization to our future work.*

# 3 Numerical studies

## 3.1 Simulation models

We use simulated curves under six different underlying population scenarios to assess the performance of the proposed method and compare it with existing approaches. Settings 1 – 5 represent scenarios where the underlying structure follows model (4) in Section 2.2 and Setting 6 is borrowed from Delaigle and Hall (2012). The proposed approach, which we name SFLDA (Sparse Functional Linear Discriminant Analysis), is compared with eight different approaches: the non-sparse version of the proposed approach (FLDA) that sets the tuning parameter $\lambda = 0$ in (7); ridge FLDA (RFLDA) proposed by Kraus and Stefanucci (2018); partial least squares (PLS) by Delaigle and Hall (2012); the three Bayes methods (B-Gauss, B-NPD, B-NPR) in Dai et al. (2017); the quadratic discriminant analysis (QDA) by Galeano et al. (2015); functional logistic regression (Logistic) (Müller, 2005).

The first four settings have a sparse underlying discriminant function. The sparse discriminant functions, $\beta_1(t)$ and $\beta_2(t)$, are generated with 33 cubic B-spline basis functions $B_{i,4}$ $(i = 1, \ldots, 33)$ with the knots $\xi_j = j/30, j = 0, \ldots, 30$, where $\xi_0$ and $\xi_{30}$ are the boundary knots. We set $\beta_1(t) = 0.2B_{5,4}(t) + 0.2B_{28,4}(t)$ for Settings 1 and 2, and $\beta_2(t) = 0.2B_{5,4}(t) - 0.2B_{24,4}(t)$ for Settings 3 and 4. For the non-sparse $\beta_3(t)$ for Setting 5, we use 8 cubic B-spline basis and set

$$\beta_3(t) = 0.1B_{1,4}(t) - 0.3B_{3,4}(t) - 0.2B_{5,4}(t) + 0.2B_{7,4}(t) - 0.1B_{8,4}(t).$$

For the covariance function $\Gamma$, we used the Matérn covariance function:

$$\gamma(s,t) = \sigma^2 \frac{2^{1-\nu}}{G(\nu)} \left\{ (2\nu)^{1/2} \frac{|s-t|}{\rho} \right\}^{\nu} K_\nu \left\{ (2\nu)^{1/2} \frac{|s-t|}{\rho} \right\},$$

where $G$ is the gamma function, $K_\nu$ is the modified Bessel function of the second kind, for which the R-function `besselK` was used, and the parameters were set as $\sigma = 1$, $\rho = 0.2$, and $\nu = 3$. We set the mean of the first group as $\mu_0(t) = 5t + \sum_{j=1}^{5}(c_j/j)\phi(t)$, where $\phi(t) = 2^{1/2}\sin(\pi j t)$, $(c_1, \ldots, c_5) = (2.19, -0.18, -0.19, -2.51, -0.56)$, and the second group mean as $\mu_1(t) = \mu_0(t) + \delta(t)$. Here the mean difference function $\delta(t)$ is determined by $\delta(t) = \int_{\mathcal{I}} \Gamma(t,s)\beta_k(s)ds$, $k = 1, 2, 3$.

The data were assumed to be available on a fine grid. We used $T = 100$ equispaced grid points on $[0, 1]$. For grid points $t_j, j = 1, \ldots, T$ , let $\tilde{\Gamma}$ be the $T \times T$ matrix with $(i, j)$ entry $\Gamma(t_i, t_j)$, and let $m_0$ and $m_1$ be the $T \times 1$ vectors with $i$th entry $\mu_0(t_i)$ and $\mu_1(t_i)$, respectively. Data were generated by $x_0 = m_0 + \tilde{\Gamma}^{1/2}e_0$ and $x_1 = m_1 + \tilde{\Gamma}^{1/2}e_1$ where $e_0$ and $e_1$ are the vectors of independent noise for each class, generated from either

Table 1: Test error rates (%) for simulated data based on 100 repetitions. Standard errors are in the parentheses.

