Interpretable discriminant analysis for functional data supported on random nonlinear domains with an application to Alzheimer’s disease

Eardi Lila¹, Wenbo Zhang¹,², Swati Rane Levendovszky³ and Alzheimer’s Disease Neuroimaging Initiative†

¹Department of Biostatistics, University of Washington, Seattle, USA  
²Department of Statistics, University of California, Irvine, USA  
³Department of Radiology, University of Washington, Seattle, USA  

Address for correspondence: Eardi Lila, Department of Biostatistics, University of Washington, 3980 15th Ave NE, Seattle, WA 98105, USA. Email: elila@uw.edu

Abstract

We introduce a novel framework for the classification of functional data supported on nonlinear, and possibly random, manifold domains. The motivating application is the identification of subjects with Alzheimer’s disease from their cortical surface geometry and associated cortical thickness map. The proposed model is based upon a reformulation of the classification problem as a regularized multivariate functional linear regression model. This allows us to adopt a direct approach to the estimation of the most discriminant direction while controlling for its complexity with appropriate differential regularization. Our approach does not require prior estimation of the covariance structure of the functional predictors, which is computationally prohibitive in our application setting. We provide a theoretical analysis of the out-of-sample prediction error of the proposed model and explore the finite sample performance in a simulation setting. We apply the proposed method to a pooled dataset from Alzheimer’s Disease Neuroimaging Initiative and Parkinson’s Progression Markers Initiative. Through this application, we identify discriminant directions that capture both cortical geometric and thickness predictive features of Alzheimer’s disease that are consistent with the existing neuroscience literature.

Keywords: functional classification, manifold data analysis, shape data analysis, neuroimaging

1 Introduction

Functional discriminant analysis, a statistical framework used to predict categorical outcomes from functional predictors, has been extensively studied within the field of Functional Data Analysis (FDA) (Ferraty & Vieu, 2006; Horváth & Kokoszka, 2012; Hsing & Eubank, 2013; Ramsay & Silverman, 2015) and has motivated a large body of literature (see, e.g. Berrendero et al., 2018; Dai et al., 2017; Delaigle & Hall, 2012; James & Hastie, 2001; Kraus & Stefanucci, 2019; Müller, 2005; Park et al., 2021; Preda, 2007). However, most of the existing methods are concerned with the classification of functions supported on one-dimensional linear
domains, which can be a limiting assumption in many modern biomedical applications (Zhu et al., 2023). On the other hand, recent work on the analysis of functional data with manifold structure has mostly focused on vector-valued functions with non-Euclidean constraints in the image space (see e.g. Dai & Müller, 2018; Dubey & Müller, 2020; Kim et al., 2021; Lin et al., 2017; Su et al., 2014).

In this paper, motivated by the analysis of modern multi-modal imaging data, we propose a novel functional discriminant analysis model that can handle functional predictors supported on non-linear sample-specific manifold domains, which we term Functions on Surfaces (FoSs) (Lila & Aston, 2020). An example of such ‘object data’ (Marron & Dryden, 2021) is provided in Figure 1a, illustrating our motivating application of identifying subjects with Alzheimer’s disease from FoSs that are subject-specific cortical surfaces coupled with cortical thickness measurements. The statistical analysis of these object data poses unique statistical challenges. This is mainly due to the non-Euclidean structure of each individual domain, which makes it difficult to define appropriate spatial regularization, and due to the more abstract non-Euclidean structure of the latent space where the random domains are supported, which further invalidates traditional linear statistical models (Dryden & Mardia, 2016; Grenander & Miller, 1998; Kendall, 1984; Younes, 2019).

Current approaches in the literature do not comprehensively address these challenges. Although various methods have been proposed to model functional data on multi-dimensional domains, they often focus on flat domains (Feng et al., 2020; Goldsmith et al., 2014; Kang et al., 2018; Wang & Zhu, 2017; Yu et al., 2021) or assume nonlinear but fixed domains (Chung et al., 2016, 2015; Lila et al., 2016; Mejia et al., 2020).

There has also been considerable work on the simpler setting of random surfaces that are not coupled with functional data. These efforts can be broadly grouped into three main approaches. The first approach leverages global parametrizations to represent surfaces, employing either an $L^2$ metric (Chung et al., 2008; Epifanio & Ventura-Campos, 2014; Ferrando et al., 2020) or a non-Euclidean metric (Jermyn et al., 2012, 2017; Kurtek & Drira, 2015; Zhang et al., 2022). The second approach, which is more closely related to the one adopted in this work, uses diffeomorphic deformation functions of the surfaces’ embedding space (Arguillére et al., 2016; Vaillant et al., 2004; Younes, 2019), allowing for the inclusion of topological constraints. The third approach, prevalently used in neuroimaging studies, employs pre-specified or spectrum-based descriptors of shape (Dong et al., 2020, 2019; Hazlett et al., 2017; Im et al., 2008; Reuter et al., 2006; Wachinger et al., 2015; Wang & Wang, 2017). A critical drawback of the latter approach is the inability to uniquely map the discrete representations back to the original space of random surfaces.

The statistical analysis of random surfaces that are coupled with functional data has not been extensively explored. One exception is the model in Zaetz and Kurtek (2015) which focuses on annotated surfaces. In addition, unsupervised models have been investigated by Charlier et al. (2017) and Lila and Aston (2020). Lee et al. (2017) have dealt with the classification problem by employing the fshape framework (Charlier et al., 2017) to represent the data and by using a linear discriminant model on the resulting finite-dimensional representations. In contrast, the statistical framework for the discriminant analysis of FoSs presented in this paper avoids any dimension reduction of the predictors and instead employs spatial penalties to regularize the discriminant direction. It tackles the non-Euclidean nature of the latent space of random domains by defining appropriate linear functional representations of FoSs, effectively reframing the problem of classifying FoSs as the problem of classifying bivariate functional data supported on general, but fixed, domains. A direct model for the estimation of a functional discriminant direction is then defined on the representation space where differential spatial regularizations are introduced to produce interpretable and well-defined estimates. A key feature of the proposed representation is its invertibility, which enables us to map estimates from the representation space back to the original space. This allows us to explore and interpret the estimated classification rule in the context of the original neurobiological objects of our motivating application.

The rest of the paper is organized as follows. In Section 2, we describe the representation model adopted to parametrize the non-Euclidean space of FoSs using linear function spaces. In Sections 3 and 4, we develop a novel discriminant analysis model on the parametrizing linear function spaces and provide theoretical guarantees for the prediction performance of the proposed model. We introduce an efficient discretization approach in Section 5 and apply the proposed model to the
Alzheimer’s Disease Neuroimaging Initiative (ADNI) and Parkinson’s Progression Markers Initiative (PPMI) datasets in Section 6. Proofs and simulations are left to the appendices.

2 Functional data supported on general random domains

The data considered in this work are a sample of triplets

\[ \{(g_i, \mathcal{M}_i, z_i), i = 1, \ldots, n\}, \tag{1} \]

where \( g_i \) is a binary label, \( \mathcal{M}_i \subset \mathbb{R}^3 \) is a sample-specific closed two-dimensional manifold embedded in \( \mathbb{R}^3 \), and \( z_i : \mathcal{M}_i \to \mathbb{R} \) is a scalar function supported on the geometric object \( \mathcal{M}_i \). We moreover assume that the points on the geometries \( \{\mathcal{M}_i\} \) of the observed FosSs are in one-to-one correspondence across subjects. We refer to the pairs \( \{(\mathcal{M}_i, z_i)\} \) as FosSs.

In Figure 1a, we display three observations of the training sample of our final application, where \( g_i \) encodes the disease state of the subject, \( \mathcal{M}_i \) encodes the geometry of the cerebral cortex, and \( z_i : \mathcal{M}_i \to \mathbb{R} \) is a real function, supported on \( \mathcal{M}_i \), describing cortical thickness (in mm). Panel b: Linear representation \((v_i, x_i)\) of each FosS \((\mathcal{M}_i, z_i)\) shown in Panel a. Here \( v_i : \mathbb{R}^3 \to \mathbb{R}^3 \) is a vector-valued function encoding the geometry of the \( i \)th individual. This is depicted as a collection of 3D vectors \((v_i(p_j))\) for a dense set of points \( \{p_j\} \subset \mathbb{R}^3 \). For clarity, the function \( v_i \) is displayed only on half of its domain \( \mathbb{R}^3 \). The function \( x_i : \mathcal{M} \to \mathbb{R} \) describes the spatially normalized thickness map of the \( i \)th individual on the fixed template \( \mathcal{M} \). Panel c: FosS \((\varphi_{v_1}(\mathcal{M}), x_1 \circ \varphi_{v_1}^{-1})\) parametrized by the associated functions \((v_i, x_i)\) in Panel b. This is a close approximation of the FosS \((\mathcal{M}_i, z_i)\) in Panel a.

Alzheimer’s Disease Neuroimaging Initiative (ADNI) and Parkinson’s Progression Markers Initiative (PPMI) datasets in Section 6. Proofs and simulations are left to the appendices.

