Penalized discriminant analysis for multivariate functional data

Xiaoyan Sun

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2022

Committee:
Eardi Lila
Chongzhi Di

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Penalized discriminant analysis for multivariate functional data

Xiaoyan Sun

Chair of the Supervisory Committee:
Eardi Lila
Biostatistics

We introduce a penalized discriminant analysis method for multivariate functional data supported on compact one-dimensional domains, motivated by an application that aims to identify subjects with poor cognitive status from diffusion MRI data. By leveraging a connection to the optimal scoring problem, we bypass minimizing complex objective functions directly and recast the problem into a penalized regression framework. The proposed formulation leads to an efficient model computationally. Simulation studies showed that the univariate classifier achieves satisfactory performance compared to existing methods. The multivariate classifier achieved adequate prediction performance on the challenging diffusion MRI dataset. As an extension of the methodology proposed, we also present a multi-class classifier that can accommodate multiple outcome categories.

Contents

1 Introduction

1.1 Background: functional data analysis

Functional data analysis (FDA) is the branch of statistics concerned with the analysis of functions, curves, and images. The term functional data analysis was coined by Ramsay (1982) and Ramsay & Dalzell (1991), but the study area is much older and dates back to 1950s at least (Grenander, 1950; Rao, 1958).

It's a fast-growing research field with wide-ranging applications. With the advancement of modern technology, more data is being recorded continuously during some time interval, such as the electroencephalogram (EEG) measurements, temperatures in an hour, or stock prices in a day. One could discretize the data and turn it into a classic multivariate modeling problem. But this is not optimal because important information would be lost. In fact, multivariate vectors can be randomly permuted without affecting its meaning, but permuting continuously indexed data will render it meaningless.

We can think of these data types as functions since they are curves supported on a one-dimensional domain, hence termed functional data. We also have functional data supported on a two-dimensional domain, such as an image of a cat or pollution levels in a city. There are other domains of interest such as manifolds, where examples include brain imaging and global temperature patterns.

1.2 Functional classification (discriminant analysis)

Functional discriminant analysis is a natural extension of linear discriminant analysis (LDA) for the classification of functional data and is well-studied in the literature (e.g., Delaigle & Hall, 2011; Ferrando, Ventura-Campos & Epifanio, 2020; Kraus & Stefanucci, 2018; Lila, Zhang, & Rane, 2021; Park, Ahn & Jeon 2021; Ramsay & Silverman, 2010). This is a challenging problem, given that, in the functional setting, each observation has an infinite number of covariates. This property can lead to overfitting. It is therefore not surprising that many methods in the FDA literature adopt a regularized approach. In this thesis, we propose a regularized discriminant analysis method for multivariate functional data. We leverage a connection to the optimal scoring objective function, and re-frame the problem under a penalized regression framework.

The thesis is organized as follows. To introduce the problem, we first review the rationale for a very popular discriminant approach in the FDA literature, i.e., the linear centroid-based classifier, and the derivation of its objective function. In Section 2 we introduce the model proposed for binary classification of univariate functional data, paving the way for the multivariate classifier. In Section 3 we introduce the multivariate classifier which is able to accommodate multiple functional predictors supported on a common domain, as required by our motivating application. In Section 4 we present an extension of the method that accommodates multi-class outcomes, whose full development is left to future work. We conduct a simulation study of univariate classification of binary outcome in Section 5. We then apply the multivariate classifier to a real-world diffusion magnetic resonance imaging (MRI) dataset in Section 6. We leave detailed derivations

from the Introduction to Appendix A, and we derive the convergence rate of the univariate classifier's excess prediction risk in Appendix B. Additional simulation study results are found in Appendix C.

1.3 Setup and Notation

We are interested in studying functional classification, where the outcome is a categorical variable and the predictors are curves supported on some one-dimensional compact domain $T \subset \mathbb{R}$. Some applications include classifying recorded speech periodograms into different phoneme classes, and predicting precipitation level based on annual temperature patterns. To preserve the inherent structure in the data, we model each predictor as a function mapping the domain to the real line, $x : T \to \mathbb{R}$. In practice the data $\{x(t_1), ..., x(t_m)\}$ is discretized as densely sampled datapoints on some grid $\{t_1, ..., t_m\} \subset T$. In the following sections, we assume $T = [0, 1]$ without loss of generality.

Assume our training data $\{(g_i, x_i)\}$ consists of n independent samples of (G, X). G is a binary random variable that denotes the two possible classes of the outcome:

$$P(G = g_1) = \pi_1, \quad P(G = g_0) = \pi_0.$$

We have n_1 samples from class 1 and n_0 samples from class 0. Without prior information we usually assume $\pi_0 = \pi_1$.

Denote with $\mathcal{L}^2(T)$ the space of square-integrable functions over T, equipped with the standard inner product

$$\langle f, g \rangle = \int f(t)g(t)\, dt$$

and norm

$$\|f\|^2 = \int f(t)^2\, dt.$$

X is a zero-mean, square-integrable random function taking values in $\mathcal{L}^2(T)$. Let

$$\mu_1(t) = E[X(t)|G = g_1], \quad \mu_0(t) = E[X(t)|G = g_0]$$

be the conditional mean functions of X, and assume $\mu_1 \neq \mu_0$. The two classes are assumed to share a common within-class covariance function:

$$C(s, t) = E[X(t)X(s)|\, G = g_1] = E[X(t)X(s)|\, G = g_0], \quad \forall s, t \in T$$

and we assume the covariance function is square-integrable, i.e.

$$\int \int C(s, t)^2\, dsdt < \infty.$$

Since the covariance function is real, symmetric, square-integrable, and non-negative, we can define the covariance operator $L_C : \mathcal{L}^2(T) \to \mathcal{L}^2(T)$ as

$$L_C f(\cdot) = \int C(t,\cdot) f(t)\, dt, \quad \forall f \in \mathcal{L}^2(T).$$

Furthermore, L_C is a compact, self-adjoint operator (Cucker & Smale, 2001). It admits a spectral representation, for eigenvalues $\theta_1 \geq \theta_2 \geq ... \geq 0$ and associated eigenfunctions $e_1, e_2, ... \subset \mathcal{L}^2(T)$ of L_C:

$$C(s,t) = \sum_{k=1}^{\infty} \theta_k e_k(s) e_k(t),$$

$$L_C\, e_k(\cdot) = \int C(t,\cdot) e_k(t)\, dt = \theta_k e_k(\cdot).$$

Moreover, $\forall f \in \mathcal{L}^2(T)$,

$$\begin{aligned} L_C f(\cdot) &= \int C(t,\cdot) f(t)\, dt \\ &= \int \sum_{k=1}^{\infty} \theta_k e_k(\cdot) e_k(t) f(t) dt \\ &= \sum_{k=1}^{\infty} \theta_k e_k(\cdot) \int e_k(t) f(t) dt \\ &= \sum_{k=1}^{\infty} \theta_k \langle e_k, f \rangle e_k(\cdot). \end{aligned}$$

Since L_C is non-negative, the inverse operator is well-defined, although in general unbounded (more on this later), with the spectral representation:

$$L_C^{-1} f(\cdot) = \int C^{-1}(t,\cdot)\, f(t)\, dt,$$

where

$$C^{-1}(s,t) = \sum_{k=1}^{\infty} \frac{1}{\theta_k} e_k(s) e_k(t).$$

1.4 Rationale for linear centroid classifier

Delaigle & Hall (2011) showed that perfect asymptotic classification is possible in functional classification problems. This remarkable phenomenon is a special property of the infinite-dimensional setting of functional data, and cannot occur in the multivariate setting except in pathological cases. They further showed that simple linear centroid classifiers, based on a well-chosen one-dimensional projection of the predictor functions, can achieve this optimal classification either exactly or in the limit. We briefly review their method with supplementing details from Kraus & Stefanucci (2018). For the sake of consistency

we follow their assumptions (in this section only), and note that some of our assumptions are different in later sections.

Assume that the functional data is Gaussian. Assume that the covariance $C(s,t)$ is strictly positive definite and uniformly bounded. Strict positive-definiteness of C is equivalent to $\theta_k > 0, \forall k$. Uniform boundedness implies that $\sum_k \theta_k < \infty$.

Let

$$\hat{\mu}_1(t) = \frac{1}{n_1} \sum_{\{i\,|G_i=1\}}^{n_1} x_i(t), \quad \hat{\mu}_0(t) = \frac{1}{n_0} \sum_{\{i\,|G_i=0\}}^{n_0} x_i(t)$$

be the sample mean of each class.

Let X^* be a new data function that we wish to classify to one of the two populations. The centroid classifier assigns X^* to class 1 if statistic

$$T(X^*) = D^2(X^*, \hat{\mu}_0) - D^2(X^*, \hat{\mu}_1) > 0$$

for some distance function D.

Delaigle & Hall showed that when D is given by the distance of the projections of the two functions on a one-dimensional space, the classifier enjoys optimality properties. Let $\beta(t) \in \mathcal{L}^2(T)$. Consider

$$D(X^*, \hat{\mu}_1) = \langle X^* - \hat{\mu}_1, \beta \rangle.$$

As n_1 and n_0 increase, the sample statistic, obtained by replacing D with the projection defined above,

$$\hat{T}(X^*) = \langle X^* - \hat{\mu}_0, \beta \rangle^2 - \langle X^* - \hat{\mu}_1, \beta \rangle^2$$

converges to the theoretical statistic

$$T(X^*) = \langle X^* - \mu_0, \beta \rangle^2 - \langle X^* - \mu_1, \beta \rangle^2.$$

This approach amounts to projecting the data onto a one-dimensional space determined by β (to be learned from the data), and then perform classification based on the projections.

1.5 Objective function

Now we will look at the misclassification error and objective function in this Gaussian setting. Let η denote the projection $\langle X, \beta \rangle$. If X belongs to class $j, (j = 0, 1)$, the distribution of η_j is Normal with mean $\langle \mu_j, \beta \rangle$ and variance $\langle \beta, L_C \beta \rangle$ (Kraus & Stefanucci, 2018). By definition every projection of Gaussian random function is Gaussian and

$$\eta_j = \langle X_j, \beta \rangle \sim N\Big(\langle \mu_j, \beta \rangle,\ \langle \beta, L_C \beta \rangle \Big),$$

where $N\Big(\langle \mu_j, \beta \rangle,\ \langle \beta, L_C \beta \rangle \Big)$ denotes a normal distribution centered at $\langle \mu_j, \beta \rangle$, and with variance $\langle \beta, L_C \beta \rangle$.

Denote the corresponding Gaussian densities by $f_{\beta,j}$. The classifier is given by:

$$\boxed{f(X) = 1\Big\{\frac{f_{\beta,1}(\eta)}{f_{\beta,0}(\eta)} > 1\Big\} = 1\Big\{\langle X-\mu_0,\beta\rangle^2 - \langle X-\mu_1,\beta\rangle^2 > 0\Big\}}$$

We want to re-write the expression towards deriving the objective function.

Lemma 1.1 For any given data function $X \in \mathcal{L}^2(T)$, the linear classifier calculates the distance between the data to the two class means in the projection space. Then X is assigned to the class with the shorter distance. The classifier can be re-written in terms of the class means' average and class means' difference as

$$f(X) = 1\Big\{\langle X-\mu_0,\beta\rangle^2 - \langle X-\mu_1,\beta\rangle^2 > 0\Big\} = 1\Big\{\langle\delta,\beta\rangle\langle X-\bar{\mu},\beta\rangle > 0\Big\},$$

where $\delta = \mu_1 - \mu_0$ and $\bar{\mu} = (\mu_0+\mu_1)/2$. The derivation is found in Appendix A1.

Lemma 1.2 Using the classifier

$$f(X) = 1\Big\{\langle\delta,\beta\rangle\langle X-\bar{\mu},\beta\rangle > 0\Big\},$$

and under the assumption that $\pi_0 = \pi_1 = 1/2$, the misclassification probability is

$$\boxed{err = \frac{P_0[f(X)=1]}{2} + \frac{P_1[f(X)=0]}{2} = 1 - \Phi\left(\frac{|\langle\mu_1-\mu_0,\beta\rangle|}{2\langle\beta,L_C\beta\rangle^{1/2}}\right)},$$

where P_0 and P_1 denote the probability distribution of η_0 and η_1, respectively, and Φ denotes the cumulative distribution function of a standard Normal random variable.

The derivation is found in Appendix A2.

Therefore, to find the best β for the linear centroid classifier, the objective function is given by minimizing the misclassification error:

$$\underset{\beta}{\text{minimize}} \quad 1 - \Phi\left(\frac{|\langle\mu_1-\mu_0,\beta\rangle|}{2\langle\beta,L_C\beta\rangle^{1/2}}\right).$$

By the property of the cumulative density function, minimizing the misclassification error above is equivalent to the following problem

$$\underset{\beta}{\text{maximize}}\, \frac{|\langle\mu_1-\mu_0,\beta\rangle|}{|\langle\beta,L_C\beta\rangle^{1/2}|}. \tag{1}$$

The solution to this problem is derived in the following lemma.

Lemma 1.3 The function β_0 that solves the maximization problem (1) is

$$\beta_0 = \underset{\beta}{\text{maximize}}\, \frac{|\langle\mu_1-\mu_0,\beta\rangle|}{|\langle\beta,L_C\beta\rangle^{1/2}|} = L_C^{-1}(\mu_1-\mu_0).$$

The derivation is found in Appendix A3.

1.6 Assumptions on convergence

The convergence or divergence of $L_C^{-1}(\mu_1 - \mu_0)$ and $L_C^{-1/2}(\mu_1 - \mu_0)$ are key to functional classification. They determine whether or not we can achieve the perfect classification described by Delaigle & Hall. Here we summarize three regimes:

Case 1: $\|L_C^{-1/2}(\mu_1 - \mu_0)\| < \infty$,

Case 2: $\|L_C^{-1/2}(\mu_1 - \mu_0)\| < \infty$ and $\|L_C^{-1}(\mu_1 - \mu_0)\| = \infty$,

Case 3: $\|L_C^{-1/2}(\mu_1 - \mu_0)\| = \infty$.

Note that the quantities involved here are population parameters that are generally unknown in a practical statistical analysis. So this is more of a theoretical issue. In our setting, mostly to be able to study the convergence rates of the model, we assume $\|L_C^{-1}(\mu_1 - \mu_0)\|$ and $\|L_C^{-1/2}(\mu_1 - \mu_0)\|$ are finite.

As shown in Lemma 1.3, the function that minimizes the misclassification error is $\beta_0 = L_C^{-1}(\mu_1 - \mu_0)$. For this choice of β_0, or any multiple of it, the misclassification probability is given by the following lemma.

Lemma 1.4 Given β_0 in (1), the misclassification probability is given by

$$err_0 = 1 - \Phi\left(\frac{\|L_C^{-1/2}(\mu_1 - \mu_0)\|}{2}\right).$$

The derivation is found in Appendix A4.

The minimum classification error that can be achieved varies depending on the assumptions on the convergence of $\|L_C^{-1}(\mu_1 - \mu_0)\|$ and $\|L_C^{-1/2}(\mu_1 - \mu_0)\|$. Recall the spectral representation

$$C(s,t) = \sum_{k=1}^{\infty} \theta_k e_k(s) e_k(t),$$

and let

$$\mu_1(t) - \mu_0(t) = \sum_{k}^{\infty} d_k e_k(t)$$

be the generalized Fourier decomposition of $\mu_1 - \mu_0$ with respect to eigenfunctions $e_1, e_2, ... \subset \mathcal{L}^2(T)$.

Note that in our derivation for Lemma 1.3, where we found the function to minimize the misclassification error, we first assumed that $\|L_C^{-1/2}(\mu_1 - \mu_0)\| < \infty$, then we assumed $\|L_C^{-1}(\mu_1 - \mu_0)\| < \infty$. The latter is more stringent since

$$\|L_C^{-1/2}(\mu_1 - \mu_0)\|_{\mathcal{L}^2}^2 = \sum_k \frac{d_k^2}{\theta_k},$$

$$\|L_C^{-1}(\mu_1 - \mu_0)\|_{\mathcal{L}^2}^2 = \sum_k \frac{d_k^2}{\theta_k^2},$$

and $1/\theta_k$ and $1/\theta_k^2$ are both diverging, but $1/\theta_k^2$ is diverging faster. Therefore d_k needs to decay faster to ensure the sum doesn't explode, so it's a stronger assumption.

Here we rephrase Theorem 1 in (Delaigle & Hall, 2011) as follows:

Case 1

If $\|L_C^{-1/2}(\mu_1 - \mu_0)\| < \infty$, then

$$err_0 = 1 - \Phi\left(\frac{\|L_C^{-1/2}(\mu_1 - \mu_0)\|}{2}\right)$$

is the minimal (Bayes) error (Berrendero et al., 2018).

Case 2

If $\|L_C^{-1/2}(\mu_1 - \mu_0)\| < \infty$ but $\|L_C^{-1}(\mu_1 - \mu_0)\| = \infty$, the Bayes risk cannot be achieved by a projection classifier based on a bounded linear functional of the form $\langle X, \beta\rangle$ for some $\beta \in \mathcal{L}^(T)$. However, approximations in the form of projections can asymptotically achieve the Bayes risk (Kraus & Stefanucci, 2018).

Case 3

If $\|L_C^{-1/2}(\mu_1 - \mu_0)\| = \infty$ then $err_0 = 0$ and perfect classification is possible.

1.7 Linear problem

Lemma 1.3 was derived using Cauchy-Schwartz inequality and checking that the solution achieved the upper bound. Alternatively, the objective function

$$\underset{\beta}{\text{maximize}}\, \frac{|\langle \mu_1 - \mu_0, \beta\rangle|}{|\langle \beta, L_C\beta\rangle^{1/2}|}$$

can be solved by maximizing $\langle \mu_1 - \mu_0, \beta\rangle$ subject to $\langle \beta, L_C\beta\rangle = 1$. Using Lagrange multipliers:

$$L(\beta, \lambda) = \langle \mu_1 - \mu_0, \beta\rangle + \lambda(1 - \langle \beta, L_C\beta\rangle),$$

taking the Frechet derivative with respect to β and setting it to zero, we obtain the equation

$$2\lambda L_C\beta = \mu_1 - \mu_0$$

(Kraus & Stefanucci, 2018). Note that λ only changes the scale, which can be normalized. Solutions of this equation, if they exist, i.e., if $\|L_C^{-1}(\mu_1 - \mu_0)\| < \infty$, yield the same optimal misclassification probability $\forall \lambda > 0$. Without loss of generality, we take $\lambda = 1/2$. Thus we translate the unconstrained quadratic optimization problem

$$\boxed{\underset{\beta}{\text{maximize}}\ \langle \mu_1 - \mu_0, \beta\rangle - \frac{1}{2}\langle \beta, L_C\beta\rangle} \tag{2}$$

into the linear problem seen in some literature:

$$L_C\beta = \mu_1 - \mu_0.$$

1.8 Final Remarks and introduction to our method

This motivates us to find a classification rule based on a linear one-dimensional projection of the data. We want to estimate β such that the projection yields good class separation. Equation (2) is particularly well suited to defining regularized versions of β_0 by replacing the population quantities with the empirical counterparts and adding a penalty term. Two recent papers used this kind of approach. Park, Ahn, Jeon (2021) combined the L1 norm of $\beta(t)$ and L2 norm of $\beta'(t)$ to achieve regularization and sparsity (in the form of domain selection). Kraus & Stefanucci (2018) extended the multivariate conjugate gradient algorithm to find the numerical solution.

Most relevant to this thesis are the following works. Hastie, Buja, Tibshirani (1994) recast the high-dimensional LDA problem into a penalized regression framework via optimal scoring (OS), which offers a convenient objective function and estimation procedure. Gaynanova (2020) extended the OS approach to multi-class categorical problems and derived convergence results. Notably, it can be shown that the discriminant vector obtained via OS is the same as the LDA estimator up to a constant. In addition, OS facilitates the use of any penalized regression technique. In light of this connection, we recast the functional discriminant analysis into a functional regression problem under a RKHS regression framework (Yuan, Cai 2010).