| Setting | SFLDA | FLDA | RFLDA | PLS | B-Gauss | B-NPD | B-NPR | Logistic | QDA |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 33.3 | 33.4 | 33.6 | 33.7 | 33.4 | 34.6 | 35.1 | 34.7 | 34.2 |
| | (0.22) | (0.22) | (0.24) | (0.23) | (0.24) | (0.25) | (0.25) | (0.27) | (0.28) |
| 2 | 33.5 | 33.5 | 33.8 | 33.7 | 33.6 | 34.7 | 35.4 | 34.9 | 34.2 |
| | (0.21) | (0.20) | (0.20) | (0.20) | (0.20) | (0.22) | (0.24) | (0.25) | (0.22) |
| 3 | 37.2 | 36.9 | 37.5 | 37.1 | 37.2 | 37.8 | 39.0 | 38.5 | 38.0 |
| | (0.23) | (0.20) | (0.20) | (0.21) | (0.22) | (0.25) | (0.29) | (0.27) | (0.25) |
| 4 | 37.7 | 37.3 | 38.1 | 37.6 | 37.5 | 39.1 | 40.3 | 39.2 | 38.2 |
| | (0.24) | (0.22) | (0.24) | (0.23) | (0.24) | (0.27) | (0.30) | (0.28) | (0.26) |
| 5 | 3.2 | 3.1 | 3.1 | 3.1 | 3.2 | 3.5 | 3.7 | 3.6 | 4.2 |
| | (0.08) | (0.07) | (0.07) | (0.07) | (0.08) | (0.10) | (0.10) | (0.14) | (0.08) |
| 6 | 3.3 | 3.2 | 3.3 | 3.1 | 4.0 | 3.4 | 2.5 | 2.5 | 3.6 |
| | (0.08) | (0.07) | (0.07) | (0.07) | (0.08) | (0.08) | (0.10) | (0.10) | (0.14) |

Table 2: Average norm difference ($\|\hat{\beta} - \beta_0\|_j$) between the estimated discriminant function and the true function. Numbers are multiplied by 100.

| Setting | $j$ | SFLDA | FLDA | RFLDA | PLS |
|---|---|---|---|---|---|
| 1 | 1 | 5.37 (0.19) | 7.97 (0.10) | 8.81 (0.13) | 8.20 (0.09) |
| | 2 | 9.64 (0.22) | 10.07 (0.11) | 11.29 (0.18) | 10.14 (0.10) |
| 3 | 1 | 5.84 (0.18) | 8.10 (0.11) | 9.00 (0.12) | 8.30 (0.11) |
| | 2 | 10.39 (0.22) | 10.40 (0.14) | 11.69 (0.18) | 10.26 (0.11) |
| 5 | 1 | 6.47 (0.14) | 4.15 (0.11) | 5.78 (0.17) | 4.50 (0.10) |
| | 2 | 8.21 (0.18) | 5.25 (0.13) | 7.23 (0.20) | 5.64 (0.11) |

a standard normal for Settings 1, 3, and 5 or a $t$-distribution with 5 degrees of freedom for Settings 2 and 4.

As mentioned earlier, Setting 6 is borrowed from Delaigle and Hall (2012). Each curve from the $i$th group is generated as $x_i = \sum_{j=1}^{40}(j^{-1}Z_{ij} + \mu_{ij})\phi_j(t)$, where $Z_{ij}$ are centred exponential variables, $(\mu_{01}, \ldots, \mu_{06}) = (0, -0.5, 1, -0.5, 1, -0.5)$, and $(\mu_{11}, \ldots, \mu_{16}) = (0, -0.75, 0.75, -0.15, 1.4, 0.1)$.

We generated $n = 100$ curves for each group to train each classifier and independently generated 300 curves for each group to evaluate the misclassification error. The tuning parameters in each method are chosen via 5-fold cross-validation within the training data based on the misclassification error.

Table 3: Mean test errors (%) from real data examples, based on 30 repetitions. Standard errors are in the parentheses.