Figure 1. Panel a: FosSs of three subjects in the training sample, where \( g_i \in \{\text{‘C’, ‘AD’}\} \) denotes the disease state of the \( i \)th individual (C, Control; AD, Alzheimer’s Disease), \( \mathcal{M}_i \) is a two-dimensional manifold encoding the geometry of the cerebral cortex, and \( z_i : \mathcal{M}_i \to \mathbb{R} \) is a real function, supported on \( \mathcal{M}_i \), describing cortical thickness (in mm). Panel b: Linear representation \((v_i, x_i)\) of each FosS \((\mathcal{M}_i, z_i)\) shown in Panel a. Here \( v_i : \mathbb{R}^3 \to \mathbb{R}^3 \) is a vector-valued function encoding the geometry of the \( i \)th individual. This is depicted as a collection of 3D vectors \((v_i(p_j))\) for a dense set of points \( \{p_j\} \subset \mathbb{R}^3 \). For clarity, the function \( v_i \) is displayed only on half of its domain \( \mathbb{R}^3 \). The function \( x_i : \mathcal{M} \to \mathbb{R} \) describes the spatially normalized thickness map of the \( i \)th individual on the fixed template \( \mathcal{M} \). Panel c: FosS \((\varphi_{v_1}(\mathcal{M}), x_1 \circ \varphi_{v_1}^{-1})\) parametrized by the associated functions \((v_i, x_i)\) in Panel b. This is a close approximation of the FosS \((\mathcal{M}_i, z_i)\) in Panel a.
cortex, and $z_i$ encodes the cortical thickness map supported on $M_i$. Our goal is to build a classifier from the given training sample that can predict the binary label $g^*$ of a previously unseen FoS $(M^*, z^*)$.

2.1 Linear functional representations

In our motivating application, the geometric objects \( \{M_i : i = 1, \ldots, n\} \) are topologically equivalent to a sphere, and therefore, they do not display holes or self-intersections. To inform our model of such physical non-Euclidean constraints, we define a convenient unconstrained representation model for the FoSs \( \{(M_i, z_i) : i = 1, \ldots, n\} \) in terms of objects belonging to linear function spaces.

Let \( M \) be a template two-dimensional manifold embedded in \( \mathbb{R}^3 \) that is topologically equivalent to a sphere. We denote by \( L^2(M) \) the space of square integrable functions over \( M \), equipped with the standard inner product \( \langle \cdot, \cdot \rangle_{L^2(M)} \) and norm \( \| \cdot \|_{L^2(M)} \), and denote by \( L^2(\mathbb{R}^3, \mathbb{R}^3) \) the space of square integrable vector-valued functions from \( \mathbb{R}^3 \) to \( \mathbb{R}^3 \), with inner product \( \langle \cdot, \cdot \rangle_{L^2(\mathbb{R}^3, \mathbb{R}^3)} \) and norm \( \| \cdot \|_{L^2(\mathbb{R}^3, \mathbb{R}^3)} \). Let \( \mathcal{V}(\mathbb{R}^3) \subset L^2(\mathbb{R}^3, \mathbb{R}^3) \) be a Reproducing Kernel Hilbert Space (RKHS) of smooth functions with compact support in \( \mathbb{R}^3 \). We then introduce a diffeomorphic operator \( \phi \) such that \( \phi_p : \mathbb{R}^3 \to \mathbb{R}^3 \) is a diffeomorphic function for every choice of \( v \in \mathcal{V}(\mathbb{R}^3) \) (Younes, 2019). We denote by \( \phi_p(M) \) the geometric object given by displacing every point \( p \in M \subset \mathbb{R}^3 \) to the new location \( \phi_p(p) \). A direct consequence of \( \phi \) being diffeomorphic is that \( \phi_p(M) \) is topologically equivalent to a sphere for every choice of \( v \in \mathcal{V}(\mathbb{R}^3) \). The construction of the diffeomorphic operator adopted in this paper and the computation of \( v_i \in \mathcal{V}(\mathbb{R}^3) \) such that \( \phi_{v_i}(M) \approx M_i \) are detailed in Appendix A.1.

Next, we use the estimated \( v_i \), and \( z_i \), to define the spatially normalized function \( x_i : M \to \mathbb{R} \) as \( x_i = z_i \circ \phi_p \). For each point \( p \in M, x_i(p) \) is given by \( z_i \circ \phi_p^{-1}(p) \), that is, the value of \( z_i \) at the corresponding point \( \phi_p^{-1}(p) \in M_i \). We can then represent each FoS \( (M_i, z_i) \) with a pair of functions \( (v_i, x_i) \) such that

\[
(M_i, z_i) \approx (\phi_{v_i}(M), x_i \circ \phi_p^{-1}),
\]

where \( v_i \in \mathcal{V}(\mathbb{R}^3) \) and \( x_i \in L^2(M) \).

We can depict the representation model introduced as follows:

\[
(v_i, x_i) \leftrightarrow (M_i, z_i),
\]

meaning that given a FoS \( (M_i, z_i) \), we can compute a loss-less representation \( (v_i, x_i) \) as described earlier, and vice-versa, given the representation \( (v_i, x_i) \), we can compute the associated FoS through equation (2). Hence, the pair of functions \( (v_i, x_i) \) provides us with a linear representation of the original FoS \( (M_i, z_i) \) where every geometric object \( M_i \) is modelled as a (diffeomorphic) deformation of the template, i.e. \( \phi_{v_i}(M) \), while the associated function \( z_i \) is given by ‘transporting’ the spatially normalized function \( x_i \) onto \( M_i \) with such a deformation.

The approach described allows us to recast the original non-Euclidean problem of learning a classifier from the training sample \( \{(g_i, M_i, z_i) : i = 1, \ldots, n\} \) as the problem of learning a classifier from

\[
\{(g_i, v_i, x_i) : i = 1, \ldots, n\},
\]

where \( v_i \in \mathcal{V}(\mathbb{R}^3) \subset L^2(\mathbb{R}^3, \mathbb{R}^3) \) and \( x_i \in L^2(M) \) are two functional predictors both belonging to linear function spaces. In Figure 1b, we show the functional linear representations associated with the three FoSs in Figure 1a.

Crucially, the representation mapping employed here is ‘invertible’, meaning that any pair of estimates \( (g^i, p^i) \in \mathcal{V}(\mathbb{R}^3) \times L^2(M) \), such as the ‘direction’ that optimally discriminates between two classes, can be mapped back to the original space of FoS using equation (2). This mapping defines the associated trajectories of FoSs

\[
\left\{ \left( \phi_{c_1 \beta c}(M), c_2 \beta \circ \phi^{-1}_{c_1 \beta c} \right), c_1, c_2 \in \mathbb{R} \right\},
\]
where \( \varphi_{ei}^{\beta i}(M) \) is guaranteed to be topologically equivalent to a sphere, thereby satisfying the physical constraints of the problem considered.

In contrast to methods that require computing shape features, such as the spectrum of the Laplace–Beltrami operator (Reuter et al., 2006; Wachinger et al., 2015), the approach adopted here provides us with interpretable discriminant directions in the space of the original neuro-biological objects, as described by equation (5). In addition, unlike approaches that work with global parametrizations of FoSs (see, e.g. Chung et al., 2008; Epifanio & Ventura-Campos, 2014; Ferrando et al., 2020; Zaetz & Kurtek, 2015), the representation model used in this study is independent of the imaging data type, as long as we can specify how \( \varphi_{ei} \) deforms our objects and a suitable similarity measure. Therefore, the framework proposed in this work has the potential to accommodate additional types of data, such as streamlines generated from diffusion tensor images, where there may not be a one-to-one correspondence across subjects, but for which optimal transport similarity measures have been developed (Feydy et al., 2017). Although our work assumes that the FoSs are in one-to-one correspondence, this is not strictly necessary. Assuming a one-to-one correspondence simplifies the definition of a similarity measure and makes it easier to compute these representations for complex objects such as cortical surfaces, as detailed in Appendix A.1. In Section 6, we compare the performance of our representation model with alternative models, in the context of our motivating application.

2.2 Discriminant analysis on the parametrizing linear function spaces

The aim of Sections 3 and 4 is to provide methodology for learning a linear classifier, from the training data \( \{(g_i, v_i, x_i) \mid i = 1, \ldots, n\} \) displayed in Figure 1b, by introducing a novel functional classification model that has the following crucial characteristics:

- Does not rely on Functional Principal Components Analysis (FPCA), or related dimension reduction methods, to reduce the dimension of the functional predictors, bypassing the intrinsic assumption that the discriminant direction is well represented by the space spanned by the first few unsupervised PC functions.
- Can be applied to bivariate, and possibly multivariate, functional predictors each supported on a different domain.
- Allows for explicit spatial regularization of the estimates on potentially nonlinear manifold domains, yielding well-defined and interpretable estimates.
- Provides a direct approach to estimating the discriminant directions without relying on prior computation of the covariance structure, which is prohibitive in our application setting.

3 Linear discriminant analysis over general domains

We begin by focusing on the sub-problem of defining a classifier for the training data \( \{(g_i, x_i)\} \), i.e. for the spatially normalized functional predictors supported on the fixed nonlinear manifold \( M \). In Section 4, we then extend the proposed model to account for the geometric component \( v_i \), in an additive fashion.