2 Univariate functional data

The population quantity we are interested in is $\beta^0 \in \mathcal{L}^2(T)$ such that

$$\boxed{L_C\beta^0 = (\mu_1 - \mu_0)} \tag{3}$$

where we assume that $\|L_C^{-1/2}\mu_1 - \mu_0\| < \infty$ and $\|L_C^{-1}\mu_1 - \mu_0\| < \infty$. The goal of our classification model is to derive an estimate $\hat{\beta}$ of β^0 using the training data, and project new data onto the estimated direction $\hat{\beta}$ to perform classification. In practice, since the population quantities C, μ_1, μ_0 are unknown, we replace them with the sample counterparts.

The sample covariance is given by

$$\hat{C}(s,t) = \frac{1}{n}\sum_{i=1}^{n} x_i(s)x_i(t), \quad s,t \in T$$

and the sample conditional mean function is the average from their respective classes.

$$\hat{\mu}_1(t) = \frac{1}{n_1}\sum_{\{i\,|G_i=1\}}^{n_1} x_i(t), \quad \hat{\mu}_0(t) = \frac{1}{n}\sum_{\{i\,|G_i=0\}}^{n_0} x_i(t).$$

We then rewrite Equation (3) as a penalized minimization problem, where we replace the conditional means and the covariance with the sample conditional means and the

sample covariance, respectively. The penalized classification objective function is given by

$$\boxed{\underset{\beta}{\text{minimize}} \ \frac{1}{2}\langle \beta, L_{\hat{C}}\beta\rangle - \langle \hat{\mu}_1 - \hat{\mu}_0, \beta\rangle + \text{Pen}(\beta)} \tag{4}$$

The penalty term $\text{Pen}(\beta)$ ensures the infinite minimization problem is well-defined and encourages the estimated function to be smooth.

One drawback of this approach is that it requires the explicit computation of $L_{\hat{C}}$, which can be prohibitive for functional data that are very densely observed. In the following we try to overcome such an issues.

2.1 Rationale for penalized regression reformulation

In the functional setting, Delaigle and Hall (2011) showed that, under appropriate assumptions,

$$\arg\min_{\beta} \left\{ E\left[H - E[H] - \int \beta(t)X(t)\,dt \right]^2 \right\},$$

where $H = 1_{\{G=g_1\}}$, is, up to constants, equivalent to minimizing eq (2), i.e.,

$$\underset{\beta}{\text{minimize}} \ \frac{1}{2}\langle \beta, L_C\beta\rangle - \langle \mu_1 - \mu_0, \beta\rangle.$$

For class 1, $H - E[H] = 1 - \pi_1 = \pi_0$. For class 0, $H - E[H] = 0 - \pi_1 = -\pi_1$. The expectation can be replaced with its sample version, i.e. $h_i = n_0/n$ if $G = g_1$ and $h_i = -n_1/n$ if $G = g_0$. Thus we obtain a model

$$\arg\min_{\beta} \left\{ n^{-1} \sum \left[h_i - \int \beta(t)x_i(t)\,dt \right]^2 \right\}.$$

This shows that the linear discriminant analysis problem can be effectively recast into a functional regression problem.

The idea can be generalized by leveraging an optimal scoring (OS) reformulation. Instead of directly solving the classification problem, we leverage the connection between LDA and OS to convert it into a regression problem.

Penalized optimal scoring introduces an auxiliary variable $\theta \in \mathbb{R}^2$ to transform the labels. For some hypothesis space $\mathcal{H}$ that we will formally define later, the objective function is defined by

$$\underset{\theta\in\mathbb{R}^2,\, \beta\in\mathcal{H}}{\text{minimize}} \ \ n^{-1}\sum_{i=1}^{n} \left[y_i^T\theta - \int x_i(t)\beta(t) \right]^2 + \lambda\,\text{Pen}(\beta),$$

under the constraint

$$\theta^T Y^T Y\theta/n = 1, \quad \theta^T Y^T \mathbf{1} = 0,$$

where $y_i \in \mathbb{R}^2$ is an indicator vector, and $Y \in \mathbb{R}^{n\times 2}$ is given by vertically concatenating y_i^T, i.e., this is an indicator matrix. $\theta \in \mathbb{R}^2$ is the auxiliary variable, $\mathbf{1} \in \mathbb{R}^2$ is a vector of 1's, and $\text{Pen}(\beta)$ is a regularization term to smooth out the β estimate.

Hastie, Buja, Tibshirani (1995) showed that the LDA and OS solutions are proportional: $\beta_{LDA} \propto \beta_{OS}$.

While this reformulation achieves the same objective as that in equation Eq. (4), it gives us a more flexible tool that can be straightforwardly generalized to the multi-class setting, as we will show in Section 4.

Since the classifier is invariant to changes in scale of the discriminant function β, finding the optimal direction that separates the two classes reduces to solving a penalized regression problem.

2.2 Regularized estimation with second derivative penalty

Given $x_i : T \to \mathbb{R}$, the mean-centered predictor function for the i^{th} sample, and g_i, the label for the i^{th} sample, OS introduces an auxiliary variable $\theta \in \mathbb{R}^{2\times 1}$. The labels are stored in an indicator matrix Y of size $n \times 2$: $Y_{i,1} = 1$ if sample $G_i = g_1$ and $Y_{i,2} = 1$ if sample $G_i = g_0$. For example,

$$Y = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \dots & \end{pmatrix}$$

The vector $Y\theta \in \mathbb{R}^n$ represents the transformed training labels.

Denote $\mathcal{H} = \mathcal{W}^2(T) \in \mathcal{L}^2(T)$ the Sobolev space with first and second distributional derivatives in $\mathcal{L}^2(T)$. It is an RKHS with an implicit symmetric and positive definite reproducing kernel. The space is equipped with the norm

$$J(f) = \left(\|f''\|_{\mathcal{L}^2}^2 + \epsilon \|f\|_{\mathcal{L}^2} \right)^{1/2} \quad \forall f \in \mathcal{W}^2(T).$$

For $\epsilon = 0$, which is our second-derivative penalty, it defines a semi-norm as opposed to a norm. In this case, we restrict ourselves to the subspace of $\mathcal{L}^2(T)$ that is orthogonal to the null space of J.

Note that in the next section, we present an alternative approach to defining smooth estimates that makes use of an explicit kernel (the exponential kernel) to define a different hypothesis RKHS space, with an implicit penalty function. For functional estimates supported on Euclidean spaces , defining a reproducing kernel explicitly is straightforward. For general non-Euclidean domains where defining a reproducing kernel is difficult (sometimes impossible), the differential penalty approach is preferable (Lila et al., 2021).

We then define the following objective function, given in the form of **penalized optimal scoring** for functional data:

$$\boxed{\underset{\theta\in\mathbb{R}^2,\ \beta\in\mathcal{H}}{\arg\min} \sum_{i=1}^{n} \left[y_i^T\theta - \int_T x_i(t)\beta(t)\,dt \right]^2 + \lambda \int_T (\beta''(t))^2\,dt}$$

under the constraints

$$\theta^T Y^T Y\theta = n, \quad \theta^T Y^T Y\mathbf{1} = 0.$$

Denote the vector of functions (mean-normalized) as $\mathbf{X}(t) = (x_1(t), \ldots, x_n(t))^T$. We approach the minimization problem in a two-step fashion.

Step 1. Estimation of θ given β.

Given β, the minimizing θ of the objective function is given by

$$\hat{\theta} = \underset{\theta\in\mathbb{R}^2}{\arg\min} \left\| Y\theta - \int_T \mathbf{X}(t)\beta(t)\,dt \right\|^2 + \lambda \int (\beta''(t))^2 dt,$$

with the two constraints on θ stated previously. In the following lemma, we will show that the optimal θ does not depend on β — that is — we don't need to iteratively solve the joint minimization problem.

Lemma 2.1

The minimizing $\hat{\theta}$ is given by

$$\hat{\theta} = \left(\sqrt{\frac{n_0}{n_1}}, -\sqrt{\frac{n_1}{n_0}} \right)$$

Proof: Using Karush–Kuhn–Tucker (KKT), or the method of lagrange multipliers,

$$L(\theta, \lambda, \mu) = \left\| Y\theta - \int_T \mathbf{X}(t)\beta(t)\,dt \right\|^2 + \int (\beta''(t))^2 - \lambda(\|Y\theta\| - n) - \mu\theta^T Y^T Y 1,$$

with some slight abuse notation since we use λ and μ here to be the standard lagrange multiplier notation. This is distinct from the regularization parameter.

We want to solve the system of equations:

$$\begin{cases} \frac{\delta L}{\delta\theta} = 0, \\ \|Y\theta\|^2 = n, \\ \theta^T Y^T Y\mathbf{1} = 0 \end{cases} .$$

The indicator matrix Y leads to an intuitive expression for $(Y^T Y)$ and its inverse, i.e.,

$$(Y^T Y) = \begin{pmatrix} n_1 & 0 \\ 0 & n_0 \end{pmatrix},$$

$$(Y^T Y)^{-1} = \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_0 \end{pmatrix}.$$

The first constraint can be rewritten as:

$$\begin{aligned}
\theta^T Y^T Y \theta &= n \\
\begin{pmatrix} \theta_1 & \theta_2 \end{pmatrix} \begin{pmatrix} n_1 & 0 \\ 0 & n_0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} &= n \\
\begin{pmatrix} n_1\theta_1 & n_0\theta_2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} &= n \\
n_1\theta_1^2 + n_0\theta_2^2 &= n
\end{aligned}$$

The solutions of θ_1 and θ_2 lie on an ellipse.
The second constraint can be rewritten as:

$$\begin{aligned}
\theta^T Y^T Y \mathbf{1} &= 0 \\
\begin{pmatrix} \theta_1 & \theta_2 \end{pmatrix} \begin{pmatrix} n_1 & 0 \\ 0 & n_0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= 0 \\
\begin{pmatrix} n_1\theta_1 & n_0\theta_2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= 0 \\
n_1\theta_1 + n_0\theta_2 &= 0
\end{aligned}$$

The solutions of θ_1 and θ_2 lie on a line.

The line can only intersect the ellipse at most two times, so we can have at most two solutions. Since the ellipse is symmetrical, the two points are reflections across the $y = x$ line, so we just need to consider one solution. Hence we have a special case that the constraints are so strong we get the solutions immediately without using the objective function in lagrange multiplier. From constraint one we get:

$$\begin{aligned}
n_1\theta_1^2 + n_0\theta_2^2 &= n \\
n_0\theta_2^2 &= n - n_1\theta_1^2 \\
\theta_2^2 &= \frac{n - n_1\theta_1^2}{n_0} \\
\theta_2 &= -\sqrt{\frac{n - n_1\theta_1^2}{n_0}}
\end{aligned}$$

From constraint two we get:

$$\begin{aligned}
n_1\theta_1 + n_0\theta_2 &= 0 \\
n_0\theta_2 &= -n_1\theta_1 \\
\theta_2 &= -\frac{n_1\theta_1}{n_0}
\end{aligned}$$

Equating these two expressions we have that

$$
\begin{aligned}
-\frac{n_1\theta_1}{n_0} &= -\sqrt{\frac{n-n_1\theta_1^2}{n_0}} \\
\frac{n_1^2}{n_0^2}\,\theta_1^2 &= \frac{n-n_1\theta_1^2}{n_0} \\
\frac{n_1^2}{n_0^2}\,\theta_1^2 &= \frac{n}{n_0} - \frac{n_1}{n_0}\theta_1^2 \\
\left(\frac{n_1^2}{n_0^2} + \frac{n_1}{n_0}\right)\theta_1^2 &= \frac{n}{n_0} \\
\frac{n_1^2 + n_1 n_0}{n_0^2}\,\theta_1^2 &= \frac{n}{n_0} \\
\theta_1^2 &= \frac{n\,n_0}{n_1^2 + n_1\,n_0} \\
\theta_1 &= \sqrt{\frac{n_0}{n_1}}
\end{aligned}
$$

Plug in to find θ_1:

$$\theta_2 = -\frac{n_1\theta_1}{n_0} = -\sqrt{\frac{n_1}{n_0}}.$$

Therefore we have

$$\hat{\theta} = \left(\sqrt{\frac{n_0}{n_1}},\, -\sqrt{\frac{n_1}{n_0}}\right),$$

which is consistent with the results obtained in (Gaynanova, 2020).

Step 2. Estimation of β given θ.

Note that we obtained θ without β. We can denote the transformed labels $Y\theta$ with $\tilde{y} \in \mathbb{R}^n$: $\tilde{y}_i = \sqrt{n_0/n_1}$ if $G_i = g_1$ and $\tilde{y}_i = -\sqrt{n_1/n_0}$ if $G_i = g_0$. Therefore, the optimal β is given by minimising

$$\hat{\beta} = \underset{\beta\in\mathcal{H}}{\arg\min}\;\; \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{y}_i - \int_T x_i(t)\beta(t)\,dt\right)^2 + \lambda\int(\beta''(t))^2 dt. \tag{5}$$

The discrete version of the objective function becomes (Yuan & Cai 2010):

$$\boxed{\underset{d\in\mathbb{R}^2,\, c\in\mathbb{R}^n}{\text{minimize}}\;\; \frac{1}{n}\|\tilde{y} - (T\mathbf{d} + \Sigma\mathbf{c})\|_{l_2}^2 + \lambda\mathbf{c}^T\Sigma\mathbf{c},}$$

where the matrices are defined as

$$T_{ij} = \int x_i(t)t^{j-1}\,dt \in \mathbb{R}^{n\times 2}$$

$$\Sigma_{ij} = \int\int x_i(s)K(s,t)x_j(t)\,ds\,dt \;\in \mathbb{R}^{n\times n}$$

For a given θ, the minimizing β of the objective function is given by (Yuan & Cai 2010):

$$\boxed{\hat{\beta} = d_1\xi_1(t) + d_2\xi_2(t) + \sum_i^n c_i \int K(t,\cdot)x_i(t)\,dt.}$$

The coefficients $\mathbf{c}, \mathbf{d}$ are given by

$$\mathbf{d} = (d_1, d_2)^T = (T'W^{-1}T)^{-1}T'W^{-1}\tilde{y},$$

$$\mathbf{c} = (c_1, ..., c_n)^T = W^{-1}\left[I - T(T'W^{-1}T)^{-1}T'W^{-1}\right]\tilde{y},$$

and

$$W = \Sigma + n\lambda I \in \mathbb{R}^{n\times n}.$$

The basis functions of the null space are given by $\xi_1 = 1, \xi_2 = t$ since they are not regularized by the second-derivative penalty, where the null space is defined as

$$\mathcal{H}_0 = \{\beta \in \mathcal{H} : \text{Pen}(\beta) = 0\}$$

and $K(s,t)$ is the reproducing kernel for $\mathcal{H}_1$, the complement of the null space.

2.3 Theory

In this section, we aim to provide theoretical guarantees for the univariate model in Section 2.2, where the unknown β_0 defined in equation (3) is a function belonging to a Sobolev space. We provide the probability bound for the out-of-sample risk, i.e. the random variable

$$E^*\left[\langle X^*, \beta_0 - \hat{\beta}\rangle\right]^2,$$

where X^* is a test sample, independent of the training data, and the expectation is taken over X^*. The excess prediction risk is a measure of how close the prediction made with estimated parameter is to the prediction made with the optimal unknown β_0.

Assume that $\epsilon > 0$. By the Sobolev embedding theorem (Brezis 2011), there exists $M \geq 0$ such that for any $t \in T$,

$$f(t) \leq \sup_t |f(t)| \leq M\Big(\|f''\|^2_{\mathcal{L}^2} + \epsilon\|f\|^2_{\mathcal{L}^2}\Big),\ \forall f \in \mathcal{W}^2(T)$$

That is, the evaluation operator is a continuous functional. A consequence is that the space $\mathcal{W}^2(T)$ equipped with the norm $J^{1/2}(\cdot) = (\|f''\|^2_{\mathcal{L}^2} + \epsilon\|f\|^2_{\mathcal{L}^2})^{1/2}$ is an RKHS. Denote $\mathcal{W}^2(T) = \mathcal{H}$. It has a symmetric, positive definite kernel function $K : T \times T \to \mathbb{R}$, with eigenvalue and eigenfunction pairs $\{\zeta_k, \psi_k\}_{k\in N}$.

$$L_K f(.) = \int K(s,.) f(s) ds$$

$$K(s,t) = \sum_k \zeta_k \psi_k(s) \psi_k(t)$$

The square root operator of L_K is defined by

$$L_K^{1/2}(\psi_k) = L_{K^{1/2}}(\psi_k) = \sqrt{\zeta_k}\psi_k$$

$$K^{1/2}(s,t) = \sum_k \sqrt{\zeta_i}\psi_k(s)\psi_k(t)$$

For $\epsilon = 0$, $J^{1/2}$ defines a semi-norm instead of a norm.

Following the notation of (Cai & Yuan 2012), we define the sandwich operator

$$T = L_{K^{1/2}CK^{1/2}} = L_K^{1/2} L_C L_K^{1/2}$$

If both $L_K^{1/2}$ and L_C are bounded linear operators, so is $T = L_{K^{1/2}CK^{1/2}}$. There exists a sequence of positive eigenvalues $\{\tau_k\}$ and corresponding eigenfunctions $\{\eta_k\}$ such that

$$K^{1/2}CK^{1/2}(s,t) = \sum_k \tau_k \eta_k(s)\eta_k(t)$$

We make the following assumptions.

Assumption 1. The norm of X is a.s. bounded. There exists $M_2 > 0$ such that

$$\|X\|_{L^2} \leq M_2 \ a.s.$$

Moreover, we denote

$$\kappa = M_2 \|L_K^{1/2}\|_{op}$$

(Lila et. al, 2021).

Assumption 2. There exists smooth $\beta_0 \in \mathcal{W}^2(T)$ such that

$$\beta_0 = L_C^{-1}(\mu_1 - \mu_0),$$

i.e., the optimal function is well-defined and can be found in this space. This implies that we are in the regime define by Case 1 in Section 1.6.

Assumption 3. The effective dimension of the sandwich operator T satisfies

$$D(\lambda) = Tr((T + \lambda I)^{-1} T) = \sum_k^{\infty} \frac{\tau_k}{\tau_k + \lambda} \leq c\lambda^{-\theta}$$

for some $c, \theta > 0$. Tr denotes the trace operator.

Note that

$$Tr(T) = \sum_{k}^{\infty} \tau_k$$

Assumption 1 allows us to use Theorem B.2. in the auxiliary results (Tong & Ng 2018) in the derivation of the probability bound. It does not have practical implications. Assumption 2 is needed so that the population quantity β_0 is well defined and belongs to the space of smooth functions $\mathcal{W}^2(T)$. Assumption 3 is related to the rate of decay of the eigenvalues of L_K, L_C, and their alignment (Cai & Yuan 2012). This assumption holds by assuming that $\tau_k \asymp k^{-2r}$ with $r > 1/2$ (Lila et al. 2021).

The following theorem provides a probability bound for the out-of-sample prediction risk.

Theorem 2.1 Under Assumptions 1 to 3, if $\lambda \asymp n^{-\frac{1}{1+\theta}}$, the estimate $\hat{\beta}$ in (5) satisfies

$$\boxed{E^*\left[\langle X^*, \beta_0 - \hat{\beta}\rangle_{\mathcal{L}^2(T)}\right]^2 = \mathcal{O}_p(n^{-\frac{1}{1+\theta}})}$$

Intuitively, this means that by increasing the number of observations n the predictions made with $\hat{\beta}$ get closer and closer to those made by using β. The speed of convergence is determined by the parameter θ, which determined how difficult the problem at hand is.

Similar rates of convergence for excess risk have been shown to hold in the regularized functional linear regression setting (Tong & Ng, 2018). However, unlike the regression setting, the random variable $\tilde{Y} - \langle X, \beta_0\rangle_{\mathcal{L}^2(T)}$ and the predictor X is not independent, due to the optimal scoring procedure. Therefore we cannot directly apply their results. Despite this dependence, we are still able to recover the functional regression rates of convergence. The proof is found in Appendix B.

2.4 Regularized estimation with reproducing kernel

In this section, we consider a modification of the model proposed in Section 2.2, where we define the unknown β to be a function belonging to an RKHS with an exponential kernel. Given that we have shown that the optimal θ does not depend on β, this modification will only require an adaptation of Step 2, estimation of β given θ.

Consider the exponential kernel

$$K(s,t) = e^{-((s-t)/\sigma)^2/2}, \ \ \forall s,t \in T, \sigma \in \mathbb{R}^+$$

and its associated RKHS $\mathcal{H}$.