| Data | SFLDA | FLDA | RFLDA | PLS | B-Gauss | B-NPD | B-NPR | Logistic | QDA |
|---|---|---|---|---|---|---|---|---|---|
| Tecator | 5.0 | 3.4 | 7.4 | 4.8 | 6.3 | 4.2 | 4.8 | 1.8 | 3.3 |
| | (0.44) | (0.41) | (0.68) | (0.46) | (0.48) | (0.42) | (0.59) | (0.27) | (0.42) |
| Wine | 10.3 | 10.0 | 9.2 | 9.2 | 11.1 | 10.4 | 9.9 | 8.7 | 11.6 |
| | (0.81) | (0.74) | (0.72) | (0.81) | (0.91) | (0.89) | (0.57) | (0.65) | (0.80) |
| Growth | 6.0 | 5.7 | 8.3 | 6.1 | 6.2 | 7.2 | 7.1 | 5.9 | 6.7 |
| | (0.39) | (0.39) | (0.43) | (0.32) | (0.44) | (0.36) | (0.40) | (0.48) | (0.40) |
| dGrowth | 7.0 | 6.4 | 7.4 | 6.3 | 7.4 | 8.1 | 7.6 | 7.1 | 9.5 |
| | (0.57) | (0.47) | (0.52) | (0.52) | (0.48) | (0.49) | (0.55) | (0.56) | (0.60) |
| Wheat | 0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 0.2 | 0.2 | 0.2 |
| | (0.09) | (0.00) | (0.33) | (0.00) | (0.09) | (0.00) | (0.13) | (0.17) | (0.13) |
| Phoneme | 19.7 | 21.6 | 20.9 | 21.4 | 23.7 | 23.8 | 24.2 | 22.1 | 24.4 |
| | (0.53) | (0.59) | (0.53) | (0.58) | (0.60) | (0.57) | (0.60) | (0.62) | (0.65) |

## 3.2 Results

Table 1 lists test errors based on 100 repetitions. The proposed approach is generally competitive in all settings, even though the differences among the methods do not stand out. A main benefit of the proposed sparse estimation can be seen in its ability to estimate zero regions in the discriminant function. Table 2 displays the norm differences of the estimated discriminant function $\hat{\beta}(t)$ by the proposed method with sparse functional penalty, its non-sparse version, ridge regularization, and partial least squares, in Settings 1, 3, and 5. The proposed method has the lowest error with respect to both norms in the two sparse settings (1 and 3), while its non-sparse version and partial least squares are better in the non-sparse setting (5). Also shown in Figure 2 are the estimated curves for each of the four methods for Settings 1, 3, and 5, along with the median in solid black and the true discriminant curve in a dot-dashed black line. The median curve of the estimators from the proposed method resembles the true function the most in Settings 1 and 3, successfully identifying the non-signal regions. The estimators with a ridge penalty show larger variability than other methods, which suggests that a ridge-type regularization in functional classification does not necessarily reduce variance, unlike in ridge regression.

## 3.3 Real data examples

We tested the same set of the functional classification methods as in the previous section for real data examples. The Tecator dataset (Delaigle and Hall, 2012) has 240 curves of near infrared absorbance spectra (850 - 1050 nm) of finely chopped meat, using a Tecator Infratec Food & Feed Analyser. Here the groups are defined according to the fat content. The Wine spectra dataset (Dai et al., 2017) contains 123 samples of mid infrared spectra (4000 - 400 $cm^{-1}$), which are divided into two groups based on the alcohol content level. The Growth dataset (Gasser et al., 1984; Sheehy et al., 1999) has height growth curves of 112 boys and 120 girls from births to the 18th year. We also analysed the velocity of the Growth curves. The Wheat data (Kalivas, 1997) have near infrared spectra of 100 wheat samples, divided into two groups according to the protein content. The Phoneme data are log-periodograms constructed from digitized speech of

Figure 2: Discriminant functions from 100 repetitions of the simulation study estimated by SFLDA, FLDA, RFLDA, and PLS, for Settings 1, 3, and 5. The thicker black line in each plot represents the median of the estimated curves, while the dot-dashed curve is the true discriminant function.