Assume the training set \( \{(g_i, x_i)\} \) consists of \( n \) independent copies of \( (G, X) \), a pair of random variables with \( X \) a zero-mean random function taking values in \( L^2(M) \) and \( G \) a binary random variable such that \( P(G = 1) = \pi_1 \) and \( P(G = 2) = \pi_2 \). Let \( \mu_1 = \mathbb{E}[X \mid G = 1] \) and \( \mu_2 = \mathbb{E}[X \mid G = 2] \) denote the conditional means of \( X \) and assume \( \mu_1 \neq \mu_2 \). Moreover, let \( C(p, q) = \mathbb{E}[X(p)X(q)] \), \( p, q \in M \) denote the covariance function of \( X \) and assume this is square integrable, i.e. \( \int_M |C(p, q)|^2 dp dq < \infty \). For a real, symmetric, square-integrable, and non-negative function \( R \in L^2(M \times M) \), let the integral operator \( L_R : L^2(M) \rightarrow L^2(M) \) be defined as

\[
L_R(\beta)(\cdot) = \int_M R(p, \cdot) \beta(p) \, dp, \quad \forall \beta \in L^2(M).
\]
Consequently, $L_C$ denotes the covariance operator of $X$, which is a compact self-adjoint operator and therefore admits the following spectral representation

$$L_C(\beta) = \sum_{k=1}^{\infty} \theta_k \langle \beta, e_k \rangle_{L^2(M)} e_k,$$

in terms of the eigenvalues $\theta_1 \geq \theta_2 \geq \ldots \geq 0$ and associated eigenfunctions $e_1, e_2, \ldots \in L^2(M)$ of $L_C$.

Let $L_C^{-1}$ denote the linear inverse covariance operator, where $L_C^{-1}(e_k) = \frac{1}{\theta_k} e_k$ for all $k \geq 1$. Assume that $\|L_C^{-1}(\mu_2 - \mu_1)\|_{L^2(M)} < \infty$ and define the population quantity $\beta^0 \in L^2(M)$ such that

$$L_C \beta^0 = \mu_2 - \mu_1.$$

Note that this is an assumption on the underlying population quantities and will not have practical implications. However, it allows us to have a unique well-defined variable $\beta^0$ to study the convergence properties of the proposed model. For a discussion on the case $\|L_C^{-1}(\mu_2 - \mu_1)\|_{L^2(M)} = \infty$, which is related to the perfect classification phenomenon, see Delaigle and Hall (2012), Berrendero et al. (2018), Chen and Jiang (2018) and Kraus and Stefanucci (2019).

The function $\beta^0$ can be understood as a functional analogue of the multivariate discriminant vector of a linear discriminant analysis (Shin, 2008). For a new observation with predictor $x^* \in L^2(M)$, it can be used to predict the associated label $g^*$ with the linear classification rule $\langle \beta^0, x^* \rangle_{L^2(M)} > c^e$, with $c^e$ an appropriately chosen threshold. Moreover, if $X$ is a Gaussian random function within each group in $G$, it can be shown that the function $\beta^0$ defines the linear classifier that minimizes the misclassification error rate, and it is therefore optimal (Delaigle & Hall, 2012). The discriminant direction $\beta^0$ can also be equivalently defined as the minimizer of the functional

$$\frac{1}{2} \langle \beta, L_C \beta \rangle_{L^2(M)} - \langle \mu_2 - \mu_1, \beta \rangle_{L^2(M)}.$$

In practice, the population quantities $\mu$, $\mu_1$, and $\mu_2$ are unknown and need to be estimated from the data. The goal of a classification model is therefore to recover $\beta^0$ from the training sample $\{(g_i, x_i) : i = 1, \ldots, n\}$ of $n$ independent copies of $(G, X)$. This can be achieved by using the sample covariance $L_{\hat{C}}$, with

$$\hat{C}(p, q) = \frac{1}{n} \sum_{i=1}^{n} x_i(p)x_i(q), \quad p, q \in M$$

and the sample conditional means $\hat{\mu}_1$ and $\hat{\mu}_2$ to replace the population counterparts in equation (7). An estimate $\hat{\beta}$ of $\beta^0$ can then be defined as a minimizer of

$$\frac{1}{2} \langle \beta, L_{\hat{C}} \beta \rangle_{L^2(M)} - \langle \hat{\mu}_2 - \hat{\mu}_1, \beta \rangle_{L^2(M)} + \mathcal{P}(\beta),$$

where a penalty term $\mathcal{P}(\beta)$ is typically added to overcome the ill-posedness of the minimization problem, which is due to the low-rank structure of $L_C$. For instance, Park et al. (2021) define a penalty $\mathcal{P}$ that encourages the estimate $\hat{\beta}$ to be smooth and sparse, while Kraus and Stefanucci (2019) consider a ridge-type penalty.

The functional discriminant model in equation (8) requires precomputing the empirical covariance function, which is generally not possible for dense functional data supported on multidimensional domains and is ultimately not feasible in our application setting. We therefore propose a direct regularized approach to estimating $\beta^0$. This will be possible thanks to the following simple
observation. As noted for instance in Delaigle and Hall (2012), the discriminant direction \( \beta^0 \) can be equivalently characterized as the solution to the minimization problem
\[
\beta^0 = \arg \min_{\beta} \mathbb{E} \left[ Y - (X, \beta)^2 \right],
\]
where \( Y \) is an auxiliary scalar random variable such that \( Y = -\frac{1}{n_1} \) if \( G = 1 \) and \( Y = \frac{1}{n_2} \) otherwise. In other words, the classification problem considered can be reformulated as a functional regression problem. This motivates the adoption of a least-squares approach to estimating \( \beta^0 \), based on the empirical counterpart of the objective function in equation (9), where an additional differential regularization term is introduced to incorporate information on the geometric domain \( \mathcal{M} \) and overcome the ill-posedness of the problem. For multivariate data, analogous least-squares formulations have also been adopted, for instance, in Hastie et al. (1994), Mai et al. (2012), and Gaynanova (2020).

3.1 Regularized estimation
Given the training sample \((y_i, x_i)\), introduce a scalar variable \( y_i \) such that \( y_i = -\frac{G}{n_1} \) if \( g_i = 1 \) and \( y_i = \frac{G}{n_2} \) otherwise, where \( n_1 \) and \( n_2 \) represent the sample sizes of class 1 and 2, respectively. Observe that \(-\frac{n}{n_1} \) and \( \frac{n}{n_2} \) are estimates of the values that can be taken by the random variable \( Y \). Let \( W^2(\mathcal{M}) \) be the Sobolev space of functions in \( \mathcal{L}^2(\mathcal{M}) \) with first and second distributional derivatives in \( \mathcal{L}^2(\mathcal{M}) \). We define an estimate \( \hat{\beta} \in W^2(\mathcal{M}) \) of the population quantity \( \beta^0 \) as the solution to the following minimization problem
\[
\hat{\beta} = \arg \min_{\beta \in W^2(\mathcal{M})} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - (x_i, \beta) \mathcal{L}^2(\mathcal{M}) \right)^2 + \lambda J(\beta),
\]
where the first term is a least-squares estimate of the objective function in equation (9) and the second term is a differential regularization term. The parameter \( \lambda \) controls the trade-off between the least-squares term of the objective function and the penalty term. Our choice of the regularization term is
\[
J(\beta) = \| \Delta_{\mathcal{M}} \beta \|_{\mathcal{L}^2(\mathcal{M})}^2 + \epsilon \| \beta \|_{\mathcal{L}^2(\mathcal{M})}^2,
\]
with \( \epsilon \geq 0 \). This is a linear combination of two terms. The first one is based on the Laplace-Beltrami operator \( \Delta_{\mathcal{M}} : W^2 \subset \mathcal{L}^2(\mathcal{M}) \rightarrow \mathcal{L}^2(\mathcal{M}) \) and quantifies the smoothness of the function \( \beta : \mathcal{M} \rightarrow \mathbb{R} \) on the nonlinear manifold domain \( \mathcal{M} \). Specifically, it allows the model estimate \( \hat{\beta} \), at any fixed point \( p \in \mathcal{M} \), to borrow strength from the other points on \( \mathcal{M} \) while constraining the ‘information’ to propagate coherently with the nonlinear manifold structure of the anatomical object \( \mathcal{M} \). The second term is a generic shrinkage-type regularization.

It is worth noting that the function space \( \mathcal{L}^2(\mathcal{M}) \) is linear, even though each function \( f \in \mathcal{L}^2(\mathcal{M}) \) is supported on a nonlinear domain. For Euclidean domains, it is common to define a smooth subspace of \( \mathcal{L}^2(\mathcal{M}) \) by forming an RKHS from a positive-definite kernel. However, constructing a positive-definite kernel that is compatible with the geometry of a manifold is a non-trivial task (Feragen et al., 2015; Jayasumana et al., 2015). To overcome this challenge, we constructively define a Sobolev norm \( J^{1/2}(\cdot) \) and an associated Sobolev space \( W^2(\mathcal{M}) \) by leveraging a local differential operator, namely the Laplace–Beltrami operator. The discrete counterpart of this local operator is a sparse matrix, reducing our problem to sparse linear algebra and enabling us to solve equation (10) for the large data of our final application. We provide more details about the relationship of our approach with the RKHS approach in Section 4.2.

3.2 Theory
The aim of this section is to provide theoretical guarantees for the performance of the proposed model. Specifically, we provide a probability bound for the out-of-sample risk, i.e. the random variable
\[ E^* \left[ (X^*, \beta^0 - \hat{\beta}) \right]_{\mathcal{L}^2(M)}^2, \]  

where \( X^* \) is a copy of \( X \) that is independent of the training data and \( E^* \) is the expectation taken over \( X^* \). Equation (12) measures the discrepancy between the prediction made with the estimated parameter \( \hat{\beta} \) and the ‘optimal’ prediction made with the unknown population quantity \( \beta^0 \).