Compared to the second derivative penalty induced kernel, the exponential kernel has an additional tuning parameter, the bandwidth, which controls how strongly neighboring observations influence each other.

The null space is empty for the exponential kernel, so we do not need to include the basis functions $\xi(t)$ of the null space in the estimate $\hat{\beta}$ like in Section 2.2. The discrete objective function reduces from

$$\underset{d\in\mathbb{R}^2,\, c\in\mathbb{R}^n}{\text{minimize}} \;\; \frac{1}{n}\|\tilde{y} - (T\mathbf{d} + \Sigma\mathbf{c})\|_{l_2}^2 + \lambda \mathbf{c}^T\Sigma\mathbf{c}$$

to

$$\boxed{\underset{\mathbf{c}\in\mathbb{R}^n}{\text{minimize}} \frac{1}{n} \;\; \|\tilde{y} - \Sigma\mathbf{c}\|_{l_2}^2 + \lambda \mathbf{c}^T\Sigma\mathbf{c}.} \tag{6}$$

Lemma 2.2 The estimate is given by

$$\boxed{\hat{\beta} = \sum_i^n c_i \int K(t,\cdot)x_i(t)\,dt}$$

where the coefficient vector is given by the solution to Eq. (6)

$$\mathbf{c} = \left(\Sigma + \frac{n\lambda}{2}I\right)^{-1}\tilde{y} \in \mathbb{R}^n.$$

Proof: Expanding the expression $\frac{1}{n} \;\; \|\tilde{y} - \Sigma\mathbf{c}\|_{l_2}^2 + \lambda \mathbf{c}^T\Sigma\mathbf{c}$ in the minimization problem

$$\tilde{y}^T\tilde{y}/n - 2\mathbf{c}^T\Sigma^T\tilde{y}/n + \mathbf{c}^T\Sigma^T\Sigma\mathbf{c}/n + \lambda\mathbf{c}^T\Sigma c,$$

and taking the derivative w.r.t $\mathbf{c}$ and setting to zero:

$$\begin{aligned}
\frac{-2\Sigma^T\tilde{y}}{n} + \frac{2\Sigma^T\Sigma\mathbf{c}}{n} + \lambda\Sigma\mathbf{c} &= 0 \\
2\Sigma^T\Sigma\mathbf{c} + n\lambda\Sigma\mathbf{c} &= 2\Sigma^T\tilde{y} \\
\mathbf{c} &= \left(\Sigma^T\Sigma + \frac{n\lambda\Sigma}{2}\right)^{-1}\Sigma^T\tilde{y} \\
\mathbf{c} &= \Sigma^T\left[\left(\Sigma^T + \frac{n\lambda}{2}I\right)\Sigma\right]^{-1}\tilde{y} \\
\mathbf{c} &= \left(\Sigma + \frac{n\lambda}{2}I\right)^{-1}\tilde{y}
\end{aligned}$$

2.5 Classification rule

The classification rule are the same for the two models considered. Once we have the estimated $\hat{\beta}(t)$ function, we project all training data x_i onto this direction by taking inner product $\int x_i(t)\hat{\beta}(t)\,dt$ to obtain their scores. Then we choose some classification threshold. This threshold could be set upon inspection of the receiver operating characteristic (ROC) curve. This is not particularly important in our final application as we are mostly concerned with producing an ROC curve.

For a new observation, if its score is higher than the threshold then we label it class 1, otherwise class 0. The tuning parameter λ (and bandwidth σ) are chosen with a validation set approach.

3 Multivariate functional data

3.1 Model setup and estimation

In many applications we have more than one functional predictor. For example, we would like to predict some disease status using multimodal imaging, such as the EEG and EKG together. Here we consider an extension of our classifier to multivariate functional data. Assume now we have p functional predictors. The training set consists of $\{x_{1i}, x_{2i}, ..., x_{pi}, g_i\}$, n independent copies of $(X_1, X_2, ..., X_p, G)$. $\forall l \in 1, ..., p, X_l \in \mathcal{L}^2(T)$ is a mean-centered predictor function on the same domain $T = [0, 1]$, and G is a binary outcome random variable.

The estimate $\hat{\mathbf{b}}$ is a stacked vector with

$$\hat{\mathbf{b}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_p \end{pmatrix}$$

where we assume each $\beta_l \in \mathcal{H}_K$, an RKHS with the exponential kernel. We chose this hypothesis space instead of the Sobolev space because the exponential kernel does not have a null space. This simplifies the matrix calculations. $\hat{\mathbf{b}}$ is given by the solution to the following objective function:

$$\boxed{\hat{\mathbf{b}} = \underset{\beta_l \in \mathcal{H}_K}{\arg\min} \ \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{y}_i - \sum_{l=1}^{p} \langle x_{li}, \beta_l \rangle \right)^2 + \lambda \sum_{l=1}^{p} \text{Pen}(\beta_l)}$$

Lemma 3.1 The representer theorem guarantees the solution is of the form

$$\hat{\beta}_l(\cdot) = \sum_{i=1}^{n} c_{li} \int K(t, \cdot) x_{li}(t) \, dt, \ \forall l = 1, 2, ..., p$$

with coefficients

$$\mathbf{c} = (\mathbf{c}_1, ..., \mathbf{c}_p)^T = (\Sigma_h' \Sigma_h + \lambda \Sigma_d)^{-1} \Sigma_h' \tilde{y},$$

where Σ_h is the horizontally stacked matrix of Σ^l from each predictor, and Σ_d is the diagonally stacked matrix of Σ^l from each predictor:

$$\Sigma_h = \left(\Sigma^1, \Sigma^2, ..., \Sigma^p \right) \in \mathbb{R}^{n \times pn}$$

and

$$\Sigma_d = \begin{pmatrix} \Sigma^1 & 0 & 0 & 0 \\ 0 & \Sigma^2 & 0 & 0 \\ ... & & & \\ 0 & 0 & 0 & \Sigma^p \end{pmatrix} \in \mathbb{R}^{pn \times pn}.$$

Proof: Recall the discrete objective function of univariate estimation with the exponential kernel in Section 2.3,

$$\hat{\mathbf{c}} = \underset{\mathbf{c}\in\mathbb{R}^n}{\text{minimize}} \frac{1}{n} \ \|\tilde{y} - \Sigma\mathbf{c}\|_{l_2}^2 + \lambda \mathbf{c}^T\Sigma\mathbf{c}.$$

We can simply build upon it with new predictors in an additive fashion (taking $p = 3$ as an example):

$$(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \hat{\mathbf{c}}_3) = \underset{c_1,c_2,c_3\in\mathbb{R}^n}{\arg\min} \ \|\tilde{y} - \Sigma^1\mathbf{c}_1 - \Sigma^2\mathbf{c}_2 - \Sigma^3\mathbf{c}_3\|^2 + \lambda\bigg(\mathbf{c}_1^T\Sigma^1\mathbf{c}_1 + \mathbf{c}_2^T\Sigma^2\mathbf{c}_2 + \mathbf{c}_3^T\Sigma^3\mathbf{c}_3\bigg),$$

where for each predictor we have a different matrix:

$$\Sigma_{ij}^l = \int\int x_{li}(s)K(s,t)x_{lj}(t)\,ds\,dt, \quad l = 1, 2, ..., p.$$

The objective function's notation can be simplified. We can stack the unknown coefficients $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_p$ into one vector $\mathbf{c} \in \mathbb{R}^{pn}$

$$\mathbf{c} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ ... \\ \mathbf{c}_p \end{pmatrix}$$

and write the objective function as:

$$\boxed{\hat{\mathbf{c}} = \underset{c\in\mathbb{R}^{pn}}{\arg\min} \|\tilde{y} - \Sigma_h\mathbf{c}\|_2^2 + \lambda\mathbf{c}^T\Sigma_d\mathbf{c}}$$

where Σ_h is the horizontally stacked matrix of Σ^l from each predictor, and Σ_p is the diagonally stacked matrix of Σ^l from each predictor.

We can solve for $\hat{\mathbf{c}}$ by using the same differentiation techniques like the univariate model with exponential kernel.

Expanding the objective function:

$$\tilde{y}'\tilde{y} - c^T\Sigma_h'\tilde{y} - \tilde{y}'\Sigma_h c + c^T\Sigma_h'\Sigma_h c + \lambda c^T\Sigma_d c.$$

Differentiating with respect to $\mathbf{c}$ and setting to zero:

$$\begin{aligned} -2\Sigma_h'\tilde{y} + 2\Sigma_h'\Sigma_h\mathbf{c} + 2\lambda\Sigma_d\mathbf{c} &= 0 \\ \Sigma_h'\Sigma_h\mathbf{c} + \lambda\Sigma_d\mathbf{c} &= \Sigma_h'\tilde{y} \\ \mathbf{c} &= (\Sigma_h'\Sigma_h + \lambda\Sigma_d)^{-1}\Sigma_h'\tilde{y}. \end{aligned}$$

In the R code implementation, there is an issue of singularity due to numeric integrations. We add a small identity matrix to $(\Sigma_h'\Sigma_h + \lambda\Sigma_d)$ to ensure we can compute the matrix inverse.

3.2 Classification rule

We estimate the discriminant directions $\hat{\beta}_l$ using the training data, and evaluate the score via the sum of projections:

$$\sum_{l=1}^{p} \langle x_{li}, \hat{\beta}_l \rangle.$$

Given a new subject with mean-centered predictors $(x_1^*, x_2^*, ..., x_p^*)$ we assign to class 1 if the score is greater than the threshold, class 0 if the score is lower. λ and σ are chosen by computing sensitivity and specificity on a test set.

Note that we use the same λ to control the regularization for all predictors in the objective function. This is a potential limitation since predictors can take values in very different scales, so this results in uneven regularization of the respective $\hat{\beta}_l$ estimates. Having p different λ parameters would account for this difference and flexibly regularize each predictor, but tuning them would be computationally time-consuming.

4 Extension: Multi-class outcome

4.1 Model setup

So far we have focused on the case where the outcome is binary. Now we consider the extension where the outcome falls into $K > 2$ classes and we have one functional predictor. Without loss of generality, we assume that there are three classes. This makes our illustration for decomposing the objective function easier. Four classes or more work in the same way.

As previously mentioned, the OS formulation can be straightforwardly extended to this case as follows.

We want to find two directions to project the data such that the classes are maximally separated. The training sample consists of $\{(x_i, g_i), i = 1, .., n\}$ from the random variable pair (X, G). $P(G = g_1) = \pi_1$, $P(G = g_2) = \pi_2$, and $P(G = g_3) = \pi_3$. $X \in \mathcal{L}^2$ is a square-integrable random function. The labels are stored in an $n \times 3$ indicator matrix Y. $y_{ik} = 1$ if $G_i = g_k$, for $k = 1, 2, 3$, and $y_{ik} = 0$ otherwise. For example, the first four samples are class 1, class 3, class 2, class 2 etc.

$$Y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \cdots & & \end{pmatrix}$$

Denote $B(t) = [\beta_1(t), \beta_2(t)]$ as a vector of functions, and $\Theta = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2] \in \mathbb{R}^{3\times 2}$ as the matrix of scores to transform the labels.

The multi-class classification objective function is given by:

$$\underset{\Theta,B}{\text{minimize}} \ (2n)^{-1} \left\| Y\Theta - \int \boldsymbol{X}(t)B(t) \right\|^2 + \text{Pen}(B(t)),$$

subject to

$$n^{-1}\Theta^T Y^T Y\Theta = I_2, \quad \Theta^T Y^T Y 1 = 0,$$

where $\boldsymbol{X}(t)$ is the $x_i(t)$ functions stacked vertically, $i = 1, ..., n$.

It is easy to show that this minimization problem is equivalent to

$$\underset{\beta_1,\beta_2,\theta_1,\theta_2}{\text{minimize}} \ \frac{1}{2n} \left\| Y\boldsymbol{\theta}_1 - \int \boldsymbol{X}(t)\beta_1(t) \right\|^2 + \frac{1}{2n} \left\| Y\boldsymbol{\theta}_2 - \int \boldsymbol{X}(t)\beta_2(t) \right\|^2 + \lambda(\text{Pen}(\beta_1) + \text{Pen}(\beta_2)).$$

subject to

$$\Theta^T Y^T Y\Theta = nI_2, \quad \Theta^T Y^T Y\mathbf{1} = 0,$$

with column vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^3$. The penalty is the sum of the second derivatives:

$$\text{Pen}(B(t)) = \int (\beta_1''(t))^2 + (\beta_2''(t))^2.$$

To see why we can separate the minimization problem column-wise, we use a toy example to illustrate the F-norm. The numbers are made up for illustration.

$$\left\| Y\Theta - \int \boldsymbol{X}(t)B(t) \right\|_F^2 = \left\| \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ ... & & \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0.1 & 0.2 \\ 0.3 & 0.3 \\ 0.6 & 0.5 \end{pmatrix} - \begin{pmatrix} \int x_1(t)\beta_1(t) & \int x_1(t)\beta_2(t) \\ ... & ... \\ \int x_n(t)\beta_1(t) & \int x_n(t)\beta_2(t) \end{pmatrix} \right\|_F^2$$

Since the F-norm is an element-wise addition, we can split it by columns

$$\left\| Y\Theta - \int \boldsymbol{X}(t)B(t) \right\|_F^2 = \left\| (Y\Theta)_{.,1} - \left[\int \boldsymbol{X}(t)B(t) \right]_{.,1} \right\|^2 + \left\| (Y\Theta)_{.,2} - \left[\int \boldsymbol{X}(t)B(t) \right]_{.,2} \right\|^2.$$

By the properties of matrix multiplication,

$$(Y\Theta)_{.,1} = Y\boldsymbol{\theta}_1, \quad \left[\int \boldsymbol{X}(t)B(t) \right]_{.,1} = \int \boldsymbol{X}(t)\beta_1(t).$$

So the F-norm becomes:

$$\left\| Y\Theta - \int \boldsymbol{X}(t)B(t) \right\|_F^2 = \left\| Y\boldsymbol{\theta}_1 - \int \boldsymbol{X}(t)\beta_1(t) \right\|^2 + \left\| Y\boldsymbol{\theta}_2 - \int \boldsymbol{X}(t)\beta_2(t) \right\|^2,$$

which is a column-wise decomposition of the norm. This conveniently reduces the multi-class problem to the binary setting!

4.2 Estimation

A closed form for Θ can be derived by applying Lemma 1 and 2 in (Gaynanova 2020), where it is shown that similarly to the binary case, the optimal Θ does not depend on B. However, Lemma 1 in (Gaynanova 2020) requires that the following property holds: $\text{Pen}(B(t)) = \text{Pen}(B(t)A)$ for any $B \in \mathcal{H}\times\mathcal{H}$ and any orthogonal matrix $A \in \mathbb{R}^{K-1\times K-1} = \mathbb{R}^{2\times 2}$.

The second derivative penalty is in fact invariant under orthogonal transformation, i.e. $\text{Pen}(B(t)) = \text{Pen}(B(t)A)$. So we can apply this result in our estimation.

Step 1 By lemma 2 in (Gaynanova 2020), let $\hat{\Theta} \in \mathbb{R}^{K\times(K-1)} = R^{3\times 2}$ have columns $\hat{\boldsymbol{\theta}}_l \in \mathbb{R}^K, l = 1, .., K-1$,

$$\hat{\boldsymbol{\theta}}_l = \left[\left(\sqrt{\frac{nn_{l+1}}{\sum_{i=1}^{l} n_i \sum_{i}^{l+1} n_i}}\right)_l, \quad -\sqrt{\frac{n\sum_{i=1}^{l} n_i}{n_{l+1}\sum_{i=1}^{l+1} n_i}}, \quad (0)_{K-1-l}\right]^T.$$

Then this matrix of scores satisfy the two constraints:

$$\hat{\Theta}^T Y^T Y \hat{\Theta} = nI_{K-1}, \quad \hat{\Theta}^T Y^T Y \mathbf{1} = 0.$$

Step 2 Given Θ, the estimate for $B_{\hat{\Theta}} = [\hat{\beta}_1, \hat{\beta}_2]$ is given by:

$$\hat{B}_{\hat{\Theta}} = \underset{\beta_1,\beta_2}{\text{minimize}} \ \left\|Y\hat{\boldsymbol{\theta}}_1 - \int_T X(t)\beta_1(t)\right\|^2 + \left\|Y\hat{\boldsymbol{\theta}}_2 - \int_T X(t)\beta_2(t)\right\|^2 + \text{Pen}(B(t)).$$

This is equivalent to minimizing them separately over β_1 and β_2:

$$\underset{\beta_1}{\text{minimize}} \left\|Y\hat{\theta}_1 - \int_T x(t)\beta_1(t)\right\|^2 + \int(\beta_1''(t))^2 dt \ + \underset{\beta_2}{\text{minimize}} \left\|Y\hat{\theta}_2 - \int_T x(t)\beta_2(t)\right\|^2 + \int(\beta_2''(t))^2 dt.$$

So similarly to binary classification results in section 2,

$$\hat{\beta}_1(t) = d_{11}\xi_1(t) + d_{12}\xi_2(t) + \sum_i^n c_{1i} \int K(t,\cdot)x_i(t)\,dt,$$

$$\hat{\beta}_2(t) = d_{21}\xi_1(t) + d_{22}\xi_2(t) + \sum_i^n c_{2i} \int K(t,\cdot)x_i(t)\,dt.$$

Once we have the two discriminant directions $\hat{\beta}_1$ and $\hat{\beta}_2$, we can use the functional analog of Fisher's LDA classification rule on the two-dimensional projections of the data. Further developments of this section are left for future work.

5 Simulation study of the univariate classifier

In this section, we perform a simulation study of the univariate classifier with a binary outcome.

Recall that the true discriminant function is $\beta_0 \in \mathcal{H}$ such that $L_C\beta_0 = \mu_1 - \mu_0$. To satify our assumptions, we will constrain our data such that $E[X] = 0$. Let $X(t)$ be a combination of 20 orthonormal basis. Consider

$$X(t) = \sum_{k=1}^{20} Z_k \phi_k(t)$$

where the basis functions are given by

$$\phi_1(t) = 1,$$

$$\phi_{k+1}(t) = \sqrt{2}\cos(k\pi t) \text{ for } k > 1$$

and the coefficients are

$$Z_k \sim N(0, \sigma_k^2),$$

where σ_k^2 are positive constants decreasing in k.

We can get an expression of the covariance and its (psuedo-)inverse using the properties of orthonormal basis:

$$\begin{aligned} C(s,t) &= E[X(s)X(t)] \\ &= E\left[\sum_{k=1}^{20} Z_k\phi_i(s) \sum_{k=1}^{20} Z_k\phi_k(t)\right] \\ &= \sum_{k=1}^{20} \sigma_k^2 \phi_k(t)\phi_k(s), \\ C^{-1}(s,t) &= \sum_{i=1}^{20} \frac{1}{\sigma_i^2}\phi_i(t)\phi_i(s). \end{aligned}$$

Since we want to work with mean-centered data, the mean function must satisfy:

$$E[X(t)] = E[E[X(t)|G]] = \pi_1\mu_1(t) + \pi_0\mu_0(t) = 0.$$

Let

$$\mu_1(t) = \sum_{k=1}^{20} a_k\phi_k(t), \quad \mu_0(t) = \sum_{k=1}^{20}(-a_k)\phi_k(t),$$

for some coefficients $a_1, ..., a_{20} \in \mathbb{R}$. Here we choose $a_k = k^{-2}, k = 1, 2, ..., 20$ and $\sigma_k = k^{-1}, k = 1, 2, ..., 20$. Note these two functions' decay completely characterize the dataset. Later we will change them to obtain different datasets.

Data from the two classes can be generated as:

$$X_1(t) = \mu_1(t) + \sum_{k=1}^{20} Z_k \phi_k(t), \quad X_2(t) = \mu_0(t) + \sum_{k=1}^{20} Z_k \phi_k(t)$$

The true β_0 is given by:

$$\begin{aligned}
\beta_0(t) &= L_C^{-1}(\mu_1 - \mu_0) \\
&= \int C^{-1}(s,t)[\mu_1(s) - \mu_0(s)]\, ds \\
&= \int C^{-1}(s,t)\, 2 \sum_{k=1}^{20} a_k \phi_k(s)\, ds \\
&= 2 \int \left[\sum_{k=1}^{20} \frac{1}{\sigma_k^2} \phi_k(t) \phi_k(s) \right] \left[\sum_{k=1}^{20} a_k \phi_k(s) \right] ds \\
&= 2 \int \sum_{k=1}^{20} \frac{a_k}{\sigma_k^2} \phi_k(t) \phi_k(s)^2\, ds \\
&= 2 \sum_{k=1}^{20} \frac{a_k}{\sigma_k^2} \phi_k(t).
\end{aligned}$$

We simulated 200 training samples for estimation, and we assume $\pi_1 = \pi_0 = 0.5$. We choose $m = 300$, which is the number of sampling points along the domain T. We want to illustrate the case where functions are finely sampled and $m > n$. We then tune the parameters on a validation set of 200 samples. The list of values we try are $\lambda, \sigma \in \{10^{-5}, 10^{-4}, 10^{-3}, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50\}$.