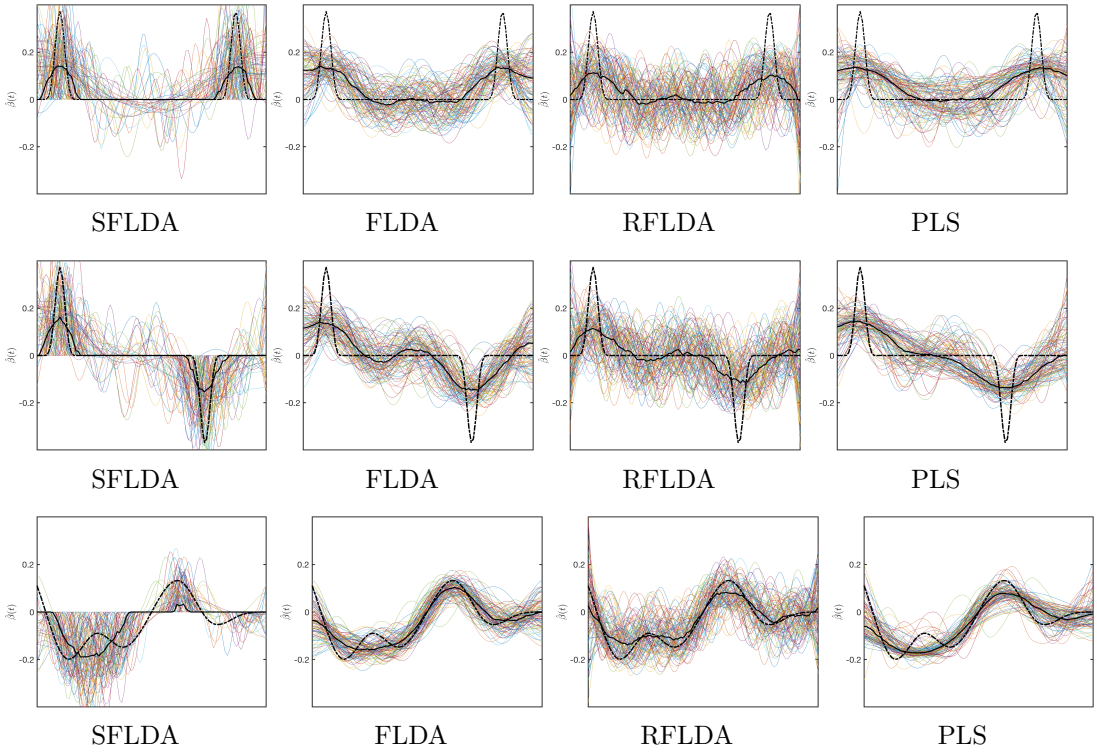
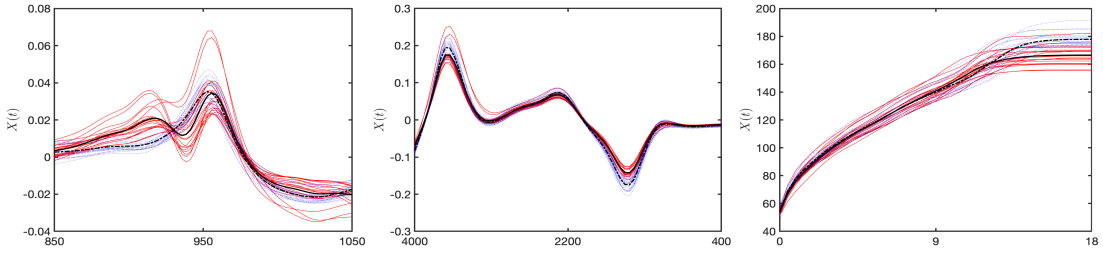
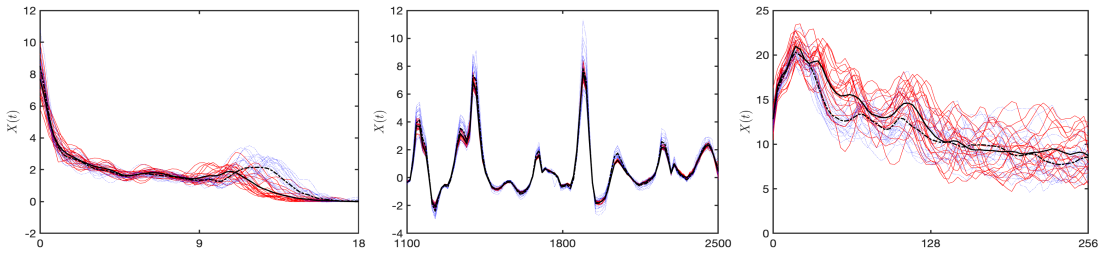


Figure 3: For each real data example, 20 sample curves from each class are shown along with average curves. Solid and dotted curves represent curves from different classes.