Assume for simplicity that \( \varepsilon > 0 \). Then, thanks to the Sobolev embedding theorem (Brezis, 2011), \( \exists M \geq 0 \) such that for any \( p \in \mathcal{M} \)

\[ f(p) \leq \sup_q |f(q)| \leq M \left( \| \Delta_M f \|_{\mathcal{L}^2(M)} + \epsilon \| f \|_{\mathcal{L}^2(M)} \right)^{1/2}, \quad \forall f \in \mathcal{W}^2(M), \]

that is, the evaluation operator is a continuous functional. A direct consequence is that the space \( \mathcal{W}^2(M) \) equipped with the norm \( f^{1/2}(\cdot) = (\| \Delta_M \cdot \|_{\mathcal{L}^2(M)} + \epsilon \| \cdot \|_{\mathcal{L}^2(M)}^{1/2} \) is an RKHS with a symmetric, positive definite kernel function \( K_M : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R} \) (Berlinet & Thomas-Agnan, 2004). The kernel function \( K_M \) is used only for theoretical investigation here and obtaining its explicit form is, in general, not computationally feasible and not necessary. We will, however, take advantage of the fact that \( L_{K_M}^{1/2}(\mathcal{L}^2(M)) = \mathcal{W}^2(M) \), where \( L_{K_M}^{1/2} \) denotes the square root of \( L_{K_M} \) (Cucker & Smale, 2002). For \( \epsilon = 0 \), the functional \( J^{1/2} \) defines a semi-norm rather than a norm and similar arguments hold by restricting ourselves to the subspace of \( \mathcal{L}^2(M) \) that is orthogonal to the null space of \( J^{1/2} \).

Next, we define the sandwich operator \( T = L_{K_M}^{1/2} L_{C} L_{K_M}^{1/2} \) (Cai & Yuan, 2012) and make the following assumptions.

**Assumption 3.1** The constant \( \kappa^2 \), defined as \( \kappa^2 = \text{ess sup} \| L_{K_M}^{1/2} X \|_{\mathcal{L}^2(M)}^2 \) is finite.

**Assumption 3.2** There exists a smooth function \( \beta^0 \in \mathcal{W}^2(M) \) such that \( \beta^0 = L_C^{-1}(\mu_2 - \mu_1) \).

**Assumption 3.3** The effective dimension of \( T \) satisfies \( D(\lambda) = \text{Tr}((T + \lambda I)^{-1}T) \leq c\lambda^{-\theta} \) for constants \( c, \theta > 0 \). Here \( \text{Tr} \) denotes the trace operator.

Assumption 3.1 allows us to use a Hoeffding-type inequality for Hilbert space valued random elements and has no practical implications. This condition is met, for example, when \( \| X \|_{\mathcal{L}^2(M)} \) is bounded. However, it is more general given that \( L_{K_M}^{1/2} X \) represents a smoothed version of \( X \). Assumption 3.2 guarantees that the population quantity \( \beta^0 \) is well defined and belongs to the space of smooth functions \( \mathcal{W}^2(M) \). Assumption 3.3 is expressed in terms of properties of the effective dimension \( D(\cdot) \). For our final choice of \( \lambda_\tau \), it is straightforward to check that this assumption holds by assuming that the eigenvalues \( \{ \tau_k \} \) of \( T \) decay as \( \tau_k \propto k^{-\tau r} \), with \( r > \frac{1}{2} \). This is a typical assumption in the literature on functional linear models (Cai & Yuan, 2012) and is related to the rate of decay of the eigenvalues of \( L_{K_M} \) and \( L_C \), and their alignment.

The following theorem provides an upper bound for the out-of-sample risk.

**Theorem 3.1** Under Assumptions 3.1–3.3, if \( \lambda \asymp n^{-\frac{1}{2m}} \), the estimator \( \hat{\beta} \) in equation (10) is such that

\[ E^* \left[ (X^*, \beta^0 - \hat{\beta}) \right]_{\mathcal{L}^2(M)}^2 = O_p \left( n^{-\frac{1}{2m}} \right). \]  

Similar rates of convergence have been shown to hold for regularized estimates in the functional linear regression setting (see, e.g. Cai & Yuan, 2012; Reimherr et al., 2018; Sun et al., 2018; Tong & Ng, 2018). However, a key difference in our model is that the residual random variable \( \varepsilon = Y - \langle X, \beta^0 \rangle_{\mathcal{L}^2(M)} \) and the functional predictor \( X \) are not independent, which prevents the direct
application of such results. Therefore, Theorem 3.1 shows that in spite of such a dependence structure we are nevertheless able to recover the functional linear model rates of convergence. The proof is provided in Appendix C.

3.3 Nonlinear extensions
To incorporate nonlinearity into the model described in equation (10), one can substitute the term $\langle x_i, \beta \rangle_{L^2(M)}$ with a nonlinear function of the data, such as a polynomial term or a single-index model, as done in the context of functional regression models in Yao and Müller (2010) and Jiang and Wang (2011), respectively. However, these extensions come at the cost of estimating additional functional parameters or optimizing a more complex objective function.

If the covariance structures of the two classes are believed to be different, the proposed functional linear discriminant model can be generalized to an approximate functional quadratic discriminant model outperforming the functional linear discriminant model when the covariance structures of the two classes differ. Examining the theoretical properties of this extension is beyond the scope of this paper and is left to future work.

4 Additive multivariate generalizations
We now consider a bivariate extension of the functional model introduced in Section 3, which incorporates the geometric component of the original data. We therefore consider the training sample $\{(g_i, v_i, x_i)\}$, where $v_i \in V(\mathbb{R}^3)$ is a vector field representing the subject-specific geometry of the $i$th subject. Recall that $(V(\mathbb{R}^3), \| \cdot \|_{V(\mathbb{R}^3)})$ is an RKHS of smooth functions with compact support in $\mathbb{R}^3$. Moreover, denote by $K_{\mathbb{R}^3}$ its reproducing kernel.

Suppose the training set $\{(g_i, v_i, x_i)\}$ consists of $n$ independent copies of $(G, V, X)$, a triplet of random variables with $V$ a zero-mean random function taking values in $V(\mathbb{R}^3)$, $X$ a zero-mean random function taking values in $L^2(M)$, and $G$ a binary random variable such that $P(G = 1) = \pi_1$ and $P(G = 2) = \pi_2$.

We now adopt the multivariate functional data notation from Happ and Greven (2018), and define $X(p) = (V(p_1), X(p_2))$, with $p = (p_1, p_2) \in D = D_1 \times D_2 = \mathbb{R}^3 \times M$. The multivariate random function $X$ takes values in a Hilbert space $\mathcal{H} = L^2(\mathbb{R}^3, \mathbb{R}^3) \times L^2(M)$ with inner product $(f, g)_\mathcal{H} = (f^{(1)}, g^{(1)})_{L^2(\mathbb{R}^3, \mathbb{R}^3)} + (f^{(2)}, g^{(2)})_{L^2(M)}$ for $f, g \in \mathcal{H}$. Here $f^{(j)}$, with $j \in \{1, 2\}$, denotes the $j$th functional component of the multivariate function $f$. For $p, q \in D_i$, define the matrix of covariances $C(p, q) = \mathbb{E}(X(p) \otimes X(q))$ with elements $C_{ij}(p_l, q_l) = \mathbb{E}[X^{(i)}(p_l)X^{(j)}(q_l)]$ where $p_l \in D_{il}$, $q_l \in D_{j}$, $l \in \{1, 2\}$, and $i \in \{1, 2\}$. Denote the conditional means of $X$ by $\mu_i = (\mu_i^{(1)}, \mu_i^{(2)}) := (\mathbb{E}[V \mid G = 1], \mathbb{E}[X \mid G = 1])$ and $\mu_2 = (\mu_2^{(1)}, \mu_2^{(2)}) := (\mathbb{E}[V \mid G = 2], \mathbb{E}[X \mid G = 2])$, and assume $\mu_1 \neq \mu_2$. The covariance operator $L_C : \mathcal{H} \to \mathcal{H}$ is such that the $j$th component of $L_C f$, for any $f \in \mathcal{H}$, is given by

$$
(L_C f)^{(j)}(p_l) = \sum_{i=1}^{2} \int_{D_i} C_{ij}(q_l, p_l) f^{(i)}(q_l) \, dq_l.
$$

(14)
Similar to the univariate case, we assume that the population quantity $\beta^0 \in \mathcal{H}$ is well-defined and satisfies the equation

$$L_{C} \beta^0 = \mu_2 - \mu_1.$$ 

This can be viewed as a multivariate generalization of the linear discriminant direction defined in the previous section. We now turn to the problem of defining an estimator for $\beta^0$.