In addition, we implemented the truncation estimator from (Delaigle & Hall, 2011). Denote the sample mean difference as $\hat{d}(t) = \hat{\mu}_1(t) - \hat{\mu}_0(t)$. Their estimator is

$$\hat{\beta}^{(r)}(t) = \sum_{k}^{r} \frac{\hat{d}_k}{\hat{\sigma}_k^2} \hat{\phi}_k(t),$$

where $\hat{\phi}_k$ and $\hat{\sigma}_k^2$ are estimates of the eigenfunctions and eigenvalues of the covariance function, estimated with empirical principal component analysis (PCA) of the entire dataset, using R function prcomp(). And $\hat{d}_k = \int_T \hat{d}(t)\hat{\phi}_k(t)$. The tuning parameter $r \in \{1, ..., 20\}$ is chosen using a validation set of 200 samples. Finally we compare the performance of the classifiers on a test set of 200 samples.

The two metrics we use are AUC and the estimation error, i.e., $\|\beta_0 - \hat{\beta}\|^2 = \int (\beta_0(t) - \hat{\beta}(t))^2 dt$. The main metric of importance to us is the AUC, since for a classification task, we mostly care about prediction, not estimation.

Fig. 1 shows the simulated data.

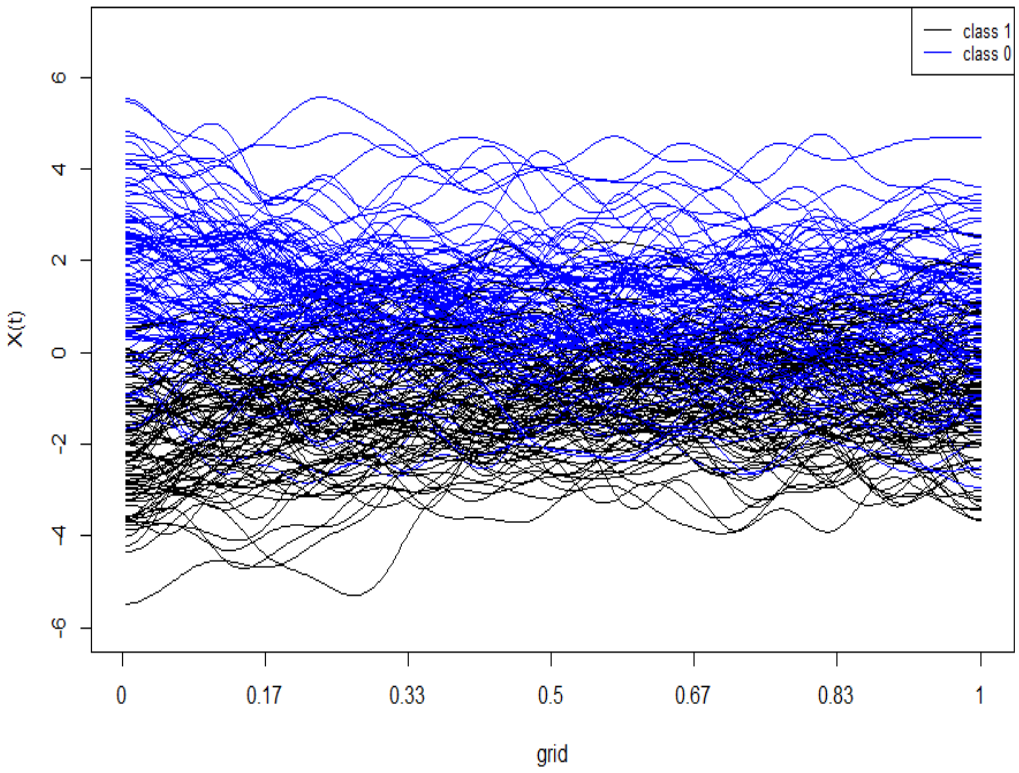

Figure 1: simulated data of 200 training samples.

Fig. 2 and Fig. 3 display the performance of the classifier on the validation set, for β belonging to the RKHS with exponential kernel. The maximum AUC is 0.964, achieved by $\lambda = 0.002, \sigma = 0.005$. The minimum estimation error is 85.2, achieved by $\lambda = 50, \sigma = 0.0001$.

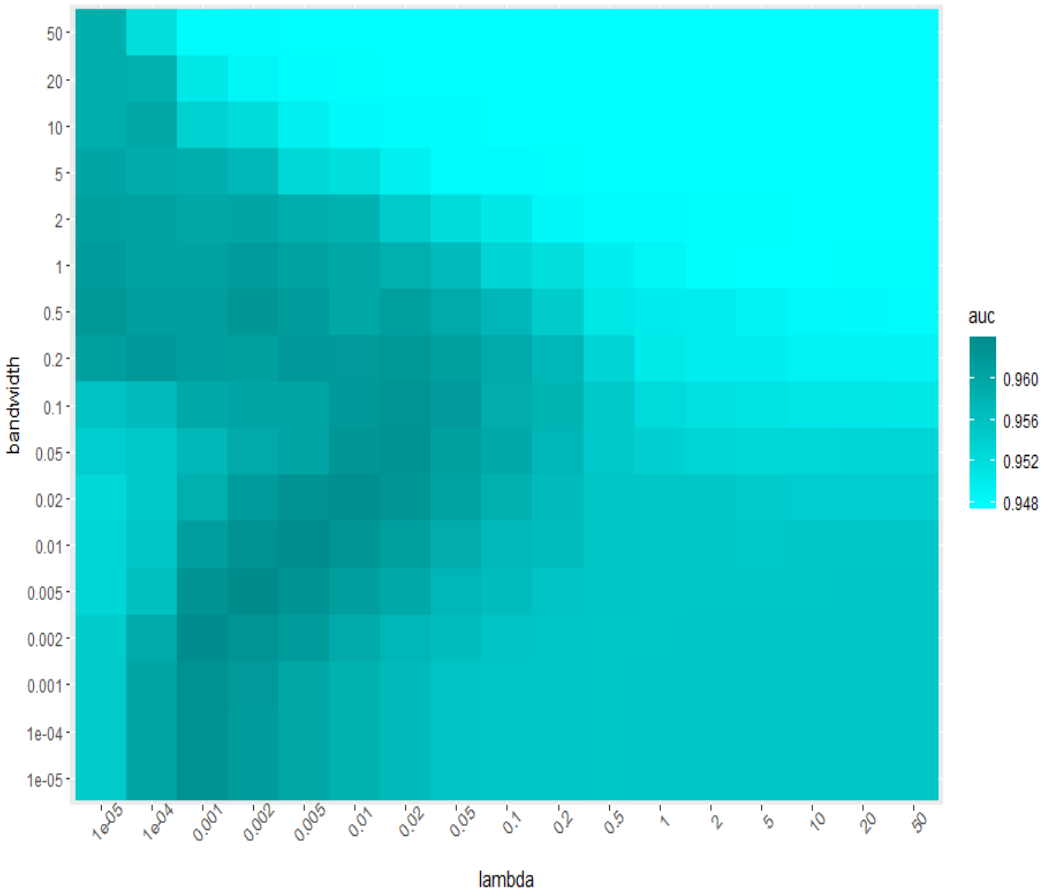

Figure 2: β is estimated with the exponential kernel. The color intensity shows the AUC of classification on the validation dataset, while varying the parameter σ and λ together. We can appreciate the fact that increasing λ or σ (and therefore the smoothness of the estimate), starts underfitting the model and leads to lower AUC.

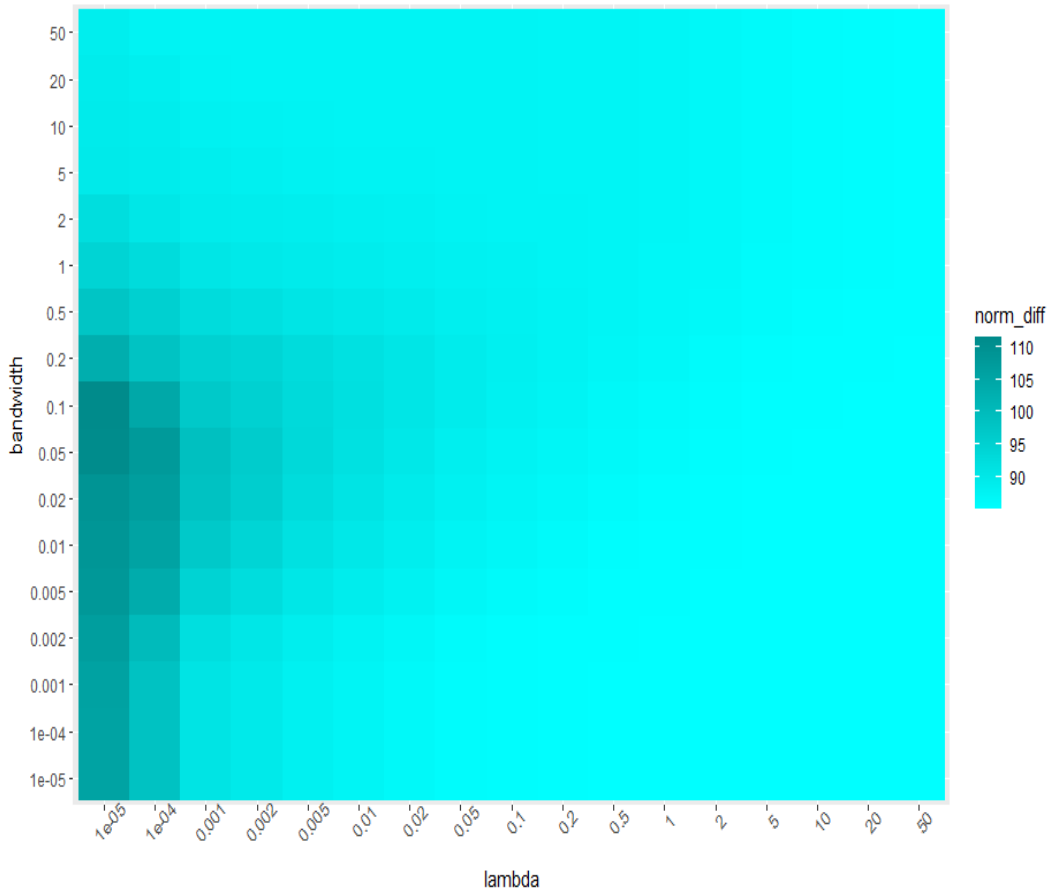

Figure 3: β is estimated with the exponential kernel. Color intensity represents the estimation error $\|\hat{\beta} - \beta_0\|$

Fig. 4 and Fig. 5 display the performance of the classifier with β belonging to the Sobolev space. The maximum AUC is 0.962, and the parameter achieving this is $\lambda = 0.0001$. The minimum estimation error is 89.4, achieved by $\lambda = 50$.

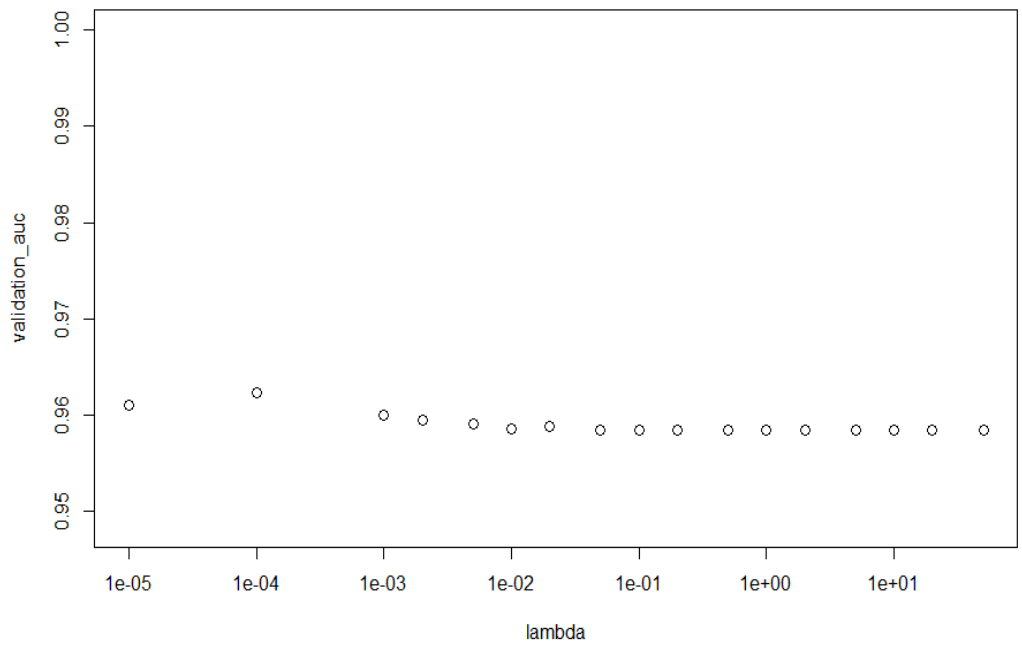

Figure 4: β is estimated with second derivative penalty. AUC of classification on the validation dataset, while varying the parameter λ.

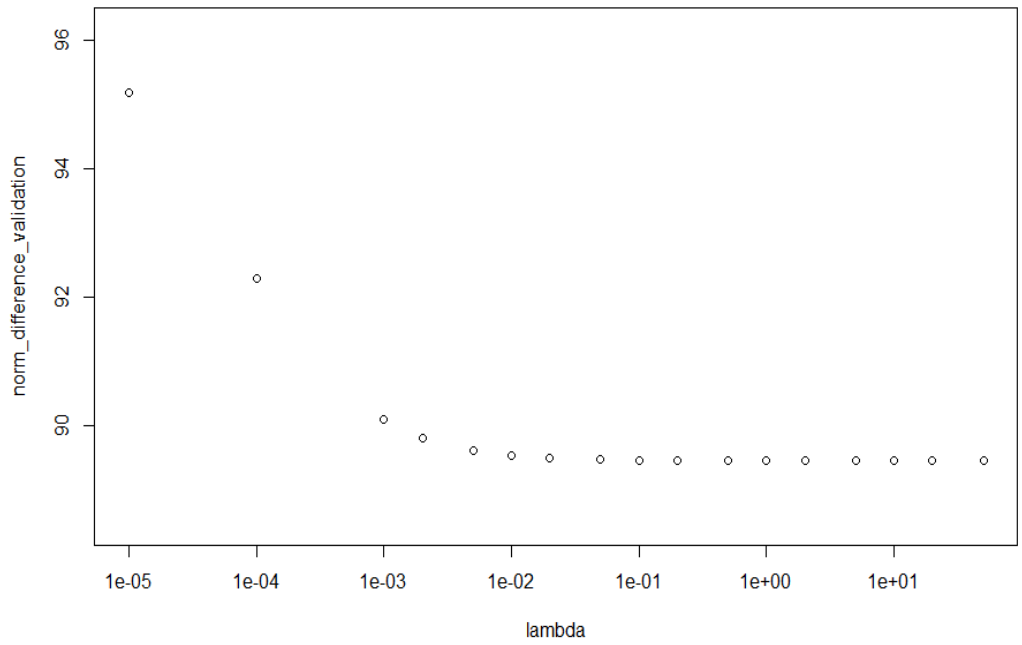

Figure 5: β is estimated with second derivative penalty. Estimation error $\|\hat{\beta} - \beta_0\|$ on the validation dataset.

Fig. 6 and Fig. 7 display the performance of the truncation estimator. The maximum AUC is 0.948, achieved by $r = 9$. The minimum estimation error is 85.2, achieved by $r = 1$.

Finally, the test AUC are 0.976, 0.976, 0.965 for the classifier estimated with exponential kernel, Sobolev space, and the truncation estimator respectively. Our methods are slightly favorable compared to the truncation estimator in this setting.

At first it seems counter-intuitive that the tuning parameters that lead to the highest AUC are not the ones that have the smallest estimation error. After plotting the best β estimates - in terms of AUC and estimation error - against the true β_0 in Fig. 8, we see that all six of the estimates are smoother compared to β_0. Therefore the difference in estimation error between the estimates are negligible in the presence of β_0. In fact, the initial sections of β_0 is very large due to the interplay between the covariance and the sample mean difference functions' decay.

We then repeated this analysis for different values of a_k and σ_k, namely we change the rate of decay of the mean difference function and the covariance function, so that they are equal to $k^{-1}, k^{-1.5}, k^{-2}, k^{-2.5}, k^{-3}$, for $k = 1, 2, ..., 20$. This results in 25 different datasets. Small a_k leads to smaller difference in the two classes, and small σ_k leads to smaller variance in the data. For each choice of the decay rate, we obtain a best estimate in terms of test AUC and test estimation error. Our methods achieved comparable prediction performance as the Delaigle & Hall truncation estimator across the board. We summarize the results in the tables in Appendix C. The main message is that we should interpret $\|\beta_0 - \hat{\beta}\|$ cautiously, because β_0 quickly explodes as σ_k decays faster. It is unstable for every choice of rate decay except $\sigma_k = k^{-1}$.

Note we only display the figures for the case where $a_k = k^{-2}$ and $\sigma_k = k^{-1}$.

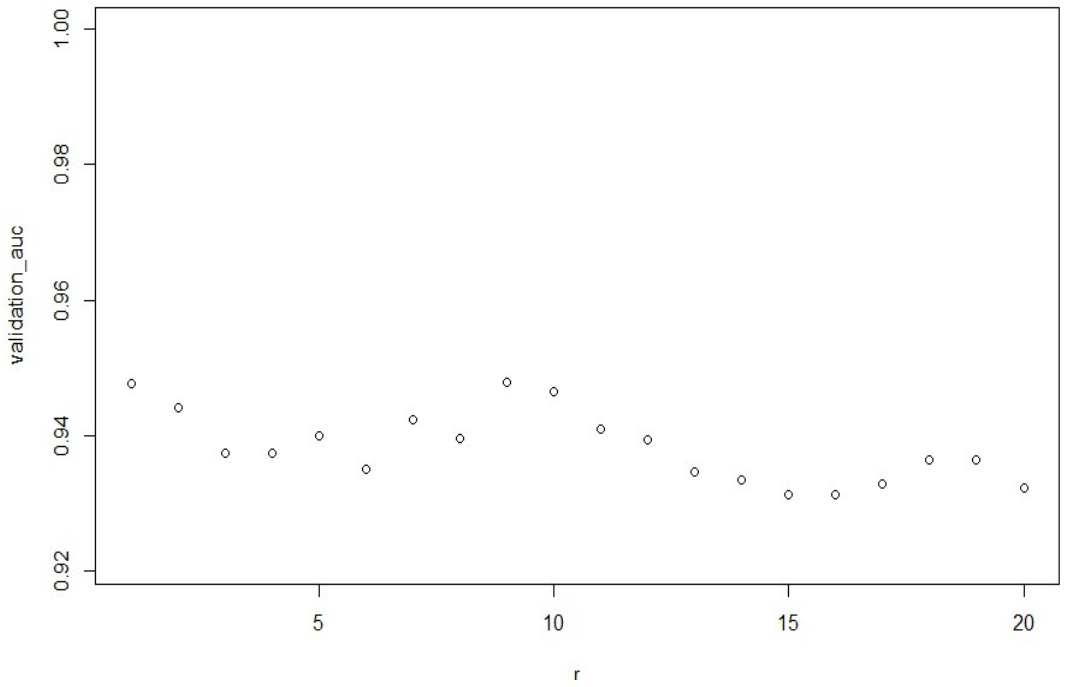

Figure 6: AUC of the truncation estimator, while varying r.