(a) Tecator (b) Wine (c) Growth



(d) dGrowth (e) Wheat (f) Phoneme

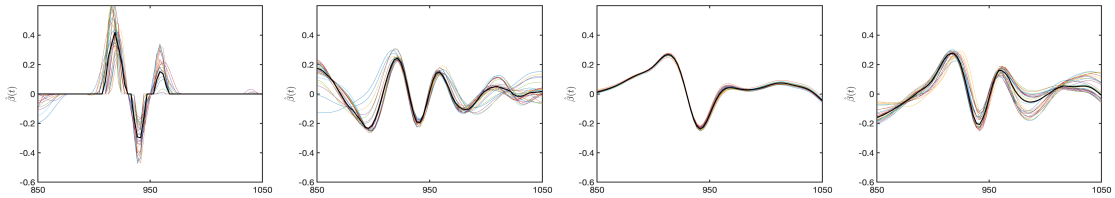
two different sounds "aa" and "ao", as described in Hastie et al. (2009). Each panel of Figure 3 respectively shows 40 randomly selected curves from each data set, with each class shown with different colors.

To estimate test errors, we split into 2/3 training and 1/3 testing data, and tune each method in the same way as the simulation study, and repeat this process 30 times. Mean test errors are displayed in Table 3. While there are no meaningful differences among the test errors of these methods, we can see that the proposed method achieves compatible accuracies with a much better ability to find localized discriminant regions, as can be seen in Figure 4. In particular, we have found that the selected regions of Tecator, Wine, and Phoneme are consistent with findings from the spectroscopy literature. Specifically, according to the body composition study (Conway et al., 1984), fat has high absorbance around 930nm, the absorbance region for ethanol is concentrated in 1200 - 850 $cm^{-1}$ (Debebe et al., 2017), and it has been found by Hastie et al. (1995) that the discriminating feature in phonemes are in the low frequencies about 500 - 1000Hz, corresponding to the frequencies 16–32. All these known regions are consistent with the estimates from the proposed method in the figure.

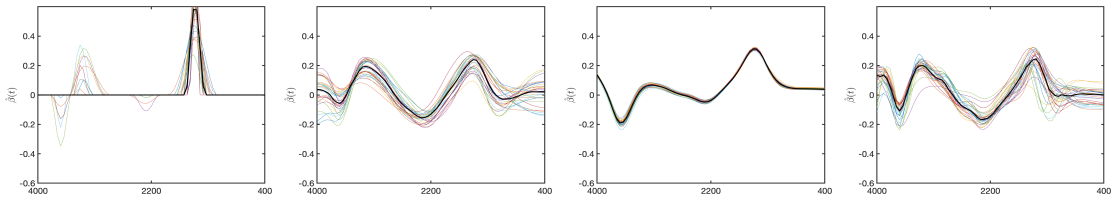
## Acknowledgment

We are grateful to Aurore Delaigle and Xiongtao Dai for sharing the Matlab codes for compared methods. We are also grateful to two referees, an associate editor and the editor for careful reading and helpful suggestions.

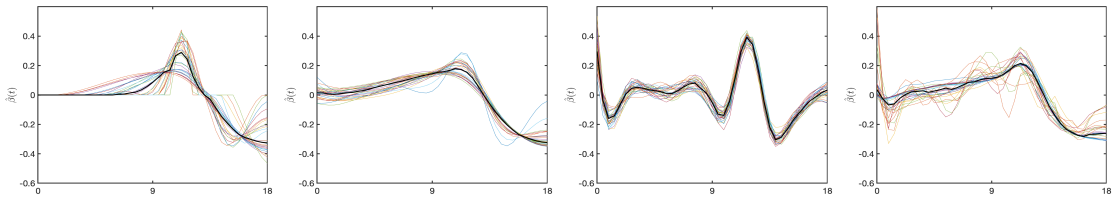
Figure 4: Estimated discriminant curves for real data examples from 30 repetitions. The median curve is shown as a thicker black solid line.