### 4.1 Regularized estimation

Let the variable $y_i$ be such that $y_i = -\frac{\mu}{n_1}$ if $g_i = 1$ and $y_i = \frac{\mu}{n_2}$ otherwise. We define the multivariate functional estimate $\hat{\beta} = (\hat{\beta}^G, \hat{\beta}^F)$ of the population quantity $\beta^0$ to be the solution to the following minimization problem

$$\begin{align*}
(\hat{\beta}^G, \hat{\beta}^F) &= \arg \min_{\beta^G \in \mathcal{V}(\mathbb{R})} \min_{\beta^F \in \mathcal{V}(\mathcal{M})} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \langle v_i, \beta^G \rangle_{L^2(\mathbb{R}^3, \mathbb{R})} - \langle x_i, \beta^F \rangle_{\mathcal{L}^2(\mathcal{M})} \right)^2 + \lambda_1 \| \beta^G \|^2_{L^2(\mathbb{R}^3)} + \lambda_2 J(\beta^F),
\end{align*}$$

(15)

where $\lambda_1, \lambda_2 > 0$ are tuning parameters.

Equation (15) extends the model proposed in Section 3 to account for both the geometric functional descriptor $v_i$ and the function $x_i$ in an additive fashion. The regularization terms in the equation enforce smoothness on the functional estimates $\hat{\beta}^G$ and $\hat{\beta}^F$ in their respective function spaces.

### 4.2 Differential regularization and kernel penalty: a unified modelling perspective

In Section 4.1, we have adopted two different approaches to produce smooth estimates $\hat{\beta}^G : \mathbb{R}^3 \to \mathbb{R}^3$ and $\hat{\beta}^F : \mathcal{M} \to \mathbb{R}$. The smoothness of $\hat{\beta}^F$ is enforced by means of a penalty $J(\cdot)$ defined in terms of a Sobolev norm, which implicitly defines a kernel $K_{\mathcal{M}}$. On the other hand, the smoothness of $\hat{\beta}^G$ is enforced by means of a norm $\| \cdot \|_{L^2(\mathbb{R}^3)}$, defined implicitly through the direct definition of a kernel $K_{\mathbb{R}^3}$. For clarity, we summarize the relevant function spaces and associated norms and kernels in Table 1.

From a methodological perspective, the reproducing kernel $K_{\mathbb{R}^3}(p, q)$ can be understood as a measure of the influence of the function value at $p \in \mathbb{R}^3$ on the function value at $q \in \mathbb{R}^3$ and vice-versa. Defining a smooth function space through a kernel has arguably an advantage when it comes to discretizing an infinite-dimensional minimization problem over that function space. In fact, thanks to the well-known representer theorem (Wahba, 1990; Yuan & Cai, 2010), under mild conditions, its exact solution can be expressed as a linear combination of the elements of a $n$-dimensional basis, which involves the explicit expression of the kernel.

Hence, it is natural to wonder whether a similar approach could be adopted for $\hat{\beta}^F : \mathcal{M} \to \mathbb{R}$. In other words, can we define a smooth real function space on $\mathcal{M}$ by explicitly defining a kernel $K_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ encoding a measure of influence that is coherent with the nonlinear geometry of $\mathcal{M}$? This, however, is a challenging task due to the positive-definiteness property that $K_{\mathcal{M}}$ must satisfy. Consider, for instance, the popular exponential kernel. Its natural extension to the manifold setting is $K_{\mathcal{M}}(p, q) = \exp (-c \cdot d_{\mathcal{M}}(p, q)^2)$, where $d_{\mathcal{M}}(p, q)$ is the geodesic distance between $p \in \mathcal{M}$ and $q \in \mathcal{M}$. Unfortunately, this kernel cannot be guaranteed to be positive definite for a general nonlinear manifold $\mathcal{M}$ (Feragen et al., 2015; Jayasumana et al., 2015).
Alternatively, we could try to compute an explicit form of the kernel \( K_M \) from \( J^{1/2}(\cdot) \) by employing the following identity (Fasshauer & Ye, 2013; Wahba, 1990)

\[
f(\tau) = \langle K_M(p, \cdot), f \rangle_{W^2(M)}, \quad \forall p \in M, f \in W^2(M),
\]

where \( \langle \cdot, \cdot \rangle_{W^2(M)} \) is the inner product that induces the norm \( J^{1/2}(\cdot) \). However, closed-form solutions to equation (16) are not available in our setting. Approximate solutions could be computed by Finite-Elements Analysis (FEA) (Quarteroni, 2009), but we would still face the challenge of storing the dense object \( K_M(\cdot, \cdot) \). As described in Section 3, we instead leverage FEA to directly discretize the function \( \hat{\beta}^F : M \to \mathbb{R} \) in equation (15).

This highlights that the choice of the two modelling approaches is not arbitrary and that, arguably, for functional estimates supported on Euclidean spaces, defining explicitly a reproducing kernel is likely the preferred choice. Meanwhile, for general non-Euclidean domains, where defining a reproducing kernel is not trivial, the differential penalization approach is preferable.

4.3 Theory

Define the diagonal matrix of reproducing kernels \( K(p, q) \) with entries \( K_{11}(p_1, q_1) = K_{22}(p_2, q_2) = K_{12}(p_2, q_2) = K_{21}(p_2, q_2) : p_1 \in D_1, p_2 \in D_2 \). Let \( L_K : \mathcal{H} \to \mathcal{H} \) be the associated integral operator and, analogously to the univariate functional setting, define the sandwich operator \( T = L_K^{1/2}L_C L_K^{1/2} \). We make the following assumptions, which are analogous to Assumptions 3.1–3.3.

Assumption 4.1 The constant \( \kappa_2^2 \), defined as \( \kappa_2^2 = \text{ess sup} \| L_K^{1/2}X \|_H^2 \) is finite.

Assumption 4.2 There exists a smooth function \( \beta^0 \in V(\mathbb{R}^3) \times W^2(M) \) such that \( \beta^0 = L_K^{-1}(\mu_2 - \mu_1) \).

Assumption 4.3 The penalty coefficient \( \lambda : = \lambda_1 = \lambda_2 \) and the effective dimension of \( T \) satisfy \( D(\lambda) = \text{Tr}((T + \lambda I)^{-1}T) \leq c\lambda^{-\theta} \) for constants \( c, \theta > 0 \).

The following theorem, which is an extension of Theorem 3.1, provides an upper bound for the out-of-sample risk.

Theorem 4.1 Under Assumptions 4.1–4.3, if \( \lambda \asymp n^{-1/m} \), the estimator \( \hat{\beta} = (\hat{\beta}^G, \hat{\beta}^F) \) in equation (15) is such that

\[
\mathbb{E}^\tau[(X^*, \beta^0 - \hat{\beta})_{\mathcal{H}}]^2 = O_n(n^{-1/m}),
\]

where \( X^* \) is a copy of \( X \) that is independent of the training data and \( \mathbb{E}^\tau \) is the expectation taken over \( X^* \).

5 Discretization

Consider a triangle mesh, denoted by \( M_T \), which is formed by the union of a finite set of triangles, \( T \). These triangles share a common set of \( s \) vertices, denoted as \( \xi_1, \ldots, \xi_s \). Let \( M_T \) be an approximate representation of the manifold \( M \). We then introduce the linear finite-element space \( W_T \) consisting of a set of globally continuous functions over \( M_T \) that are affine within each triangle \( \tau \) in \( T \), i.e.

\[
W_T = \{ w \in C^0(M_T) : |w|_\tau \text{ is affine } \forall \tau \in T \}.
\]

The space \( W_T \) is spanned by the Finite-Elements (FE) basis \( \psi_1, \ldots, \psi_s \), where \( \psi_l(\xi_i) = 1 \), if \( l = i \), and \( \psi_l(\xi_i) = 0 \) otherwise. In Figure 2, we show one element of this basis. Moreover, define \( \psi \) as
the vector-valued function \( \psi(\cdot) = (\psi_1(\cdot), \ldots, \psi_s(\cdot))^T \). Our goal is to find an approximate solution \( \hat{\beta}_T^F \) of the form

\[
\hat{\beta}_T^F(\cdot) = \sum_{i=1}^{s} c_i^F \psi_i(\cdot) = (c_i^F)^T \psi(\cdot),
\]

where \( c_i^F = (c_i^1, \ldots, c_i^F)^T \).

Let now \( M \) and \( S \) be the sparse mass and stiffness \( s \times s \) matrices defined as \( (\mathcal{M})_{jifinmath} = \int_{\mathcal{M}_T} \psi_j \psi_{jifinmath} \sum \text{ and } (\mathcal{S})_{jifinmath} = \int_{\mathcal{M}_T} \nabla \psi_j \cdot \nabla \psi_{jifinmath} \sum, \text{ where } \nabla \psi_{jifinmath} \sum \text{ is the gradient on the mesh } \mathcal{M}_T. \) For \( \hat{\beta}_T^F \) of the form given in equation (18), we have that the penalty term \( f(\cdot) \) can be approximated as \( (c_i^F)^T D_{\mathcal{M}_T} c_i^F \), with \( D_{\mathcal{M}_T} = S M^{-1} S + \varepsilon M \) (Lila et al., 2016). Further, following an approach often adopted in FEA (Fried & Malkus, 1975; Hinton et al., 1976), we replace the dense matrix \( M^{-1} \) with the sparse matrix \( \hat{M}^{-1} \), where \( \hat{M} \) is the diagonal matrix such that \( \hat{M}_{jj} = \sum_j M_{jj} M_{jj}^{\text{dual}} \sum \text{ with the mesh } \mathcal{M}_T. \) This results in the sparse penalty matrix \( D_{\mathcal{M}_T} = \hat{M}^{-1} S + \varepsilon M. \) In practice, each functional observation \( x_i \) is also of the form given in equation (18). Therefore, denoting by \( \% \) the \( n \times s \) matrix where each row consists of the basis coefficients of \( x_i, \text{ the terms } \{ (x_i, \beta_i^F)_{\mathcal{M}_T} \} \text{ can be approximated by the entries of the vector } \% M c_i^F. \)