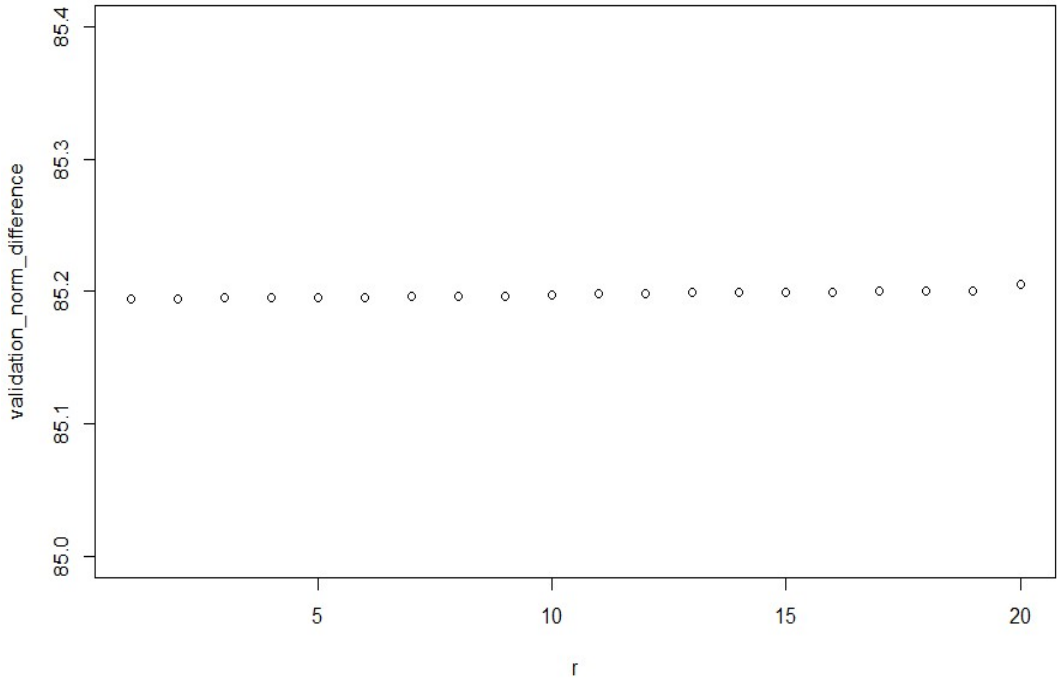

Figure 7: Estimation error $\|\hat{\beta} - \beta_0\|$ of the truncation estimator, while varying r. The estimation error essentially stays the same.

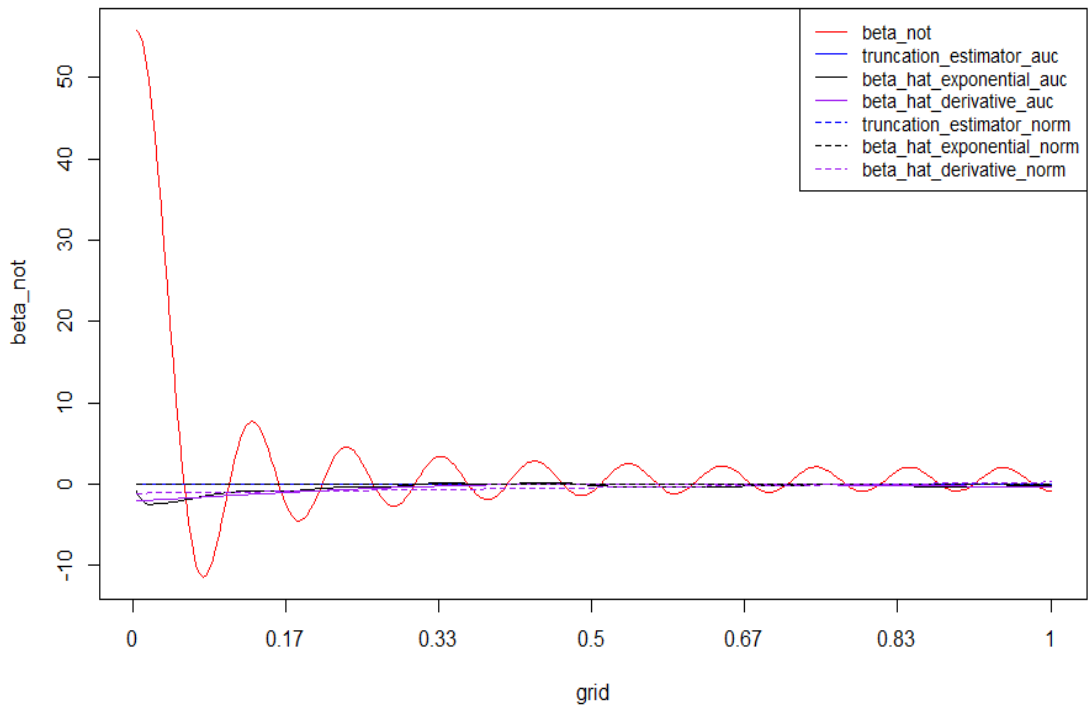

Figure 8: β_0 compared to the six estimators. For each method, the estimates are chosen by best AUC or best estimation error. β_0 is less smooth due to the interplay between the decay of the class means' difference and covariance functions. Particularly, it is highly sensitive to the covariance.

6 Application of multivariate classifier

We apply our multivariate classifier on a real-world diffusion magnetic resonance imaging (dMRI) dataset, provided by Prof. Ariel Rokem at UW Psychology. We briefly provide some background on the data application.

6.1 Background

6.1.1 Anisotropy and diffusion tensor

For most fluids and some homogeneous solid materials like gels, diffusion is the same in every direction. These substances are called isotropic and are characterized by a single diffusion coefficient (D). Biological tissues, on the other hand, are highly structured and typically have different diffusion coefficients along different directions. They are called anisotropic. White matter, which is found in the deeper tissues of the brain, contains nerve fibers (axons), which, roughly speaking, connect the neuronal cell bodies, mostly located on the cortical surface. White matter is highly anisotropic because of the parallel orientation of its nerve fibers, or axons.

In anisotropic materials, diffusion is described by a 3×3 array called the diffusion tensor (DT):

$$\begin{pmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yx} \\ D_{zx} & D_{zy} & D_{zz} \end{pmatrix}$$

The three diagonal elements (D_{xx}, D_{yy}, D_{zz}) of the DT represent diffusion coefficients measured along each of the principal (x, y, z) directions. The six off-diagonal terms reflect correlation between random motions along the principal directions.

6.1.2 Diffusion tensor magnetic resonance imaging

Diffusion tensor magnetic resonance imaging (DTI) is a popular technology for imaging the white matter of the brain. (Jensen & Helpern, 2010) The DT was originally proposed for use in magnetic resonance imaging (MRI) in 1994. Prior to the introduction of the DT model, the orientation of the axons in a tissue sample had to be known to measure anisotropic diffusion, which limited its usage. The DT model allowed, for the first time, a rotationally invariant description of water diffusion. The invariance to rotation was crucial because it enabled application of the DTI to the complex anatomy of the human brain fiber tracts.

DTI has been applied to a variety of neuroscience studies including schizophrenia, traumatic brain injury, multiple sclerosis, autism, and aging. Anatomy studies have investigated the location, asymmetry, and variability of the fiber tracts. Recent efforts were attempted to model the human "connectome" by analyzing structural versus functional brain connectivity, measured by DTI and functional MRI.

The DT model describes the diffusion of water molecules using a Gaussian distribution. It is a 3×3 symmetric, positive-definite covariance matrix, hence it has three orthogonal

eigenvectors and three positive eigenvalues. The major eigenvector of the DT points in the principal diffusion direction (the direction of the fastest diffusion). In anisotropic tissues, the major eigenvector also defines the fiber tract axis of the tissue. Thus the three orthogonal eigenvectors can be thought of as a local fiber coordinate system. However, this interpretation is only strictly true in regions where fiber tracts do not cross, fan, or branch.

The three positive eigenvalues of the tensor give the diffusivity in the direction of each eigenvector. Together, the eigenvectors and eigenvalues define an ellipsoid that represents an isosurface of Gaussian diffusion probability: the axes of the ellipsoid are aligned with the eigenvectors and their lengths are proportional to the eigenvalues.

To measure diffusion using MRI, magnetic field gradients are employed to create an image that is sensitized to diffusion in a particular direction. By repeating this process of diffusion weighting in multiple directions, a three-dimensional DT can be estimated.

Several scalar summaries can be obtained from DTI (Jensen & Helpern, 2010). The simplest is the average of the DT's eigenvalues, referred to as the mean diffusivity (MD). Another measure is called the fractional anisotropy (FA), which is the fraction of the diffusion that is anisotropic. Intuitively, it is the difference of the DT ellipsoid's shape from the shape of a perfect sphere (which represents isotropic diffusion). FA's formula is proportional to a normalized variance of the eigenvalues. MD and FA are considered a measures of "white matter integrity".

6.1.3 Diffusional kurtosis imaging

In DTI, the water diffusion probability distribution is assumed to be Gaussian. However, it has been observed that, the water diffusion distribution in the brain is not precisely Gaussian (O'Donnell & Westin, 2011). This could arise from barriers to diffusion such as cell membranes and organelles, as well as the distinct compartments with differing diffusivity. Quantification of diffusional non-Gaussianity can be useful in characterizing the associated tissue structures.

The kurtosis is a dimensionless statistical metric for quantifying non-Gaussianity of an arbitrary distribution. It has been shown that diffusion-weighted imaging can be used to estimate the kurtosis of water diffusion distribution in the brain. This method is referred to as diffusional kurtosis imaging (DKI). It is a natural extension of DTI. With DKI, one not only obtains the standard DTI metrics such as MD and FA (with modified formulas), but also additional metrics including mean kurtosis (MK) and axonal water fraction (AWF). (Fieremans et al., 2011) MK corresponds to the average kurtosis over all directions. AWF is the fraction of MRI visible water in the axons relative to the total visible water signal. It is a function of the maximum kurtosis over all directions.

6.2 Tractography

Tractography is a novel technique that allows us to reconstruct 3D curves that represent the orientation and shape of nerve tracts using data collected by diffusion MRI. These

curves are clustered in bundles, i.e., groups of nerve tracts linking different parts of the brain.

For each bundle, a centerline is computed and the anisotropy measurements above are measured along this centerline.

We will focus on the joint analysis of a centerline 'shape' and the associated measurements, shown in Fig. 9 and Fig. 10.

6.3 Application

The data provided to us contain multiple bundles of the brain. Here we choose to focus on bundle ARC_L, which is implicated in cognition. We have seven functional predictors; three are location variables of the centerline of the bundle in x, y, z coordinates, and four are diffusion-related measurements along the bundle: FA, MD, MK, AWF. All predictor functions are discretized with 100 sampling points along the bundle ($m = 100$). The data matrix stacks the predictors column-wise. For each subject (each row), we have

$$\begin{pmatrix} x & y & z & dki_{fa} & dki_{md} & dki_{mk} & dki_{awf} \end{pmatrix}$$

where each predictor is a 1×100 vector. The overall data matrix dimension is $n \times 700$.

The coordinate vectors may carry information on translation and rotation, which are not physiologically meaningful. We therefore perform rigid alignment. We register the 3D coordinate data with the R package called shapes, using function procGPA() (generalized procrustis analysis). The data registration step removes translation and rotation information. The data registration reduces the variance.

To obtain the neurological outcome variable, we leveraged the Human Connectome Project (HCP) database, which contains scans and biological variables from 1200 healthy adults. Our outcome of interest is the NIH Toolbox Picture Sequence Memory Test Unadjusted Scale Score, one of many variables from the HCP.

The Picture Sequence Memory Test is a measure developed for the assessment of episodic memory for ages 3-85 years. It involves recalling increasingly lengthy series of illustrated objects and activities that are presented in a particular order on the computer screen. The participants are asked to recall the sequence of pictures that is demonstrated over two learning trials; sequence length varies from 6-18 pictures, depending on age. The test takes approximately 7 minutes to administer. This test score will serve as a surrogate measurement of subjects' episodic memory. Note that it is an extremely reductive and narrow surrogate of a complex cognition function.

The DKI dataset consists of 1038 subjects in total. We match the subjects ID with the available HCP data. Since we are interested in classifying poor memory vs. good memory, we sample the tails of the memory score distribution. We filter the dataset to include people above the top 10% memory score or below the bottom 10%, resulting in 216 subjects in the final dataset. We plot the seven predictors in Fig. 9 and Fig. 10.

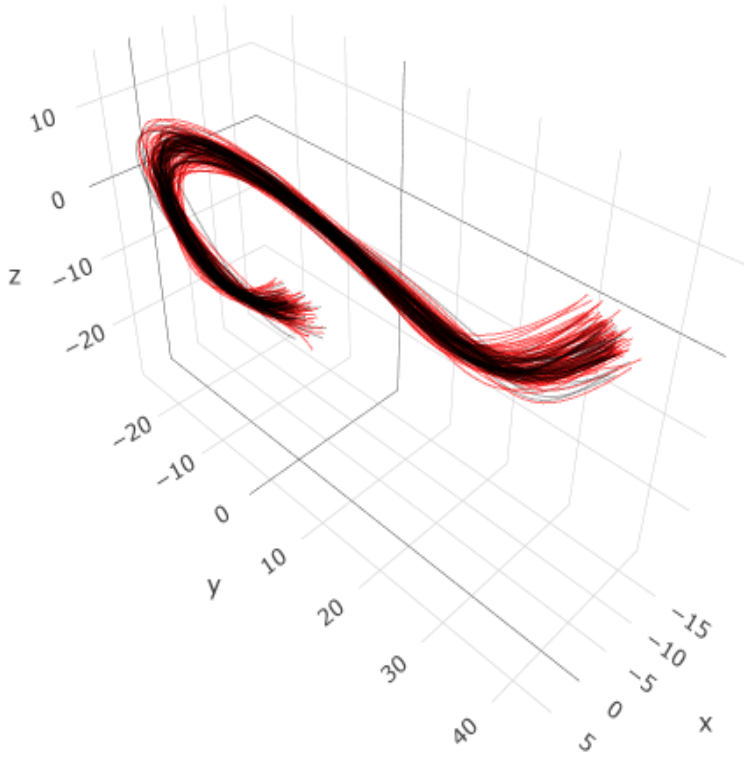

Figure 9: Spatial coordinates of 216 subjects' neural bundles. Black representing class 0 (good memory), red representing class 1 (poor memory).

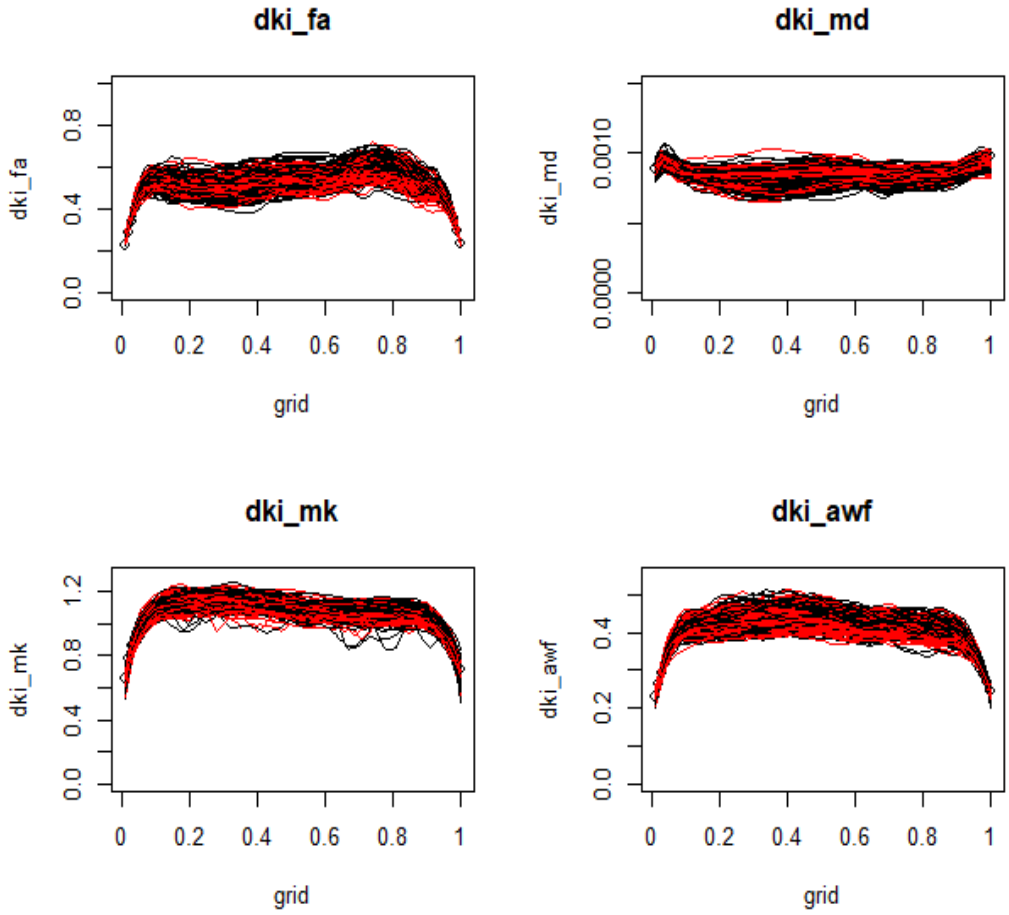

Figure 10: four functional predictors measuring diffusion alignment along the bundle tracks. Black representing class 0 (good memory), red representing class 1 (poor memory).

Even though we sampled the distribution tails of the memory scores, there is very little separation of the bundle measurements, as shown by Fig. 9 and Fig. 10, making the problem very challenging.

Note the y-axis in Fig. 10. The predictor dki_{md} is 1000 times smaller than than the other three predictors. Using the same λ to control the regularization of all β functions is a limitation. For example, if the value of λ is too small, it will adequately regularize the β function corresponding to dki_{md}, but not the other predictors. We address this problem in an ad-hoc way: we multiply the values of dki_{md} by 1000 so that it's on the same scale as the other predictors.

We split the data into 120 training samples and 96 test samples, and fitted the classifier on the training samples.

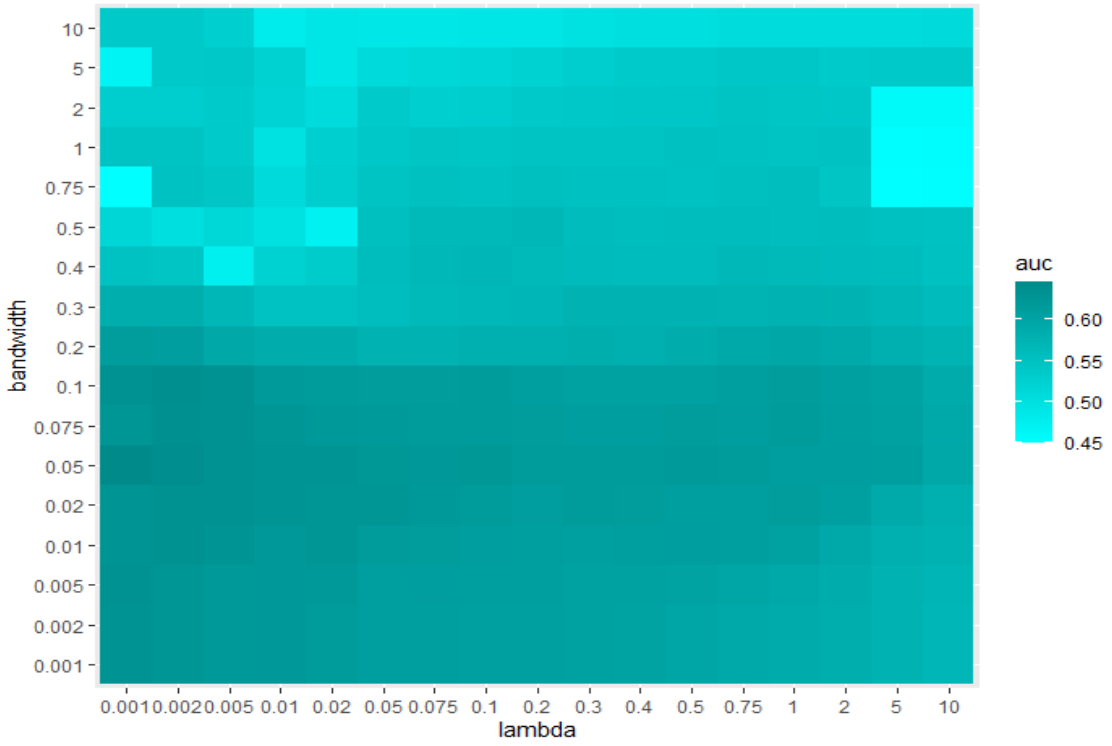

Figure 11: AUC of classification on 96 test samples, with different lambda and bandwidth parameters.