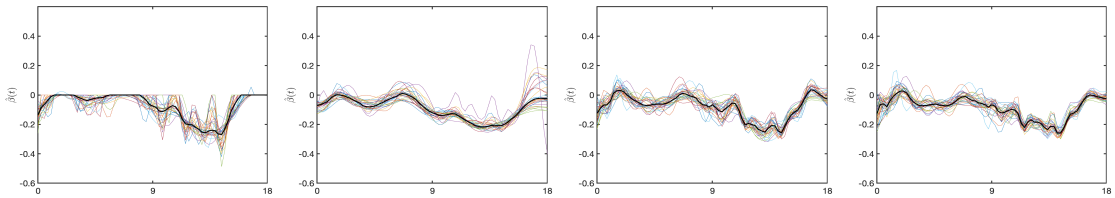
Tecator - SFLDA FLDA RFLDA PLS



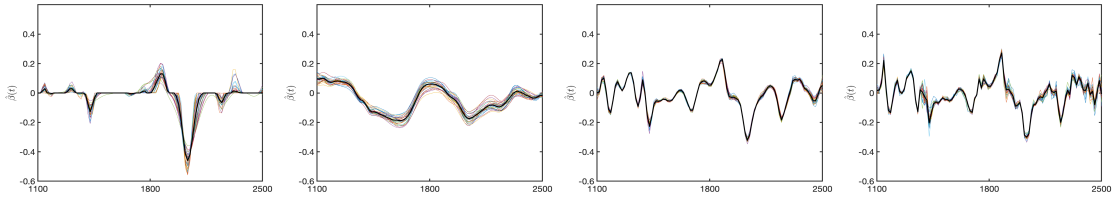
Wine - SFLDA FLDA RFLDA PLS



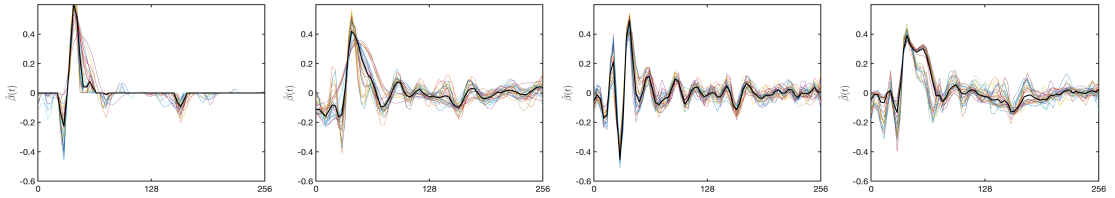
Growth - SFLDA FLDA RFLDA PLS



Dgrowth - SFLDA FLDA RFLDA PLS



Wheat - SFLDA FLDA RFLDA PLS



Phoneme - SFLDA FLDA RFLDA PLS

## Supplementary material

Supplementary material includes additional details on relevant background material and proofs, as well as an example of matlab codes for implementation.

## References


Berrendero, J. R., A. Cuevas, and J. L. Torrecilla (2018). On the use of reproducing kernel Hilbert spaces in functional classification. *J. Am. Statist. Assoc. 113(523)*, 1210–1218.

Bongiorno, E. G. and A. Goia (2016). Classification methods for Hilbert data based on surrogate density. *Comput. Stat. Data Anal. 99*, 204–222.

Brezis, H. (2011). *Functional analysis, Sobolev spaces and partial differential equations*. Springer.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Stat. Probabil. Lett. 45(5)*, 11–22.

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

Cardot, H., A. Mas, and P. Sarda (2007). CLT in functional linear regression models. *Probab. Theory Rel. 138*, 325–361.

Conway, J. M., K. H. Norris, and C. Bodwell (1984). A new approach for the estimation of body composition: infrared interactance. *Am. J. Clin. Nutr. 40*(6), 1123–1130.

Crambes, C., A. Kneip, and P. Sarda (2009). Smoothing spline estimators for functional linear regression. *Ann. Statist. 37(1)*, 35–72.

Dai, X., H.-G. Müller, and F. Yao (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika 104(3)*, 545–560.

Debebe, A., M. Redi-Abshiro, and B. S. Chandravanshi (2017). Non-destructive determination of ethanol levels in fermented alcoholic beverages using fourier transform mid-infrared spectroscopy. *Chemistry Central Journal 11*(1), 1–8.

Delaigle, A. and P. Hall (2010). Defining probability density for a distribution of random functions. *Ann. Statist. 38(2)*, 1171–1193.

Delaigle, A. and P. Hall (2012). Achieving near perfect classification for functional data. *J. R. Statist. Soc.* B *74(2)*, 267–286.

Fu, W. J. (2012, Feb). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Stat. 7*(3), 397–416.

Gabushin, V. (1967). Inequalities for the norms of a function and its derivatives in metric $l_p$. *Mathematical Notes of the Academy of Sciences of the USSR 1*, 194–198.