We now turn our attention to the estimate \( \hat{\beta}_G^G \in \mathcal{V}(\mathbb{R}^3). \) Since for \( \mathcal{V}(\mathbb{R}^3) \) we have an explicit form of the associated reproducing kernel \( K_{\mathbb{R}^3}, \) we employ the representer theorem (Wahba, 1990; Yuan & Cai, 2010) and take \( \hat{\beta}_G^G(\cdot) \) of the form

\[
\hat{\beta}_G^G(\cdot) = \sum_{i=1}^{n} c_i^G \int_{\mathbb{R}^3} K_{\mathbb{R}^3}(p, \cdot) \psi_i(p) \, dp.
\]

On the right-hand side of Figure 2, we show an example of a basis function \( \int_{\mathbb{R}^3} K_{\mathbb{R}^3}(p, \cdot) \psi_i(p) \, dp. \) Then, we have that \( \| \beta^G \|^2_\mathcal{V} = (c_i^G)^T \Sigma c_i^G, \) where \( c_i^G = (c_i^1, \ldots, c_i^G)^T \) and \( \Sigma \) is a \( n \times n \) matrix with entries

\[
\Sigma_{ij} = \mathcal{F} \psi_i(p)^T K_{\mathbb{R}^3}(p, q) \psi_j(q) \, dp \, dq.
\]

As a result, the coefficients of the approximate solution of the model in equation (15) are given by

\[
(c_i^G, c_i^F) = \arg \min_{c_i^G, c_i^F} \| y - \Sigma c_i^G - \% M c_i^F \|^2_2 + \lambda_1 (c_i^G)^T \Sigma c_i^G + \lambda_2 (c_i^F)^T D_{\mathcal{M}_T} c_i^F,
\]

Figure 2. On the left-hand side, we show an element of the FE basis \( \{ \psi_i : \mathcal{M}_T \to \mathbb{R}, i = 1, \ldots, s \}. \) This is a scalar affine function within each triangle of the mesh \( \mathcal{M}_T \) that takes value 1 on a fixed vertex and value 0 on every other vertex. On the right-hand side, we show an element of the basis \( \{ \int_{\mathbb{R}^3} K_{\mathbb{R}^3}(p, \cdot) \psi_i(p) \, dp, i = 1, \ldots, n \}. \) This is a smooth vector-valued function from \( \mathbb{R}^3 \) to \( \mathbb{R}^3 \).
where \( y = (y_1, \ldots, y_n)^T \) is the vector of auxiliary response variables. It is easy to check that this minimization problem can be equivalently written as the following augmented quadratic least-squares problem

\[
(\hat{c}^G, \hat{c}^F) = \arg \min_{c^G \in \mathbb{R}^n, c^F \in \mathbb{R}^s} \| \begin{bmatrix} y \\ 0 \end{bmatrix} - A \begin{bmatrix} c^G \\ c^F \end{bmatrix} \|_2^2,
\]

(21)

where 0 is the zero-vector of length 2s + n and with

\[
A = \begin{bmatrix}
\Sigma & XM \\
0 & \tilde{M}^{-1}S \\
0 & \tilde{M}^{-1}S \tilde{e}^s e^s M^s \\
\tilde{e}^t \Sigma t & 0
\end{bmatrix}.
\]

Note that for \( n \ll s \), which is the setting of our application, the matrix \( A \) is sparse. Therefore, the minimization problem (21) can be efficiently solved by conjugate gradients, or its variations, e.g. LSQR (Paige & Saunders, 1982), without requiring the explicit computation of the high-dimensional normal matrix \( A^T A \)—a quantity related to the covariance structure of the functional predictors.

An approximate solution to the univariate model in equation (10) follows as a special case of the multivariate case considered here.

6 Application

6.1 Data and preprocessing

We analyze a cohort of \( n = 484 \) subjects from the ADNI and PPMI studies. On the basis of the ATN classification scheme (Jack et al., 2016), each subject was assigned to one of the two diagnostic categories—C, Control (\( n_1 = 100 \)); AD, Alzheimer’s Disease (\( n_2 = 384 \)). Here, we focus on data collected at the baseline visit, which includes, among other imaging modalities, structural T1-weighted MRI.

The T1-weighted images were preprocessed using FreeSurfer (Dale et al., 1999; Fischl et al., 1999). Specifically, white matter, grey matter, and cerebrospinal fluid were segmented and used to extract the outer and inner surfaces of the cerebral cortex. From these two surfaces, we generated a central surface interpolating the midpoints between the outer and inner layers, which offers an accurate representation of the two-dimensional anatomical structure of the cerebral cortex. This representation has the benefit of encoding a notion of distance between brain regions that is neurologically more relevant than the original volumetric representation. The cortical surface can moreover be coupled with one or more maps describing complementary structural or functional properties of the cortex, such as cortical thickness measurements (Fischl & Dale, 2000), fMRI signals, or connectivity maps (Smith et al., 2013; Yeo et al., 2011). In this study, we focus on cortical thickness, which is estimated from the distances between the outer and inner surfaces of the cerebral cortex. Next, the \( n \) surfaces were registered and sub-sampled.

As a result of the preprocessing stage, we obtain \( n = 484 \) triangle meshes \( \{M^T_i\} \) consisting of \( s = 64K \) vertices in correspondence across subjects, along with a set of triangular faces defining how these vertices are connected to delineate the surfaces. By classical Generalized Procrustes Analysis (Dryden & Mardia, 2016), translation, rigid rotation, and scale were removed from the data, while jointly estimating a template \( M^T \), which is also a triangle mesh with \( s = 64K \) vertices in correspondence with those of the individual subjects. Each surface has been moreover coupled with cortical thickness measurements at the mesh vertices, which we model as a real piecewise linear continuous function \( z^T_i \) over the mesh \( M^T_i \).

The preprocessing stage results in a set of FoSs \( \{(M^T_i, z^T_i), i = 1, \ldots, n\} \), which are discretized versions of the continuous idealized objects \( \{(M_i, z_i), i = 1, \ldots, n\} \) introduced in Section 2. To simplify the notation, we drop the superscript \( T \). Moreover, we denote the diagnostic labels by
We fit the proposed model for different choices of the parameters \( \lambda_1 \) and \( \lambda_2 \). Recall that \( \lambda_1 \) controls the regularity of the geometric discriminant direction and \( \lambda_2 \) that of the thickness discriminant direction, with high values virtually constraining the solution to be the zero function. The final choice of \( \lambda_1 \) and \( \lambda_2 \) is the result of a compromise between classification accuracy on the test set and the consistency of the estimated discriminant directions with the neurodegenerative nature of the disease (see also Discussion). The test Area under the ROC Curve (AUC) of the selected model is 0.7006.
AD. These results are consistent with the typical thickness signature of AD observed to date (see, e.g. Bondareff et al., 1989; Dickerson et al., 2009; Sabuncu et al., 2011). The geometric component $\hat{\beta}_G$ is instead a vector field in $\mathbb{R}^3$. This is a linear representation of the morphological variations, with respect to the population average cortex geometry, that are associated with AD. While a full understanding of its meaning is only possible by mapping $\hat{\beta}_G$ back to the space of FoSs, i.e. by examining $\varphi_{-c_1\hat{\beta}_G}(M)$ for different choices of $c_1 \in \mathbb{R}$, the magnitude of the vector field $\hat{\beta}_G$, at any fixed point, offers a rough indication of the cortical regions whose morphological variations are most relevant to the classification problem.

On the right-hand side of Figure 3, we show the FoSs associated with the linear representations $\hat{\beta}_G$ and $\hat{\beta}_F$, that is, $(\varphi_{e_1c_1\hat{\beta}_G}(M), c_2\hat{\beta}_F \circ \varphi^{-1}_{e+1,\hat{\beta}_G}(M))$ with $c_1, c_2 \in \mathbb{R}$. These describe the most predictive patterns of AD in terms of the original neurobiological objects. We have circled a specific area of the brain to ease comparison and highlight the morphological patterns that the model deems relevant to the classification problem.