The maximum test AUC achieved is 0.65, where $\lambda = 0.001$ and bandwidth $\sigma = 0.05$. This is comparable to classification attempts in the literature (e.g., Ferrando et al. 2020, Ghiassian et al. 2016)

7 Conclusion & Discussion

We introduced penalized discriminant analysis for multivariate functional data supported on a common, compact 1D domains. Thanks to the connection to optimal scoring, we

recast this problem into a penalized regression framework. The resulting model was very efficient. We applied the multivariate classifier to a diffusion MRI dataset, with the goal of classifying people into good vs. poor episodic memory groups. We obtained adequate prediction performance. Future studies could investigate other neural bundles (since we only focused on the ARC_L bundle), as well as different outcome variables in the HCP to identify potential associations. We noted that a limitation of the multivariate classifier was requiring one regularization parameter for all functional predictors, which lacked flexibility. This can be overcome by scaling the data by hand, but more efficient and flexible ways to regularize different functional predictors could be developed in the future.

Appendix A Derivations

Here we develop the lemmas from Section 1.

A.1 Lemma 1.1

For any given data function $X \in \mathcal{L}^2(T)$, the linear classifier calculates the distance between the data to the two class means in the projection space. Then X is assigned to the class with the shorter distance. The classifier can be re-written in terms of the class means' average and class means' difference.

$$f(X) = 1\Big\{\langle X-\mu_0,\beta\rangle^2 - \langle X-\mu_1,\beta\rangle^2 > 0\Big\} = 1\Big\{\langle \delta,\beta\rangle\langle X-\bar{\mu},\beta\rangle > 0\Big\}$$

where $\delta = \mu_1 - \mu_0$ and $\bar{\mu} = (\mu_0+\mu_1)/2$.

Proof:

$$\begin{aligned}
&\langle X-\mu_0,\beta\rangle^2 - \langle X-\mu_1,\beta\rangle^2 \\
&= \Big(\langle X,\beta\rangle^2 - 2\langle\mu_0,\beta\rangle\langle X,\beta\rangle + \langle\mu_0,\beta\rangle^2\Big) - \Big(\langle X,\beta\rangle^2 - 2\langle\mu_1,\beta\rangle\langle X,\beta\rangle + \langle\mu_1,\beta\rangle^2\Big) \\
&= \langle\mu_0,\beta\rangle^2 - 2\langle\mu_0,\beta\rangle\langle X,\beta\rangle + 2\langle\mu_1,\beta\rangle\langle X,\beta\rangle - \langle\mu_1,\beta\rangle^2 \\
&= 2\langle\mu_1-\mu_0,\beta\rangle\langle X,\beta\rangle + \langle\mu_0,\beta\rangle^2 - \langle\mu_1,\beta\rangle^2 \\
&= 2\langle\mu_1-\mu_0,\beta\rangle\langle X,\beta\rangle + \Big(\langle\mu_0,\beta\rangle + \langle\mu_1,\beta\rangle\Big)\Big(\langle\mu_0,\beta\rangle - \langle\mu_1,\beta\rangle\Big) \\
&= 2\langle\mu_1-\mu_0,\beta\rangle\langle X,\beta\rangle + \langle\mu_0+\mu_1,\beta\rangle\langle\mu_0-\mu_1,\beta\rangle \\
&= 2\langle\mu_1-\mu_0,\beta\rangle\langle X,\beta\rangle - \langle\mu_0+\mu_1,\beta\rangle\langle\mu_1-\mu_0,\beta\rangle \\
&= \langle\mu_1-\mu_0,\beta\rangle\Big(2\langle X,\beta\rangle - \langle\mu_0+\mu_1,\beta\rangle\Big) \\
&= \langle\mu_1-\mu_0,\beta\rangle\langle 2X-\mu_0-\mu_1,\beta\rangle
\end{aligned}$$

Hence

$$\langle X-\mu_0,\beta\rangle^2 - \langle X-\mu_1,\beta\rangle^2 = \langle\mu_1-\mu_0,\beta\rangle\langle 2X-\mu_0-\mu_1,\beta\rangle$$

and

$$1\Big\{\langle X-\mu_0,\beta\rangle^2-\langle X-\mu_1,\beta\rangle^2\Big\}=1\Big\{\langle\mu_1-\mu_0,\beta\rangle\langle X-\frac{(\mu_0+\mu_1)}{2},\beta\rangle\Big\},$$

since indicator functions ignore constants.

Therefore

$$f(X)=1\{\langle X-\mu_0,\beta\rangle^2-\langle X-\mu_1,\beta\rangle^2>0\}=1\{\langle\delta,\beta\rangle\langle X-\bar{\mu},\beta\rangle>0\},$$

where $\delta=\mu_1-\mu_0$ and $\bar{\mu}=(\mu_0+\mu_1)/2$.

A.2 Lemma 1.2

Using the classifier

$$f(X)=1\Big\{\langle\delta,\beta\rangle\langle X-\bar{\mu},\beta\rangle>0\Big\},$$

and under the assumption that $\pi_0=\pi_1=1/2$, the misclassification probability is

$$\boxed{err=\frac{P_0[f(X)=1]}{2}+\frac{P_1[f(X)=0]}{2}=1-\Phi\left(\frac{|\langle\mu_1-\mu_0,\beta\rangle|}{2\langle\beta,L_C\beta\rangle^{1/2}}\right)}$$

Proof: Under the assumption that $\pi_0=\pi_1=1/2$, the misclassification probability is

$$err=\frac{P_0[f(X)=1]}{2}+\frac{P_1[f(X)=0]}{2}.$$

Since misclassifications are symmetrical, we can consider one case.

$$\begin{aligned}
err&=\frac{P_0[f(X)=1]}{2}+\frac{P_1[f(X)=0]}{2}\\
&=P_0[f(X)=1]\\
&=P_0\left[\langle X-\bar{\mu},\beta\rangle\langle\delta,\beta\rangle>0\right]\\
&=P_0\left[\langle X-(\mu_0+\mu_1)/2,\beta\rangle\langle\mu_1-\mu_0,\beta\rangle>0\right]\\
&=P_0\left[\Big(\langle X-\mu_0/2,\beta\rangle-\langle\mu_1/2,\beta\rangle\Big)\langle\mu_1-\mu_0,\beta\rangle>0\right]\\
&=P_0\Big[\langle X-\mu_0/2,\beta\rangle\langle\mu_1-\mu_0,\beta\rangle>\langle\mu_1/2,\beta\rangle\langle\mu_1-\mu_0,\beta\rangle\Big]
\end{aligned}$$

We can't divide over the term without considering the two cases.
case 1: $\langle\mu_1-\mu_0,\beta\rangle>0$.

$$
\begin{aligned}
err &= P_0\left[\langle X-\mu_0/2,\beta\rangle\langle\mu_1-\mu_0,\beta\rangle > \langle\mu_1/2,\beta\rangle\langle\mu_1-\mu_0,\beta\rangle\right] \\
&= P_0\left[\langle X-\mu_0/2,\beta\rangle > \langle\mu_1/2,\beta\rangle\right] \\
&= P_0\left[\langle X-\mu_0+\mu_0/2,\beta\rangle > \langle\mu_1/2,\beta\rangle\right] \\
&= P_0\left[\langle X-\mu_0,\beta\rangle + \langle\mu_0/2,\beta\rangle > \langle\mu_1/2,\beta\rangle\right] \\
&= P_0\left[\langle X-\mu_0,\beta\rangle > \langle\mu_1/2,\beta\rangle - \langle\mu_0/2,\beta\rangle\right] \\
&= P_0\left[\langle X-\mu_0,\beta\rangle > \langle\mu_1/2-\mu_0/2,\beta\rangle\right]
\end{aligned}
$$

case 2: $\langle\mu_1-\mu_0,\beta\rangle < 0$.

$$
\begin{aligned}
err &= P_0\left[\langle X-\mu_0/2,\beta\rangle\langle\mu_1-\mu_0,\beta\rangle > \langle\mu_1/2,\beta\rangle\,\langle\mu_1-\mu_0,\beta\rangle\right] \\
&= P_0\left[\langle X-\mu_0/2,\beta\rangle < \langle\mu_1/2,\beta\rangle\right] \\
&= P_0\left[\langle X-\mu_0,\beta\rangle < \langle\mu_1/2,\beta\rangle - \langle\mu_0/2,\beta\rangle\right] \\
&= P_0\left[\langle X-\mu_0,\beta\rangle < \langle\mu_1/2-\mu_0/2,\beta\rangle\right]
\end{aligned}
$$

Each case can be written as the tail probability of a Gaussian variable, we have

$$
err = P_0\left[\langle X-\mu_0,\beta\rangle > \frac{|\langle\mu_1-\mu_0,\beta\rangle|}{2}\right] = 1-\Phi\left(\frac{|\langle\mu_1-\mu_0,\beta\rangle|}{2\langle\beta,L_C\beta\rangle^{1/2}}\right),
$$

since P_0 is the distribution of the projections of class 0. Φ is the standard Normal cumulative distribution.

A.3 Lemma 1.3

The function that solves the maximization problem is

$$
\beta_0 = \underset{\beta}{\text{maximize}}\,\frac{|\langle\mu_1-\mu_0,\beta\rangle|}{|\langle\beta,L_C\beta\rangle^{1/2}|} = L_C^{-1}(\mu_1-\mu_0)
$$

Proof: If $\|L_C^{-1/2}(\mu_1-\mu_0)\| < \infty$,

$$
\begin{aligned}
\frac{|\langle(\mu_1-\mu_0),\beta\rangle|}{\langle\beta,L_C\beta\rangle^{1/2}} &= \frac{|\langle L_C^{-1/2}(\mu_1-\mu_0),L_C^{1/2}\beta\rangle|}{\langle\beta,L_C\beta\rangle^{1/2}} \\
&\le \frac{\|L_C^{-1/2}(\mu_1-\mu_0)\|\,\|L_C^{1/2}\beta\|}{\langle\beta,L_C\beta\rangle^{1/2}} \\
&= \frac{\|L_C^{-1/2}(\mu_1-\mu_0)\|\,\|L_C^{1/2}\beta\|}{\langle L_C^{1/2}\beta,L_C^{1/2}\beta\rangle^{1/2}} \\
&= \|L_C^{-1/2}(\mu_1-\mu_0)\|
\end{aligned}
$$

The first equality comes from the self-adjointness of operator L_C. The inequality in the second line follows from Cauchy-Schwartz:

$$|\langle L_C^{-1/2}(\mu_1 - \mu_0), L_C^{1/2}\beta\rangle| \leq \|L_C^{-1/2}(\mu_1 - \mu_0)\| \, \|L_C^{1/2}\beta\|.$$

If $\|L_C^{-1}(\mu_1 - \mu_0)\| < \infty$, equality is achieved for

$$\beta = L_C^{-1}(\mu_1 - \mu_0),$$

.

which can be seen by plugging it in:

$$\left|\left\langle L_C^{-1/2}(\mu_1 - \mu_0), L_C^{1/2}L_C^{-1}(\mu_1 - \mu_0)\right\rangle\right| = \left|\left\langle L_C^{-1/2}(\mu_1 - \mu_0), L_C^{-1/2}(\mu_1 - \mu_0)\right\rangle\right|,$$

$$\left\|L_C^{-1/2}(\mu_1 - \mu_0)\right\| \, \left\|L_C^{1/2}L_C^{-1}(\mu_1 - \mu_0)\right\| = \left\|L_C^{-1/2}(\mu_1 - \mu_0)\right\| \, \left\|L_C^{-1/2}(\mu_1 - \mu_0)\right\|.$$

In other words,

$$\frac{|\langle \mu_1 - \mu_0, \beta\rangle|}{|\langle \beta, L_C\beta\rangle^{1/2}|}$$

has an upper bound, but it can be achieved when $\beta = L_C^{-1}(\mu_1 - \mu_0)$. Hence this is the solution to the problem of misclassification error minimization.

A.4 Lemma 1.4

The function that minimizes the misclassification error is $\beta_0 = L_C^{-1}(\mu_1 - \mu_0)$. For this choice of β_0 or any multiple of it, the misclassification probability is given by

$$err_0 = 1 - \Phi\left(\frac{\|L_C^{-1/2}(\mu_1 - \mu_0)\|}{2}\right).$$

Proof:

$$
\begin{aligned}
err_0 &= 1 - \Phi\left(\frac{|\langle \mu_1 - \mu_0, \beta_0 \rangle|}{2\langle \beta_0, L_C \beta_0 \rangle^{1/2}}\right) \\
&= 1 - \Phi\left(\frac{|\langle \mu_1 - \mu_0, L_C^{-1}(\mu_1 - \mu_0) \rangle|}{2\langle L_C^{-1}(\mu_1 - \mu_0), L_C L_C^{-1}(\mu_1 - \mu_0) \rangle^{1/2}}\right) \\
&= 1 - \Phi\left(\frac{|\langle L_C^{-1/2}(\mu_1 - \mu_0), L_C^{-1/2}(\mu_1 - \mu_0) \rangle|}{2\langle L_C^{-1}(\mu_1 - \mu_0), (\mu_1 - \mu_0) \rangle^{1/2}}\right) \\
&= 1 - \Phi\left(\frac{\left\| L_C^{-1/2}(\mu_1 - \mu_0) \right\|^2}{2\langle L_C^{-1/2}(\mu_1 - \mu_0), L_C^{-1/2}(\mu_1 - \mu_0) \rangle^{1/2}}\right) \\
&= 1 - \Phi\left(\frac{\left\| L_C^{-1/2}(\mu_1 - \mu_0) \right\|^2}{2\left\| L_C^{-1/2}(\mu_1 - \mu_0) \right\|}\right) \\
&= 1 - \Phi\left(\frac{\left\| L_C^{-1/2}(\mu_1 - \mu_0) \right\|}{2}\right)
\end{aligned}
$$

Appendix B Excess Prediction Risk

In this section we will prove Theorem 2.1. We want to find the probability bound for the out-of-sample risk, i.e. the random variable

$$E^*[\langle X^*, \beta_0 - \hat{\beta} \rangle]^2$$

where X^* is a test sample, independent of the training data, and the expectation is taken over all X^*.

The proof is adapted from Lila et al. (2021).

B.1 Setup

In the main text we used $\tilde{y}$ to denote the OS-transformed labels to distinguish it from the class indicator variable. For easier notation, throughout the proof we take Y to be the random variable for transformed labels, instead of using $\tilde{Y}$.

Our hypothesis space is the Sobolev space $\mathcal{W}^2(T)$, endowed with the norm $(\|\beta''\|^2_{\mathcal{L}^2(T)} + \epsilon\|\beta\|^2_{\mathcal{L}^2(T)})^{1/2}$. The sandwich operator T is defined as $T = L_K^{1/2} L_C L_K^{1/2}$. The operator norm of a linear operator $A : \mathcal{L}^2 \to \mathcal{L}^2$ is defined as $\|A\|_{op} = \sup_{f:\|f\|=1} \|Af\|$.

Recall our three assumptions:

Assumption 1. There exists $M_2 > 0$ such that $\|X\|_{L^2} \leq M_2$ *a.s.* We denoted $\kappa = M_2 \|L_K^{1/2}\|_{op}$ (Lila et al., 2021).

Assumption 2. There exists smooth $\beta_0 \in \mathcal{W}^2(T)$ such that $\beta_0 = L_C^{-1}(\mu_1 - \mu_0)$, i.e. the optimal function is well-defined and can be found in this space.

Assumption 3. The effective dimension of T satisfies

$$D(\lambda) = Tr((T + \lambda I)^{-1}T) = \sum_k^{\infty} \frac{\tau_k}{\tau_k + \lambda} \leq c\lambda^{-\theta}$$

for some $c, \theta > 0$.

B.2 Notation

Recall β_0 can be defined with the classification objective, $\beta_0 = L_C^{-1}(\mu_1 - \mu_0)$ which is the "standard" function. It can also be defined with the regression objective function:

$$\beta_0 = \arg\min_{\beta} E[Y - \langle X, \beta \rangle_{L^2}]^2.$$

Depending on how we define Y (the auxiliary variable), the solutions β_0 will have different scale i.e. contain different constants in their expressions. For example, Lila et al. (2021) defined $y_i = -n/n_1$ if $g_i = g_1$ and $y_i = n/n_0$ otherwise, which is different from our auxiliary variables derived in Section 2. Since the centroid classifier is invariant to changes in scale, the constants in β_0 does not matter. The solution from the regression problem (with different auxiliary variables) will still give the same discriminant function up to a constant.

Therefore, to make our subsequent calculations easier, we abuse notation and consider

$$\beta_0 = \sqrt{\pi_1 \pi_0} L_C^{-1}(\mu_1 - \mu_0).$$

The reason we introduce this constant is to simplify some algebra in later calculations.

B.3 Relationship with $\mathcal{L}^2$ as hypothesis space

Recall our objective function is

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle x_i, \beta \rangle_{\mathcal{L}^2})^2 + \lambda(\|\beta''\|_{\mathcal{L}^2}^2 + \epsilon \|\beta\|_{\mathcal{L}^2}^2)$$

The norm is given by $\|\beta\|_{\mathcal{H}} = \left(\|\beta\|_{\mathcal{L}^2}^2 + \epsilon\|\beta\|_{\mathcal{L}^2}^2\right)^{1/2}$. When $\epsilon = 0$ it is a semi-norm instead of a norm (which is what we have been using).

We have this relationship between the two function hypothesis spaces (Cucker & Smale, 2001):

$$L_K^{1/2}(\mathcal{L}^2(T)) = \mathcal{W}^2(T).$$

Hence there exist $f_0, \hat{f} \in \mathcal{L}^2$ such that

$$\beta_0 = L_K^{1/2} f_0, \ \hat{\beta} = L_K^{1/2} \hat{f}.$$

Converting to the $\mathcal{L}^2$ space facilitates the prediction error derivation (Yuan & Cai, 2010).

Assume that $\mathcal{H}$ is dense in $\mathcal{L}_2$ which ensures that f_0 and $\hat{f}$ are uniquely defined.