Galeano, P., E. Joseph, and R. E. Lillo (2015). The Mahalanobis distance for functional data with applications to classification. *Technometrics 57*(2), 281–291.

Gasser, T., H. G. Müller, W. Köhler, L. Molinari, and A. Prader (1984). Nonparametric regression analysis of growth curves. *Ann. Statist. 12*, 210–229.

Gaynanova, I. and T. Wang (2019). Sparse quadratic classification rules via linear dimension reduction. *J. Multivar. Anal. 169*, 278–299.

Glowinski, R. (1984). *Numerical methods for nonlinear variational problems*. Springer Series in Computational Physics. Springer-Verlag.

Hall, P. and G. Hooker (2016). Truncated linear models for functional data. *J. R. Statist. Soc.* B *78(3)*, 637–653.

Hall, P., D. S. Poskitt, and B. Presnell (2001). A functional data-analytic approach to signal discrimination. *Technometrics 43(1)*, 1–9.

Hastie, T., A. Buja, and R. Tibshirani (1995). Penalized discriminant analysis. *Ann. Statist. 23*, 73–102.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (the 2nd ed.). Springer.

Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley series in probability and statistics. Wiley.

James, G. M. and T. J. Hastie (2001). Functional linear discriminant analysis for irregularly sampled curves. *J. R. Statist. Soc.* B *63(3)*, 533–550.

James, G. M., J. Wang, and J. Zhu (2009). Functional linear regression that's interpretable. *Ann. Statist. 37*, 2083–2108.

Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems 37*(2), 255–259.

Kneip, A., D. Poß, and P. Sarda (2016). Functional linear regression with points of impact. *Ann. Statist. 44(1)*, 1–30.

Kraus, D. and M. Stefanucci (2018). Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika 106*(1), 161–180.

Li, T. and G. Leoni (2018). $L^1$ regularization for compact support. *SIAM Undergraduate Research Online 11*, 101–121.

Li, Y. and T. Hsing (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functionallongitudinal data. *Ann. Statist. 38*(6), 3321–3351.

Lin, Z., J. Cao, L. Wang, and H. Wang (2017). Locally sparse estimator for functional linear regression models. *J. Comput. Graph. Stat. 26(2)*, 306–318.

Lin, Z., L. Wang, and J. Cao (2016). Interpretable functional principal component analysis. *Biometrics 72*, 846–854.

Martin-Barragan, B., R. Lillo, and J. Romo (2014). Interpretable support vetor machines for functional data. *Eur. J. Oper. Res. 232*, 146–155.

Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scand. J. Stat. 32*(2), 223–240.

Picheny, V., R. Servien, and N. Villa-Vialaneix (2019). Interpretable sparse sir for functional data. *Stat. Comput. 29*, 255–267.

Reyes, J. D. L. (2015). *Numerical PDE-constrained optimization*. Springer briefs in optimization. Springer.

Roche, A. (2018). Variable selection and estimation in multivariate functional linear regression via the lasso. Technical report, hal-01725351.

Sheehy, A., T. Gasser, L. Molinari, and R. H. Largo (1999). An analysis of variance of the pubertal and midgrowth spurts for length and width. *Ann. Hum. Biol. 26*(4), 309–331.

Shin, H. (2008). An extension of Fisher's discriminant analysis for stochastic processes. *J. Multivar. Anal. 99*, 1191–1216.

Tian, T. S. and G. M. James (2013). Interpretable dimension reduction for classifying functional data. *Comput. Stat. Data Anal. 57*, 282–296.

Tu, C. Y., J. Park, and H. Wang (2020). Estimation of functional sparsity in non-parametric varying coefficient models for longitudinal data analysis. *Stat. Sinica 29*, 439–465.

Wang, H. and B. Kai (2015). Functional sparsity: Global versus local. *Stat. Sinica 25*, 1337–1354.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc. 100*(470), 577–590.

Zhang, X. and J.-L. Wang (2016). From sparse to dense functional data and beyond. *Ann. Statist. 44(5)*, 2281–2321.

Zhou, J., N.-Y. Wang, and N. Wang (2013). Functional linear model with zero-value coefficient function at sub-regions. *Stat. Sinica 23*, 25–50.

Zhou, Y., R. Ogden, and P. Reiss (2012). Wavelet-based LASSO in functional linear regression. *J. Comput. Graph. Stat. 21(3)*, 600–617.