6.5 Comparison against alternative approaches

In this section, we compare the test AUC of our proposed classification method with alternative models and evaluate different representation models for FoSs. In addition to the functional linear discriminant model (FLDA) that we propose, we also consider the following alternatives: (i) FPCA+LDA: The geometry-aware FPCA model proposed in Lila et al. (2016), followed by multivariate LDA (Hastie et al., 2009) on the PC scores; (ii) Lasso: A logistic regression model with lasso regularization (Tibshirani, 1996); (iii) Ridge: A logistic regression model with an $\ell^2$ regularization (Hoeffding & Kennard, 1970); (iv) FQDA: The approximate functional quadratic discriminant model defined in Section 3.3; (v) RF: A Random forest model (Breiman, 2001); (vi)
Table 2. The test AUC of the classification methods applied to the data of our final application

| Representation model                      | Linear methods                  | Nonlinear methods |
|------------------------------------------|---------------------------------|-------------------|
|                                          | FLDA                            | FPCA +LDA         | Lasso | Ridge | FQDA | RF   | SVM  | NN   |
| Thickness                                | 0.7626                          | 0.7583            | 0.7487 | 0.7632 | 0.7710 | 0.6043 | 0.7597 | 0.7678 |
| Thickness & Displacement                 | 0.6623                          | –                 | 0.6626 | 0.6571 | –     | 0.6742 | 0.6861 | 0.6771 |
| Thickness & Shape spectrum               | –                               | –                 | 0.7832 | 0.6638 | –     | 0.5797 | 0.7606 | 0.6878 |
| Proposed FoSs representation             | 0.7716                          | –                 | 0.7484 | 0.7646 | –     | 0.7132 | 0.7600 | 0.7443 |

Note. Four different representation models have been considered: (i) the registered thickness map \( x_i : M \to \mathbb{R} \) without geometric information; (ii) the parametrization \( h_i : M \to \mathbb{R}^3 \), where the first three components are the surface coordinates, and the last component is thickness; (iii) the registered thickness map \( x_i : M \to \mathbb{R} \) and the first 200 eigenvalues of the Laplace–Beltrami operator computed on the surface \( M_i \); and (iv) the proposed representation model \( (x_i, v_i) \). The symbol ‘-’ indicates that although the method could be adapted to accommodate the specific FoS representation model, its implementation is beyond the scope of this paper and is left to future work.

SVM: A support vector machine with a squared exponential kernel (Cortes & Vapnik, 1995); (vii) NN: A multilayer feedforward neural network (Hastie et al., 2009).

Furthermore, besides the proposed representation model \((v_i, x_i)\) for FoSs, we also consider the following representations: (i) Thickness: Spatially normalized thickness maps \( x_i : M \to \mathbb{R} \) without geometric information; (ii) Thickness & Displacement: The parametrizations \( h_i : M \to \mathbb{R}^4 \), where the first three components are the surface coordinates, and the last component is the (spatially normalized) thickness map; (iii) Thickness & Shape spectrum: The spatially normalized thickness maps \( x_i : M \to \mathbb{R} \) and the first 200 eigenvalues of the Laplace–Beltrami operator computed on the surface \( M_i \), i.e. a spectral representation of shape (Reuter et al., 2006).

To evaluate the listed methods and representation models, we split the dataset into three sets, namely the training set, validation set, and test set, comprising 50%, 20%, and 30% of the data, respectively. While a Monte Carlo evaluation of these methods would be desirable, it is computationally prohibitive, so we defer that analysis to the simulation setting in Appendix B. However, we use the same exact data split for all methods. The models are trained on the training set, hyperparameters are chosen to maximize the AUC on the validation set and the selected model is tested on the test set, resulting in the AUC scores presented in Table 2.

The results are shown in Table 2, from which we can make several observations. Firstly, if the goal is to maximize prediction accuracy, then the best-performing model is a lasso-penalized generalized linear model applied to thickness maps and the first 200 eigenvalues of the Laplace–Beltrami operator of the surfaces. However, as mentioned in the Introduction, this ‘lossy’ shape representation cannot be mapped back to the original space of neurobiological objects, leading to a less interpretable model. Additionally, our results show that in the context of our application, the representation \((v_i, x_i)\) performs better across all methods than using the representation \( h_i : M \to \mathbb{R}^4 \). The latter appears to be more susceptible to overfitting, resulting in inferior performance even when compared to models that use thickness only. Finally, classification using thickness alone produces satisfactory results. The top-performing models are the proposed FLDA and FQDA, and the ridge logistic regression model. One possible explanation for this is that the registered thickness maps may include some geometric information due to misregistration.
incorporating geometric information into the model may lead to only minor improvements in classification performance, as shown in Figure 4, the estimated discriminant direction can be significantly different between the two models. Although the ground truth is unknown, the estimated discriminant direction when geometric information is included is more consistent with the neurodegenerative nature of the disease, as explained in the next section.

6.6 Discussion

The results in Figure 3 identify the typical AD thickness signature. Several studies that focus on identifying AD-vulnerable areas include the regions found in our analysis (see, e.g. Bondareff et al., 1989; Dickerson et al., 2009; Sabuncu et al., 2011). However, there is some variability in the estimated regions. For instance, Sabuncu et al. (2011) used a dynamic model and found strongest changes in the inferior parietal regions and the posterior cingulate. It should be noted that these studies typically consist of massive univariate analyses between the cortical thickness at each vertex, or each parcel, and the diagnostic label. They are therefore taking a feature-centric perspective on the problem. It is not clear how these findings would generalize to out-of-sample data (Li & Tong, 2020).

To demonstrate the importance of modelling cortical geometry, we compare our results to those obtained by fitting a ridge logistic regression model and the proposed model in equation (10), i.e. by discarding inter-subject geometric differences. We compare these estimates in Figure 4. What we observe is that ridge logistic regression yields estimated discriminant directions that are more difficult to interpret, due to the high spatial variability. Except for the entorhinal cortex, the functional model in equation (10) is able to capture the main areas where cortical thinning is associated with AD. However, this model also suggests that a thicker cortex in certain regions (dark red) is associated with AD, contradicting the neurodegenerative nature of AD. Interestingly, introducing the geometric component in the model reduces such effects. This may also be caused by the geometric component now capturing systematic misregistration. In order to verify such a hypothesis, further validation of the estimated geometric component is required in controlled settings where registration is more reliable, e.g. in the longitudinal setting.
7 Conclusions

We introduce a framework for the discriminant analysis of functional data supported on random manifold domains, i.e. FoSs. To this aim, we adopt linear representations of these objects that are bivariate functional data belonging to linear spaces. We then define a functional linear classification model on the parametrizing space. Thanks to a penalized least-squares formulation, the proposed model is able to estimate the most discriminant direction in the data without requiring the explicit computation of the covariance function of the predictors or low-rank approximations thereof. This allows us to reduce the memory requirements by five orders of magnitude and ultimately be able to run our model on a standard workstation. The complexity of the solution is controlled by means of differential penalties that are aware of the geometry of the domain where the functional data are supported.

We apply the proposed model to the analysis of modern multi-modal neuroimaging data. Specifically, we estimate interpretable discriminant directions that are able to leverage both geometric and thickness features of the cerebral cortex to identify subjects with AD. Our results are consistent with those in the neuroscience literature.

The model proposed can be applied to several imaging settings that lead to FoSs representations, such as musculoskeletal imaging (Gee et al., 2018) or cardiac imaging (Biffi et al., 2018). It is also important to highlight that the proposed model is not a mere generalization of existing models for functional data supported on one-dimensional domains to multidimensional domains. We believe that its application to one-dimensional functional data, where the bivariate representation is given by the registered functions and associated registration maps, leads to a novel classification approach in this simplified setting.

Acknowledgments

The authors are grateful to the Editor, Associate Editor, and the three reviewers for their invaluable feedback on previous versions of this manuscript.

Conflict of interest: None declared.

Funding

This work was partially supported by U.S. NIH grant EB033001 and U.S. NSF grant DMS-2210064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or any other funding agency.

Data Availability

Data used in the preparation of this article were obtained from two sources: Alzheimer’s Disease Neuroimaging Initiative (ADNI) and Parkinson’s Progression Markers Initiative (PPMI). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cognstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; Euroimmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.
The main requirement for choosing the kernel is that the associated space \( \mathcal{V}(\mathbb{R}^3) \) is an admissible space (Younes, 2019). Therefore, there must exist a positive constant \( M \) such that for all
The following inequality holds:

$$\|v\|_{1,\infty} \leq M \|v\|_{\mathcal{V}(\mathbb{R}^3)}$$

where $\|\cdot\|_{1,\infty}$ is the canonical norm of the space $C^1(\mathbb{R}^3)$. This condition guarantees that $\varphi_v$ is diffeomorphic for any $v \in \mathcal{V}(\mathbb{R}^3)$.

Nonetheless, within the set of admissible spaces, the specific choice of the kernel can significantly influence the quality of the estimated representations. We have found the approach proposed in Bruveris et al. (2012) to be effective. This uses a mixture of isotropic Gaussian kernels with different variances, which intuitively allows for a multi-scale representation of the surfaces. In our application, we use a mixture of six Gaussian kernels with variance parameters set to $(\sigma_1^2, \ldots, \sigma_6^2) = (64, 16, 4, 1, 0.25, 0.01)$. Our choice was informed by visual inspection of the differences between the estimated $\varphi_{v_i}(M^T)$ and $M^T_i$.

### A.4 Template estimation

In this section, we introduce an algorithm designed to estimate a template leveraging the (formal) Riemannian structure of the manifold of diffeomorphisms adopted to model random manifold domains. One approach is to define the template so that the average of the linear representations $\{v_i\}$, located on the tangent space at the identity map, is zero. The details of this iterative centroid approach are outlined in Algorithm 1. For an overview of alternative approaches see Cury et al. (2014).

Despite leveraging GPU acceleration for computing RKHS norms and associated gradients, the process of computing $v_i$ for each subject still takes about 40 minutes in our application. This makes the process of estimating the template computationally prohibitive given the necessity of multiple iterations. Therefore, we have chosen to use a fixed template in our final application.

### Appendix B. Simulations

In this section, we conduct simulations to assess the finite sample classification performance of the model proposed when compared to the models introduced in Section 6.5. Here we focus on the functional univariate setting described in Section 3.