We can find the $\mathcal{L}^2$ counterparts of β_0:

$$\begin{aligned} f_0 &= L_K^{-1/2}\beta_0 \\ &= L_K^{-1/2}\sqrt{\pi_1\pi_0}\; L_C^{-1}(\mu_1 - \mu_0) \\ &= L_K^{-1/2}L_C^{-1}L_K^{-1/2}L_K^{1/2}(\mu_1 - \mu_0) \\ &= \sqrt{\pi_1\pi_0}\; T^{-1}L_K^{1/2}(\mu_1 - \mu_0) \end{aligned}$$

We can reformulate the objective function in $\mathcal{H}$ as an objective in $\mathcal{L}^2$:

$$\hat{f} = \underset{f\in\mathcal{L}^2}{\text{minimize}}\, \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle x_i, L_K^{1/2}f\rangle_{L^2})^2 + \lambda\|f\|_{L^2}^2$$

Since L_K is self-adjoint, $L_K^{1/2}$ is self-adjoint, so we can re-write:

$$\hat{f} = \underset{f\in\mathcal{L}^2}{\text{minimize}}\, \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle L_K^{1/2}\,x_i, f\rangle_{L^2})^2 + \lambda\|f\|_{L^2}^2$$

Notice the problem becomes exactly a ridge regression, therefore the solution is given by

$$\hat{f} = (T_n + \lambda I)^{-1}\,\frac{1}{n}\sum_{i=1}^{n} y_i\, L_K^{1/2}\,x_i.$$

B.4 Decomposition of risk into bias and variance

To facilitate the derivation, we rewrite the out-of-sample risk as a norm:

$$E^*[\langle X^*, \beta^0 - \hat{\beta}\rangle_{\mathcal{L}^2}]^2 = \|T^{1/2}(\hat{f} - f_0)\|^2$$

$$
\begin{aligned}
E^*[\langle X^*, \beta^0 - \hat{\beta}\rangle_{\mathcal{L}^2}]^2 &= E^*\left(\int X^*(t)(\beta^0(t) - \hat{\beta}(t))\, dt \right)^2 \\
&= E^* \int\int X^*(s)X^*(t) \left[\beta^0(s) - \hat{\beta}(s)\right] \left[\beta^0(t) - \hat{\beta}(t)\right] dsdt \\
&= \int\int C(s,t) \left[\beta^0(s) - \hat{\beta}(s)\right] ds \left[\beta^0(t) - \hat{\beta}(t)\right] dt \\
&= \int \left[L_C(\beta_0 - \hat{\beta})\right](t) \left[\beta^0(t) - \hat{\beta}(t)\right] dt \\
&= \langle L_C(\beta_0 - \hat{\beta}),\ (\beta_0 - \hat{\beta})\rangle \\
&= \left\langle L_C(L_K^{1/2} f_0 - L_K^{1/2}\hat{f}),\ (L_K^{1/2} f_0 - L_K^{1/2}\hat{f}) \right\rangle \\
&= \left\langle L_C L_K^{1/2}(f_0 - \hat{f}),\ L_K^{1/2}(f_0 - \hat{f}) \right\rangle \\
&= \left\langle L_K^{1/2} L_C L_K^{1/2}(f_0 - \hat{f}),\ (f_0 - \hat{f}) \right\rangle \\
&= \langle T(\hat{f} - f_0),\ (\hat{f} - f_0)\rangle \\
&= \langle T^{1/2}(\hat{f} - f_0),\ T^{1/2}(\hat{f} - f_0)\rangle \\
&= \left\| T^{1/2}(\hat{f} - f_0) \right\|^2
\end{aligned}
$$

which is a random variable since $\hat{f}$ is random. We can start decomposing the norm:

$$
\begin{aligned}
T^{1/2}(\hat{f} - f_0) &= T^{1/2}\left[(T_n + \lambda I)^{-1} \left(\frac{1}{n}\sum_{i=1}^n y_i\, L_K^{1/2}\, x_i \right) - f_0 \right] \\
&= T^{1/2}\left[(T_n + \lambda I)^{-1}\, L_K^{1/2} \left(\frac{1}{n}\sum_{i=1}^n y_i\, x_i \right) - \sqrt{\pi_1\pi_0}\; T^{-1} L_K^{1/2}(\mu_1 - \mu_0) \right] \\
&= T^{1/2}\left[(T_n + \lambda I)^{-1}\, L_K^{1/2} \left(\frac{1}{n}\sum_{i=1}^n y_i\, x_i - \sqrt{\pi_1\pi_0}(\mu_1 - \mu_0) \right) \right] \\
&+ T^{1/2}\left[(T_n + \lambda I)^{-1} L_K^{1/2} \sqrt{\pi_1\pi_0}(\mu_1 - \mu_0) - \sqrt{\pi_1\pi_0}\; T^{-1} L_K^{1/2}(\mu_1 - \mu_0) \right]
\end{aligned}
$$

Let

$$
\hat{d} = \frac{1}{n}\sum_{i=1}^n Y_i X_i,
$$

and let

$$
d = \sqrt{\pi_1\pi_0}(\mu_1 - \mu_0).
$$

Notice $E[\hat{d}] = d$, since

$$\begin{aligned}
E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i X_i\right] &= \frac{1}{n}\sum_{i=1}^{n} E[Y_i X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \pi_1 E[Y_i X_i | \text{ class } 1] + \pi_0 E[Y_i X_i | \text{ class } 0] \\
&= \frac{1}{n}\sum_{i=1}^{n} \pi_1 \sqrt{\frac{\pi_0}{\pi_1}}\mu_1 - \pi_0\sqrt{\frac{\pi_1}{\pi_0}}\mu_0 \\
&= \frac{1}{n}\sum_{i=1}^{n} \sqrt{\pi_1\pi_0}\mu_1 - \sqrt{\pi_1\pi_0}\mu_0 \\
&= \sqrt{\pi_1\pi_0}(\mu_1 - \mu_0) \\
&= d
\end{aligned}$$

We made an approximation from line 2 to 3. Recall the auxiliary variable in optimal scoring is defined as:

$$Y = \left(\sqrt{\frac{n_0}{n_1}}, -\sqrt{\frac{n_1}{n_0}}\right)^T$$

$$E[Y|\text{class } 1] = \sqrt{\frac{n_0}{n_1}} = \sqrt{\frac{\pi_0}{\pi_1}} + o(1), \quad E[Y|\text{class } 0] = -\sqrt{\frac{n_1}{n_0}} = -\sqrt{\frac{\pi_1}{\pi_0}} + o(1).$$

Note that $\sqrt{n_0/n_1}$ is an approximation of $\sqrt{\pi_0/\pi_1}$. However in the rest of the proof, we ignore the deterministic approximation error to simplify the analysis (Gaynanova 2020).

By triangle inequality, and using the new notation, we can decompose the out of sample risk into the variance term and bias term.

$$\begin{aligned}
&\|T^{1/2}(\hat{f} - f_0)\| \\
&= \left\| T^{1/2}\left[(T_n + \lambda I)^{-1} L_K^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\, x_i - \sqrt{\pi_1\pi_0}(\mu_1 - \mu_0)\right)\right]\right. \\
&\left. + T^{1/2}\left[(T_n + \lambda I)^{-1} L_K^{1/2}\sqrt{\pi_1\pi_0}(\mu_1 - \mu_0) - \sqrt{\pi_1\pi_0}\; T^{-1} L_K^{1/2}(\mu_1 - \mu_0)\right]\right\| \\
&\leq \left\| T^{1/2}(T_n + \lambda I)^{-1} L_K^{1/2}(\hat{d} - d)\right\| + \left\| T^{1/2}\left[(T_n + \lambda I)^{-1} L_K^{1/2} d - T^{-1} L_K d\right]\right\| \\
&= I_1 + I_2
\end{aligned}$$

B.5 Variance

B.5.1 Decompose variance

$$
\begin{aligned}
I_1 &= \left\| T^{1/2}(T_n + \lambda I)^{-1} L_K^{1/2}(\hat{d} - d) \right\| \\
&= \left\| T^{1/2}(T_n + \lambda I)^{-1/2}(T_n + \lambda I)^{-1/2}(T + \lambda I)^{1/2}(T + \lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d) \right\| \\
&\leq \left\| T^{1/2}(T_n + \lambda I)^{-1/2} \right\|_{op} \left\| (T_n + \lambda I)^{-1/2}(T + \lambda I)^{1/2} \right\|_{op} \left\| (T + \lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d) \right\| \\
&\leq \left\| (T + \lambda I)^{1/2}(T_n + \lambda I)^{-1/2} \right\|_{op} \left\| (T_n + \lambda I)^{-1/2}(T + \lambda I)^{1/2} \right\|_{op} \left\| (T + \lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d) \right\| \\
&= \left\| (T + \lambda I)^{1/2}(T_n + \lambda I)^{-1/2} \right\|_{op}^2 \left\| (T + \lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d) \right\| \\
&\leq \left\| (T + \lambda I)(T_n + \lambda I)^{-1} \right\|_{op} \left\| (T + \lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d) \right\|
\end{aligned}
$$

From line 2 to 3, we use the operator norm and vector norm inequality

$$\|Ax\| \leq \|A\|_{op} \|x\|.$$

From line 3 to 4, we use

$$\|T^{1/2}(T_n + \lambda I)^{-1/2}\|_{op} \leq \|(T + \lambda I)^{1/2}(T_n + \lambda I)^{-1/2}\|_{op}$$

From line 5 to 6, we have the result from (Blanchard & Kramer, 2010):

$$\forall 0 < \gamma < 1, \|A^\gamma B^\gamma\|_{op} \leq \|AB\|_{op}^\gamma.$$

Therefore for $\gamma = 1/2$,

$$\left\| (T + \lambda)^{1/2}[(T_n + \lambda I)^{-1}]^{1/2} \right\|^2 \leq \left\| (T + \lambda)(T_n + \lambda I)^{-1} \right\|^{1/2 * 2}.$$

Hence we have

$$I_1 \leq \left\| (T + \lambda I)(T_n + \lambda I)^{-1} \right\|_{op} \left\| (T + \lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d) \right\|.$$

We can bound the first term $\left\| (T + \lambda I)(T_n + \lambda I)^{-1} \right\|_{op}$ using the results from the literature.

B.5.2 Auxiliary results

Recall Assumption 1 states that, there exists $M_2 > 0$ such that $\|X\|_{L^2} \leq M_2 \, a.s.$ We defined $\kappa = M_2 \|L_K^{1/2}\|_{op}$.

Theorem B1. (Tong & Ng, 2018). Under Assumption 1, for any $0 < \delta < 1$, with confidence at least $1 - \delta$,

$$\boxed{\|(T + \lambda I)^{-1/2}(T - T_n)\|_{op} \leq B_{n,\lambda} \log(2/\delta)}$$

$$\boxed{\|(T + \lambda I)(T_n + \lambda I)^{-1}\|_{op} \leq \left(\frac{B_{n,\lambda} \log(2/\delta)}{\sqrt{\lambda}} + 1 \right)^2}$$

where

$$B_{n,\lambda} = \frac{2\kappa}{\sqrt{n}} \left(\frac{\kappa}{\sqrt{n\lambda}} + \sqrt{D(\lambda)} \right).$$

B.5.3 Variance second term

Now we will look at the second term in the variance. We want to bound $E\left[\left\| (T + \lambda I)^{-1/2} \, L_K^{1/2} (\hat{d} - d) \right\| \right]$ and use Chebyshev's inequality.

Recall $\{\tau_k, \eta_k\}$ are the eigenvalue, eigenfunction pairs of the sandwich operator T. Since $\{\eta_k\}$ forms an orthonormal basis of $\mathcal{L}^2$,

$$L_K^{1/2}(\hat{d} - d) = \sum_k \eta_k \langle L_K^{1/2}(\hat{d} - d), \, \eta_k \rangle.$$

$$E \left\| (T + \lambda I)^{-1/2} L_K^{1/2} (\hat{d} - d) \right\|^2 = E \left\| (T + \lambda I)^{-1/2} \sum_k \eta_k \left\langle L_K^{1/2}(\hat{d} - d), \eta_k \right\rangle \right\|^2$$

We want to re-write $\langle L_K^{1/2}(\hat{d} - d), \, \eta_k \rangle$ into a more manageable expression. To do that, first we look at what $E[\hat{d} - d \otimes \hat{d} - d]$ is.

$$
\begin{aligned}
&E[\hat{d} - d \otimes \hat{d} - d] \\
&= cov\bigg([\hat{d}(t) - d(t)], [\hat{d}(s) - d(s)] \bigg) \\
&= E\left[\big(\hat{d}(t) - d(t)\big) \big(\hat{d}(s) - d(s)\big) \right] \\
&= E\left[\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i X_i(t) - \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(t) \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i X_i(s) - \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(s) \right\} \right] \\
&= \pi_1 E\left[\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i X_i(t) - \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(t) \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i X_i(s) - \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(s) \right\} | class1 \right] \\
&+ \pi_0 E\left[\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i X_i(t) - \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(t) \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i X_i(s) - \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(s) \right\} | class0 \right]
\end{aligned}
$$

For class 1, by linearity of covariance,

$$
\begin{aligned}
&\pi_1 E\left[\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i X_i(t) - \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(t) \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i X_i(s) - \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(s) \right\} | class1 \right] \\
&= \pi_1 E\left[\frac{1}{n^2} \sum_{i=1}^{n} Y_i X_i(t) \sum_{i=1}^{n} Y_i X_i(s) | class1 \right] - \pi_1 E\left[\frac{1}{n} \sum_{i=1}^{n} Y_i X_i(s) \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(t) | class1 \right] \\
&- \pi_1 E\left[\frac{1}{n} \sum_{i=1}^{n} Y_i X_i(t) \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(s) | class1 \right] + \pi_1 E[\pi_1 \pi_0 (\mu_1 - \mu_0)(t)(\mu_1 - \mu_0)(s) | class1] \\
&= \pi_1 \frac{1}{n^2} \sum \frac{\pi_0}{\pi_1} C(t,s) - \pi_1 \frac{1}{n} \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(t) \sum \sqrt{\frac{\pi_0}{\pi_1}} \mu_1(s) \\
&- \pi_1 \frac{1}{n} \sqrt{\pi_1 \pi_0}(\mu_1 - \mu_0)(s) \sum \sqrt{\frac{\pi_0}{\pi_1}} \mu_1(t) + \pi_1^2 \pi_0 (\mu_1 - \mu_0)(t)(\mu_1 - \mu_0)(s) \\
&= \frac{\pi_0}{n} C(t,s) - \pi_1 \pi_0 (\mu_1 - \mu_0)(t) \mu_1(s) - \pi_1 \pi_0 (\mu_1 - \mu_0)(s) \mu_1(t) + \pi_1^2 \pi_0 (\mu_1 - \mu_0)(t)(\mu_1 - \mu_0)(s)
\end{aligned}
$$

Similarly for class 0,

$$\pi_0 E\left[\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i X_i(t) - \sqrt{\pi_1\pi_0}(\mu_1-\mu_0)(t)\right\}\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i X_i(s) - \sqrt{\pi_1\pi_0}(\mu_1-\mu_0)(s)\right\}|class0\right]$$
$$= \pi_0 E\left[\frac{1}{n^2}\sum_{i=1}^{n} Y_i X_i(t)\sum_{i=1}^{n} Y_i X_i(s)|class0\right] - \pi_0 E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i X_i(s)\sqrt{\pi_1\pi_0}(\mu_1-\mu_0)(t)|class0\right]$$
$$- \pi_0 E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i X_i(t)\sqrt{\pi_1\pi_0}(\mu_1-\mu_0)(s)|class0\right] + \pi_0 E[\pi_1\pi_0(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s)|class0]$$
$$= \pi_0\frac{1}{n^2}\sum\frac{\pi_1}{\pi_0}C(t,s) + \pi_0\frac{1}{n}\sqrt{\pi_1\pi_0}(\mu_1-\mu_0)(t)\sum\sqrt{\frac{\pi_1}{\pi_0}}\mu_0(s)$$
$$+ \pi_0\frac{1}{n}\sqrt{\pi_1\pi_0}(\mu_1-\mu_0)(s)\sum\sqrt{\frac{\pi_1}{\pi_0}}\mu_0(t) + \pi_1\pi_0^2(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s)$$
$$= \frac{\pi_1}{n}C(t,s) + \pi_1\pi_0(\mu_1-\mu_0)(t)\mu_0(s) + \pi_1\pi_0(\mu_1-\mu_0)(s)\mu_0(t) + \pi_1\pi_0^2(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s)$$

Adding the conditional expectations together, we have

$$cov\Big([\hat{d}(t)-d(t)],[\hat{d}(s)-d(s)]\Big)$$
$$= \frac{\pi_0}{n}C(t,s) - \pi_1\pi_0(\mu_1-\mu_0)(t)\mu_1(s) - \pi_1\pi_0(\mu_1-\mu_0)(s)\mu_1(t) + \pi_1^2\pi_0(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s)$$
$$+ \frac{\pi_1}{n}C(t,s) + \pi_1\pi_0(\mu_1-\mu_0)(t)\mu_0(s) + \pi_1\pi_0(\mu_1-\mu_0)(s)\mu_0(t) + \pi_1\pi_0^2(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s)$$
$$= \frac{1}{n}C(t,s) - \pi_1\pi_0(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s) - \pi_1\pi_0(\mu_1-\mu_0)(s)(\mu_1-\mu_0)(t)$$
$$+ (\pi_1^2\pi_0 + \pi_1\pi_0^2)(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s)$$
$$= \frac{1}{n}C(t,s) - 2\pi_1\pi_0(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s) + \pi_1\pi_0(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s)(\pi_1+\pi_0)$$
$$= \frac{1}{n}C(t,s) - \pi_1\pi_0\,(\mu_1-\mu_0)(t)(\mu_1-\mu_0)(s)$$
$$= \frac{1}{n}C(t,s) - d(t)d(s)$$

We define

$$L_{E[(\hat{d}-d)\otimes(\hat{d}-d)]}f(\cdot) = \int(\hat{d}-d)(\cdot)(\hat{d}-d)(s)f(s)\,ds,$$

$$L_{\mu_1\otimes\mu_1}f(\cdot) = \int\mu_1(\cdot)\mu_1(s)f(s)\,ds.$$

Now that we have an expression for $E[(\hat{d}-d)\otimes(\hat{d}-d)]$, we can work on the expectation:

$$
\begin{aligned}
&E\left[\left\langle L_K^{1/2}(\hat{d}-d),\, \eta_k\right\rangle^2\right] \\
&= E\left[\int L_K^{1/2}(\hat{d}-d)(t)\eta_k(t)dt \int L_K^{1/2}(\hat{d}-d)(s)\eta_k(s)ds\right] \\
&= E\left[\int\int L_K^{1/2}(\hat{d}-d)(t)\eta_k(t)\; L_K^{1/2}(\hat{d}-d)(s)\eta_k(s)\, dtds\right] \\
&= E\left[\int L_K^{1/2}\left[\int (\hat{d}-d)(t)(\hat{d}-d)(s)L_K^{1/2}\eta_k(t)\, dt\right]\, \eta_k(s)\, ds\right] \\
&= \int L_K^{1/2}\left[L_{E[(\hat{d}-d)\otimes(\hat{d}-d)]}L_K^{1/2}\eta_k(s)\right]\eta_k(s)\, ds \\
&= \left\langle L_K^{1/2}\, L_{E[(\hat{d}-d)\otimes(\hat{d}-d)]}\, L_K^{1/2}\, \eta_k,\, \eta_k\right\rangle \\
&= \langle L_k^{1/2}L_{\frac{1}{n}C}L_k^{1/2}\eta_k, \eta_k\rangle - \langle L_K^{1/2}L_{d\otimes d}L_K^{1/2}\eta_k, \eta_k\rangle \\
&= \frac{1}{n}\langle T\eta_k, \eta_k\rangle - \langle L_K^{1/2}L_{d\otimes d}L_K^{1/2}\eta_k, \eta_k\rangle \\
&= \frac{1}{n}\langle \tau_k\eta_k, \eta_k\rangle - \int L_K^{1/2}L_{d\otimes d}L_k^{1/2}\eta_k(t)\eta_k(t)\, dt \\
&= \frac{1}{n}\tau_k - \int\left[\int L_K^{1/2}d(t)d(s)L_K^{1/2}\eta_k(s)\, ds\right]\, \eta_k(t)\, dt \\
&= \frac{1}{n}\tau_k - \langle L_K^{1/2}d, \eta_k\rangle\langle L_K^{1/2}d, \eta_k\rangle
\end{aligned}
$$

Recall $\{\tau_k, \eta_k\}$ are the eigenvalues, eigenfunction pairs of sandwich operator T. $\{\eta_k\}$ forms an orthonormal basis of $\mathcal{L}^2$.

$$
\begin{aligned}
&E\left\|(T+\lambda I)^{-1/2}L_K^{1/2}(\hat{d}-d)\right\|^2 \\
&= \left\|(T+\lambda I)^{-1/2}\sum_k \eta_k\left\langle L_K^{1/2}(\hat{d}-d), \eta_k\right\rangle\right\|^2 \\
&= \left\|\sum_k (T+\lambda I)^{-1/2}\,\eta_k\left[\frac{\tau_k}{n} - \langle L_K^{1/2}d, \eta_k\rangle^2\right]^{1/2}\right\|^2 \\
&\leq \sum_k \frac{1}{n}\left(\frac{\tau_k}{\tau_k+\lambda}\right) - \frac{\|L_K^{1/2}d\|^2}{\tau_k+\lambda} \\
&\leq \frac{1}{n}\sum_k \frac{\tau_k}{\tau_k+\lambda} \\
&= \frac{1}{n}D(\lambda)
\end{aligned}
$$

Recall in assumption 3, the effective dimension of T satisfies

$$D(\lambda) = Tr((T+\lambda I)^{-1}T) = \sum_{k}^{\infty} \frac{\tau_k}{\tau_k + \lambda} \leq c\lambda^{-\theta}$$

for some $c, \theta > 0$.

With this bound, we can apply Chebyshev's inequality:

$$P(|X - \mu| \geq a\sigma) \leq \frac{1}{a^2}.$$

Let $a^2 = 2/\delta$.

$$P(|X - \mu| \leq a\sigma) \geq 1 - \frac{1}{a^2} = 1 - \frac{\delta}{2},$$

$$P\left(\left\|(T+\lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d)\right\| \leq \sqrt{\frac{2}{\delta}}\sigma\right) \geq 1 - \frac{\delta}{2}.$$

Since $\sigma \leq \sqrt{\frac{D(\lambda)}{n}}$,

$$P\left(\left\|(T+\lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d)\right\| \leq \sqrt{\frac{2}{\delta}}\sqrt{\frac{D(\lambda)}{n}}\right) \geq 1 - \frac{\delta}{2}.$$

Therefore with probability at least $1 - \delta/2$,

$$\left\|(T+\lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d)\right\| \leq \sqrt{\frac{2}{\delta}}\sqrt{\frac{D(\lambda)}{n}} \leq \frac{1}{\kappa\delta} B_{n,\lambda},$$

where

$$B_{n,\lambda} = \frac{2\kappa}{\sqrt{n}}\left(\frac{\kappa}{\sqrt{n\lambda}} + \sqrt{D(\lambda)}\right), \quad \kappa = M_2 \|L_K^{1/2}\|_{op}.$$

B.5.4 Combining two variance terms

Recall the results from (Tong & Ng, 2018):

$$\|(T+\lambda I)(T_n + \lambda I)^{-1}\|_{op} \leq \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}} + 1\right)^2$$

with confidence at least $1 - \delta$.

Putting the two terms together, with probability at least $1 - \delta$, we have

$$\begin{aligned} I_1 &= \|T^{1/2}(T_n + \lambda I)^{-1} L_K^{1/2}(\hat{d} - d)\| \\ &\leq \left\|(T+\lambda I)(T_n + \lambda I)^{-1}\right\|_{op} \left\|(T+\lambda I)^{-1/2} L_K^{1/2}(\hat{d} - d)\right\| \\ &\leq \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}} + 1\right)^2 \left(\frac{1}{2\kappa\delta} B_{n,\lambda}\right) \end{aligned}$$

We have a bound for the variance term:

$$\boxed{I_1 \leq \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}} + 1\right)^2 \left(\frac{1}{2\kappa\delta} B_{n,\lambda}\right)}$$

B.6 Bias

Next we move on to the bias term.

$$\begin{aligned}
I_2 &= \left\| T^{1/2}\left[(T_n+\lambda I)^{-1} L_K^{1/2} d - T^{-1} L_K d\right] \right\| \\
&= \left\| T^{1/2}\left[(T_n+\lambda I)^{-1} T\, T^{-1} L_K^{1/2} d - T^{-1} L_K d\right] \right\| \\
&= \left\| T^{1/2}\left[(T_n+\lambda I)^{-1} T - I\right](T^{-1} L_K d) \right\| \\
&= \left\| T^{1/2}\Big((T_n+\lambda I)^{-1} T - (T_n+\lambda I)^{-1}(T_n+\lambda I)\Big)(T^{-1} L_K d) \right\| \\
&= \left\| T^{1/2}\Big((T_n+\lambda I)^{-1}(T - T_n - \lambda I\Big)(T^{-1} L_K d) \right\| \\
&\leq \left\| T^{1/2}(T_n+\lambda I)^{-1}[T - T_n - \lambda I] \right\|_{op} \|T^{-1} L_K d\|
\end{aligned}$$

Now we can apply the same steps as I_1.

$$\begin{aligned}
&= \left\| T^{1/2}(T_n+\lambda I)^{-1/2}(T_n+\lambda I)^{-1/2}(T+\lambda I)^{1/2}(T+\lambda I)^{-1/2}[T - T_n - \lambda I] \right\|_{op} \|f_0\| \\
&\leq \left\| T^{1/2}(T_n+\lambda I)^{-1/2} \right\|_{op} \left\| (T_n+\lambda I)^{-1/2}(T+\lambda I)^{1/2} \right\|_{op} \left\| (T+\lambda I)^{-1/2}[T - T_n - \lambda I] \right\|_{op} \|f_0\| \\
&\leq \left\| (T+\lambda I)^{1/2}(T_n+\lambda I)^{-1/2} \right\|_{op} \|(T_n+\lambda I)^{-1/2}(T+\lambda I)^{1/2} \right\|_{op} \left\| (T+\lambda I)^{-1/2}[T - T_n - \lambda I] \right\|_{op} \|f_0\| \\
&\leq \left\| (T+\lambda I)(T_n+\lambda I)^{-1} \right\|_{op} \left\| (T+\lambda I)^{-1/2}[T - T_n - \lambda I] \right\|_{op} \|f_0\| \\
&\leq \left\| (T+\lambda I)(T_n+\lambda I)^{-1} \right\|_{op} \left\| (T+\lambda I)^{-1/2}[T - T_n] \right\|_{op} \|f_0\| \\
&+ \left\| (T+\lambda I)(T_n+\lambda I)^{-1} \right\|_{op} \lambda \left\| (T+\lambda I)^{-1/2} \right\|_{op} \|f_0\|
\end{aligned}$$

Combining the auxiliary results from (Tong & Ng, 2018), which state that

$$\|(T+\lambda I)^{-1/2}(T - T_n)\|_{op} \leq B_{n,\lambda}\log(2/\delta),$$

$$\|(T+\lambda I)(T_n+\lambda I)^{-1}\|_{op} \leq \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^2,$$

where the inequalities hold with confidence at least $1-\delta$ under Assumption 1. In addition, we have

$$\lambda\left\|(T+\lambda I)^{-1/2}\right\|_{op} \leq \frac{\lambda}{\sqrt{\lambda}} = \sqrt{\lambda}.$$

Putting everything together, we have that with probability at least $1-\delta$,

$$\begin{aligned}
I_2 &= \left\| T^{1/2}\left[(T_n+\lambda I)^{-1}L_K^{1/2}d - T^{-1}L_K d\right]\right\| \\
&\leq \left\|(T+\lambda I)(T_n+\lambda I)^{-1}\right\|_{op}\left\|(T+\lambda I)^{-1/2}[T-T_n]\right\|_{op}\|f_0\| \\
&+ \left\|(T+\lambda I)(T_n+\lambda I)^{-1}\right\|_{op}\lambda\left\|(T+\lambda I)^{-1/2}\right\|_{op}\|f_0\| \\
&\leq \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^2\left(B_{n,\lambda}\log(2/\delta)+\sqrt{\lambda}\right)\|f_0\|
\end{aligned}$$

We have a bound for the bias term:

$$\boxed{I_2 \leq \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^2\left(B_{n,\lambda}\log(2/\delta)+\sqrt{\lambda}\right)\|f_0\|}$$

B.7 Out-of-sample risk

Recall we re-wrote the expression for the out-of-sample risk and decomposed into the bias and variance terms:

$$E^*[\langle X^*, \beta^0-\hat{\beta}\rangle_{\mathcal{L}^2}]^2 = \|T^{1/2}(\hat{f}_\lambda - f_0)\|^2$$

$$\begin{aligned}
\|T^{1/2}(\hat{f}_\lambda - f_0)\| &\leq \left\|T^{1/2}(T_n+\lambda I)^{-1}\,L_K^{1/2}(\hat{d}-d)\right\| + \left\|T^{1/2}\left[(T_n+\lambda I)^{-1}L_K^{1/2}d - T^{-1}L_K d\right]\right\| \\
&= I_1 + I_2
\end{aligned}$$

$$
\begin{aligned}
&\left\|T^{1/2}(\hat{f}-f_0)\right\|^2 \\
&\leq 2\left\|T^{1/2}(T_n+\lambda I)^{-1}\,L_K^{1/2}(\hat{d}-d)\right\|^2 + 2\left\|T^{1/2}\left[(T_n+\lambda I)^{-1}L_K^{1/2}d - T^{-1}L_K d\right]\right\|^2 \\
&\leq 2\left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^4\left(\frac{1}{2\kappa\delta}B_{n,\lambda}\right)^2 \\
&+ 2\left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^4\left(B_{n,\lambda}\log(2/\delta)+\sqrt{\lambda}\right)^2\|f_0\|^2 \\
&= 2\frac{\lambda}{\delta^2}\left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^4\left(\frac{B_{n,\lambda}}{2\kappa\sqrt{\lambda}}\right)^2 \\
&+ 2\lambda\left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^4\left(\frac{B_{n,\lambda}}{\sqrt{\lambda}}\log(2/\delta)+1\right)^2\|f_0\|^2 \\
&= 2\frac{\lambda}{\delta^2}\left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^4\left(\frac{B_{n,\lambda}}{2\kappa\sqrt{\lambda}}\right)^2 + 2\lambda\left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^6\|f_0\|^2 \\
&\leq C\frac{(\log(2/\delta))^6}{\delta^2}\,n^{-\frac{1}{1+\theta}}
\end{aligned}
$$

In the last inequality, we choose $\lambda = n^{-\frac{1}{1+\theta}}$. Recall in Assumption 3. we assumed that the effective dimension of T satisfies $D(\lambda) \leq c\lambda^{-\theta}$ for some $c, \theta > 0$.

$$
B_{n,\lambda} = \frac{2\kappa}{\sqrt{n}}\left(\frac{\kappa}{\sqrt{n\lambda}}+\sqrt{D(\lambda)}\right) \leq 2\kappa\left(\frac{\kappa}{n\sqrt{\lambda}}+\sqrt{\frac{c}{n\lambda^{\theta}}}\right) = 2\kappa\sqrt{\lambda}\left(\frac{\kappa}{n\lambda}+\sqrt{\frac{c}{n\lambda^{\theta+1}}}\right).
$$

Looking at the denominators:

$$
n\lambda = n^{-\frac{1}{1+\theta}}n = n^{\frac{1+\theta-1}{1+\theta}} = n^{\frac{\theta}{1+\theta}} \geq 1,
$$

$$
\lambda^{\frac{\theta+1}{2}}n^{1/2} = n^{-\frac{1}{1+\theta}\frac{\theta+1}{2}}n^{1/2} = n^{-1/2}n^{1/2} = 1,
$$

we have this inequality for $B_{n,\lambda}$:

$$
B_{n,\lambda} \leq 2\kappa\sqrt{\lambda}(\kappa+\sqrt{c}).
$$

Hence

$$
\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}} \leq 2\kappa(\kappa+\sqrt{c}).
$$

All the terms involving $B_{n,\lambda}$ and $\|f_0\|^2$ can be absorbed into the constant C.

Finally, we have the bound for the out-of-sample risk:

$$\boxed{E^*[\langle X^*, \beta^0 - \hat{\beta}\rangle_{\mathcal{L}^2}]^2 \leq C\frac{(\log(2/\delta))^6}{\delta^2}\, n^{-\frac{1}{1+\theta}}}$$

B.8 Auxiliary results

In Appendix B5.2, we cited the results from (Tong & Ng, 2018) that helped us bound the first term in the variance. We introduced them in the middle of the proof to help the flow of the arguments. We provide more context here.

Theorem B.2. Let $\mathcal{H}$ be a Hilbert space endowed with a norm $\|\|_{\mathcal{H}}$ and let X be a random variable taking values in $\mathcal{H}$. Let $X_1, ..., X_n$ be a sequence of n independent copies of X. Assume that $\|X\|_{\mathcal{H}} \leq M(a.s.)$, then for $0 < \delta < 1$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right\|_{\mathcal{H}}^2 \leq \frac{2M\log(2/\delta)}{n} + \sqrt{\frac{2E[\|X\|_{\mathcal{H}}^2\log(2/\delta)]}{n}}$$

this inequality holds with probability at least $1-\delta$.

Using this theorem, Tong & Ng derived the two results that we cited:

Under Assumption 1, for any $0 < \delta < 1$, with confidence at least $1-\delta$,

$$\|(T+\lambda I)^{-1/2}(T - T_n)\|_{op} \leq B_{n,\lambda}\log(2/\alpha)$$

$$\|(T+\lambda I)(T_n+\lambda I)^{-1}\|_{op} \leq \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}} + 1\right)^2$$

$$B_{n,\lambda} = \frac{2\kappa}{\sqrt{n}}\left(\frac{\kappa}{\sqrt{n\lambda}} + \sqrt{D(\lambda)}\right)$$

Moreover, their confidence set are the same.

Appendix C Simulation study results

Here we provide additional results for the univariate classifier simulation study. Recall that for our simulation study, data from the two classes were generated as:

$$X_1(t) = \mu_1(t) + \sum_{k=1}^{20} Z_k\phi_k(t), \quad X_2(t) = \mu_0(t) + \sum_{k=1}^{20} Z_k\phi_k(t)$$

where

$$\mu_1(t) = \sum_{k=1}^{20} a_k\phi_k(t), \quad \mu_0(t) = \sum_{k=1}^{20}(-a_k)\phi_k(t)$$

for some coefficients $a_1, ..., a_{20} \in \mathbb{R}$.

The coefficients are Normal random variables:

$$Z_k \sim N(0, \sigma_k^2)$$

Therefore, the data is completely characterized by Z_k and a_k. We generated dataset for $a_k = k^{-u}$ and $\sigma_k = k^{-u}$ where $u = 1, 1.5, 2, 2.5, 3$. For each dataset, we went through the training-validation-test procedure to determine the best estimator in terms of AUC or estimation error. This analysis also shows we should take the estimation error $\|\beta_0 - \hat{\beta}\|$ with a grain of salt, since β_0 is too easily swayed by small covariance (i.e. fast decaying σ_k).

AUC					
$d_k \setminus \sigma_k$	1	1.5	2	2.5	3
1	1.000	1.000	1.000	1.000	1.000
1.5	0.966	1.000	1.000	1.000	1.000
2	0.919	0.987	0.999	1.000	1.000
2.5	0.897	0.912	0.973	0.987	1.000
3	0.922	0.942	0.917	0.947	0.991

Table 1: Estimator with exponential kernel

AUC					
$d_k \setminus \sigma_k$	1	1.5	2	2.5	3
1	1.000	1.000	1.000	1.000	1.000
1.5	0.962	0.997	1.000	1.000	1.000
2	0.939	0.917	0.968	0.999	1.000
2.5	0.900	0.931	0.941	0.984	0.999
3	0.938	0.927	0.951	0.949	0.971

Table 2: Estimator with Sobolev space

AUC					
$d_k \setminus \sigma_k$	1	1.5	2	2.5	3
1	1.000	1.000	1.000	1.000	1.000
1.5	0.926	1.000	1.000	1.000	1.000
2	0.926	0.902	0.994	1.000	1.000
2.5	0.944	0.920	0.937	0.994	1.000
3	0.922	0.912	0.913	0.947	0.982

Table 3: (Delaigle & Hall) Truncation estimator

Estimation error					
$d_k \setminus \sigma_k$	1	1.5	2	2.5	3
1	12067	8.916e+08	9.865e+13	1.259e+19	1.727e+24
1.5	85.195	3.000e+06	2.890e+11	3.482e+16	4.631e+21
2	6.529	1.207e+04	8.916e+08	9.865e+13	1.259e+19
2.5	4.349	8.520e+01	2.999e+06	2.890e+11	3.482232e+16
3	4.076	6.529e+00	1.206e+04	8.916e+08	9.865e+13

Table 4: Estimator with exponential kernel

Estimation error					
$d_k \setminus \sigma_k$	1	1.5	2	2.5	3
1	12080.313	8.916e+08	9.865e+13	1.259e+19	1.727e+24
1.5	89.346	3.000e+06	2.890e+11	3.482e+16	4.631e+21
2	9.142	1.209e+04	8.916e+08	9.865e+13	1.259e+19
2.5	6.849	9.060e+01	3.000e+06	2.890e+11	3.482e+16
3	6.379	8.388e+00	1.210e+04	8.916e+08	9.865e+13

Table 5: Estimator with Sobolev space

Estimation error					
$d_k \setminus \sigma_k$	1	1.5	2	2.5	3
1	12066.905	8.916e+08	9.864e+13	1.259e+19	1.727e+24
1.5	85.195	3.000e+06	2.890e+11	3.482e+16	4.631e+21
2	6.529	1.207e+04	8.915e+08	9.864e+13	1.259e+19
2.5	4.348	8.512e+01	2.994e+06	2.884e+11	3.470e+16
3	4.076	6.529e+00	1.191e+04	8.634e+08	9.495e+13

Table 6: (Delaigle & Hall) Truncation estimator

8 References

Berrendero, J. R., Cuevas, A., & Torrecilla, J. L. (2018). On the use of reproducing Kernel Hilbert spaces in functional classification. *Journal of the American Statistical Association*, 113(523), 1210–1218. https://doi.org/10.1080/01621459.2017.1320287

Blanchard, G. & Kramer, N. (2010). Optimal learning rates for kernel conjugate gardient regression. *Advances in Neural Information Processing Systems 23:24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010.* https://doi.org/10.48-550/arXiv.1009.5839

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

Cai, T. T., & Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. Journal of the American Statistical Association, 107(499), 1201–1216. https://doi.org/10.1080/01621459.2012.716337

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1), 1–49. https://doi.org/10.1090/s0273-0979-01-00923-5

Delaigle, A., & Hall, P. (2011). Achieving near perfect classification for Functional Data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2), 267–286. https://doi.org/10.1111/j.1467-9868.2011.01003.x

Ferrando, L., Ventura-Campos, N., & Epifanio, I. (2020). Detecting and visualizing differences in brain structures with SPHARM and functional data analysis. NeuroImage, 222, 117209. https://doi.org/10.1016/j.neuroimage.2020.117209

Fieremans, E., Jensen, J. H., & Helpern, J. A. (2011). White matter characterization with diffusional kurtosis imaging. NeuroImage, 58(1), 177–188. https://doi.org/10.1016/j.neuroimage.2011.06.006

Gaynanova, I. (2020). Prediction and estimation consistency of sparse multi-class penalized optimal scoring. *Bernoulli*, 26(1). https://doi.org/10.3150/19-bej1126

Ghiassian, S., Greiner, R., Jin, P., & Brown, M. R. (2016). Using functional or structural magnetic resonance images and personal characteristic data to identify ADHD and autism. PLOS ONE, 11(12). https://doi.org/10.1371/journal.pone.0166934

Grenander, U. (1950). Stochastic processes and statistical inference. Arkiv För Matematik, 1(3), 195–277. https://doi.org/10.1007/bf02590638

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 23(1). https://doi.org/10.1214/aos/1176324456

HCP-ya data dictionary- updated for the 1200 subject release. HCP wiki. (n.d.). Retrieved June 4, 2022, from https://wiki.humanconnectome.org/display/PublicData/HCP-YA+Data+Dictionary-+Updated+for+the+1200+Subject+Release

Isotropic vs Anisotropic. Questions and Answers in MRI. (n.d.). Retrieved July 20, 2022, from https://mriquestions.com/iso-anisotropic-diffusion.html

Jensen, J. H., & Helpern, J. A. (2010). MRI quantification of non-Gaussian Water Diffusion by kurtosis analysis. NMR in Biomedicine, 23(7), 698–710. https://doi.org/10.1002/nbm.1518

Kraus, D., & Stefanucci, M. (2018). Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika*, 106(1), 161–180. https://doi.org/10.1093/biomet/asy060

Lila, E., Zhang, W., & Rane, S. (2021). Interpretable discriminant analysis for functional data supported on random non-linear domains. *arXiv* preprint arXiv:2112.02712.

O'Donnell, L. J., & Westin, C.-F. (2011). An introduction to diffusion tensor image analysis. Neurosurgery Clinics of North America, 22(2), 185–196. https://doi.org/10.1016/j.nec.2010.12.004

Ramsay, J. O. (1982). When the data are functions. Psychometrika, 47(4), 379–396. https://doi.org/10.1007/bf02293704

Ramsay, J. O., & Dalzell, C. J. (1991). Some tools for functional data analysis. Journal of the Royal Statistical Society: Series B (Methodological), 53(3), 539–561. https://doi.org/10.1111/j.2517-6161.1991.tb01844.x

Ramsay, J. O., & Silverman, B. W. (2010). Functional Data Analysis. Springer.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. Biometrics, 14(1), 1. https://doi.org/10.2307/2527726

Park, J., Ahn, J., & Jeon, Y. (2021). Sparse functional linear discriminant analysis. *Biometrika*, 109(1), 209–226. https://doi.org/10.1093/biomet/asaa107

Tong, H., & Ng, M. (2018). Analysis of regularized least squares for functional linear regression model. *Journal of Complexity*, 49, 85–94. https://doi.org/10.1016/j.jco.2018.08.001

Yuan, M., & Cai, T. T. (2010). A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6). https://doi.org/10.1214/09-aos772