We use a triangle mesh $M_T$ with 642 nodes that is an approximation of a brainstem. On this triangulated surface, we generate the orthonormal functions $\{v_i: i = 1, 2, \ldots, 40\}$ consisting of 40 eigenfunctions of the Laplace–Beltrami operator computed on $M_T$. Then, we generate two sets of smooth functional data supported on $M_T$, with identical within-group covariance structures, as follows:

\begin{align*}
\text{Group 1: } x_{i1} &= w_{i1}v_1 + w_{i2}v_2 + \ldots + w_{i40}v_{40} & i &= 1, \ldots, n, \\
\text{Group 2: } x_{i2} &= a_i + u_{i1}v_1 + u_{i2}v_2 + \ldots + u_{i40}v_{40} & i &= 1, \ldots, n,
\end{align*}

(24)
where $u_{ij}$ and $w_{ij}$ are zero-mean independent random variables that represent the scores and are distributed according to a normal distribution with variance $\sigma_j^2$ decreasing in $j$. The function $\mu$ is the groups’ difference, chosen to be a fixed linear combination of the eigenfunctions $\{\nu_l: l = 1, 2, \ldots, 40\}$, and its magnitude is controlled by a parameter $\alpha > 0$ defining the ‘difficulty’ of the classification problem, that is, the signal-to-noise ratio. We then identify three regimes, with $\alpha \in \{0.2, 0.4, 0.6\}$. Given a new function $x^*$, the aim is to recover the group this observation belongs to.

We compare the proposed FLDA method against the models introduced in Section 6.5, e.g. FPCA followed by LDA, Lasso and Ridge logistic regression, random forests, support vector machines, and fully-connected feed-forward neural networks. For standard multivariate models, we use the values of $x$ at the vertices of $M_T$ to construct the data matrix. We evaluate the performance of each method for different sample sizes of the training data: $n = 128, 256, 512, 1024$. For every $n$, besides the training set, we generate a validation set of size $n$ and a test set including 20K samples, and repeat the experiment 50 times. To select the hyperparameters of the models, we employ a validation set approach. We summarize the classification performances on the test set in Figure A1.

Figure A1 shows that by increasing $\alpha$, or the sample size $n$, the performances of all models tend to improve. This is expected, given that a larger $\alpha$ makes the classification task easier. In addition, a
larger sample size allows for a more accurate estimation of the unknown model parameters. In the setting of equal within-class covariance structures, it is well known that the best classifier is linear, so it is not surprising that some of the nonlinear models, e.g. RF, SVM, and NN, show worse performances. In particular, FLDA appears to perform better than the other methods. For large sample sizes, the influence of the regularization terms becomes negligible and the difference in performance starts vanishing.

Next, we explore the performance of the different methods in the setting where the two groups have different within-group covariance structures. Specifically, we generate the data as in equation (24) with \( u_i \) zero-mean independent random variables distributed according to a normal with variance \( \sigma_i^2 \), decreasing in \( i \), and \( w_j \) zero-mean independent random variables distributed according to a normal with variance \( \sigma_j^{2-0} \), and therefore, increasing in \( j \). The results of the simulations are shown in Figure A2.

In this setting, it is well known that the best classifier is not linear. Hence, it is not surprising that the linear methods tend to perform worse than the nonlinear ones and that this difference does not vanish by simply increasing the sample size. Moreover, a larger \( \alpha \) leads to better performance across all the tested methods. Specifically, the proposed FQDA model outperforms all the linear models, but other nonlinear approaches, such as SVM, perform even better. Importantly, the proposed FLDA model performs best among the linear models even when it is misspecified.

**Figure A2.** Results of the simulation study for heterogeneous covariance structures across different sample sizes (n = 128, 256, 512, 1024) and signal-to-noise ratios (\( \alpha = 0.2, 0.4, 0.6 \)), where \( \alpha \) reflects the strength of the discriminant signal. The prediction accuracy was evaluated through AUC and the simulations were repeated 50 times for each setting.
Classification performance is not the only relevant metric. In the application motivating this work, we are particularly interested in accurately estimating the classification rule, which for linear models is fully described by a parameter $\hat{\beta}$. Therefore, in the setting of homogeneous covariances, we compare the performance of the different models with respect to the metric $\|\hat{\beta} - \beta^0\|_{L^2(M)}^2$, with $\hat{\beta}$ an appropriately normalized version of the estimate of the true functional parameter $\beta^0$.

Classification performance is not the only relevant metric. In the application motivating this work, we are particularly interested in accurately estimating the classification rule, which for linear models is fully described by a parameter $\hat{\beta}$. Therefore, in the setting of homogeneous covariances, we compare the performance of the different models with respect to the metric $\|\hat{\beta} - \beta^0\|_{L^2(M)}^2$, with $\hat{\beta}$ an appropriately normalized version of the estimate of the true functional parameter $\beta^0$.

Note that the estimates from a linear discriminant analysis and a logistic regression model can be compared, as they assume the same model but estimate their parameters differently (Hastie et al., 2009). For the standard multivariate methods, $\hat{\beta}$ is constructed by interpolating a piecewise linear function to its estimated discrete counterpart. The results, shown in Figure A3, indicate that the proposed FLDA model does not only yield more accurate predictions but also more accurate estimates of the underlying functional parameters.

**Appendix C. Proofs**

Recall that we denote by $C$ the covariance function of the functional predictor and by $\hat{C}$ its empirical counterpart. Moreover, $K_M$ denotes the evaluation kernel of the Sobolev space $W^2(M)$ endowed with the norm $\left(\|\Delta_M \hat{\beta}\|_{L^2(M)}^2 + \|\hat{\beta}\|_{L^2(M)}^2\right)^{1/2}$. The sandwich operator $T$ is defined as $T = L_{K_M}^{1/2} L_C L_{C, M}^{1/2}$. To
simplify the notation, we drop the subscripts from \( \| \cdot \|_{C^2(M)} \) and \( \langle \cdot, \cdot \rangle_{C^2(M)} \). We define \( \| A \|_{\text{op}} = \sup_{f : \| f \|_1} \| Af \| \) to be the operator norm of a linear operator \( A : L^2(M) \to L^2(M) \).

Thanks to the fact that \( L^2_{K,M}(L^2(M)) = W^2(M) \) \cite{Cucker2002}, it is clear that we can reformulate the problem in equation (10) as

\[
\min_{\hat{f}} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \langle x_i, L^{\frac{1}{2}}_{K,M} \hat{f} \rangle \right)^2 + \lambda \| f \|_2^2,
\]

whose solution \( \hat{f}_i \) is given by

\[
\hat{f}_i = \left( T_n + \lambda I \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} y_i L^{\frac{1}{2}}_{K,M} x_i,
\]

where \( T_n = L^{\frac{1}{2}}_{K,M} L_n L^{\frac{1}{2}}_{K,M} \). Moreover, the out-of-sample risk \( \mathbb{E}^*[(X^*, \hat{\beta}^0 - \beta)]^2 \) can be easily rewritten as \( \| T^2(\hat{f}_i - f_0) \|_2^2 \), with \( f_0 \in L^2(M) \) such that \( f_0 = T^{-1} L^{\frac{1}{2}}_{K,M}(\mu_2 - \mu_1) \).

Observe that

\[
T^2(\hat{f}_i - f_0) = T^2 \left[ (T_n + \lambda I)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} y_i L^{\frac{1}{2}}_{K,M} x_i \right) - f_0 \right]
= T^2 \left[ (T_n + \lambda I)^{-1} L^{\frac{1}{2}}_{K,M} \left( \frac{1}{n} \sum_{i=1}^{n} y_i x_i - (\mu_2 - \mu_1) \right) \right]
+ (T_n + \lambda I)^{-1} L^{\frac{1}{2}}_{K,M} (\mu_2 - \mu_1) - T^{-1} L^{\frac{1}{2}}_{K,M} (\mu_2 - \mu_1)
\]

Let \( \hat{d} = \frac{1}{n} \sum_{i=1}^{n} y_i x_i \) and \( d = \mu_2 - \mu_1 \), and notice that \( \mathbb{E}[\hat{d}] = d + o(1) \). Then, we have

\[
\| T^2(\hat{f}_i - f_0) \| \leq \| T^2(\hat{d} - d) \| + \| T^2(\hat{f}_i - f_0) \| \| T^2(\hat{d} - d) \|
= I_1 + I_2
\]

We first derive a bound for the term \( I_1 \) in equation (27), i.e. the variance term, and then proceed with bounding the term \( I_2 \), i.e. the bias term. To accomplish this, we will also use Theorems C.2–C.3, stated in Section C.4.

C.1 Variance

Simple calculations show that

\[
I_1 = \| T^2(T_n + \lambda I)^{-1} L^{\frac{1}{2}}_{K,M} (\hat{d} - d) \|
\leq \| T^2(T_n + \lambda I)^{-\frac{1}{2}} \|_{\text{op}} \| (T_n + \lambda I)^{-\frac{1}{2}} (T + \lambda I)^{-\frac{1}{2}} \|_{\text{op}} \| (T + \lambda I)^{-\frac{1}{2}} L^{\frac{1}{2}}_{K,M} (\hat{d} - d) \|
\leq \| T^2(T_n + \lambda I)^{-\frac{1}{2}} \|_{\text{op}} \| (T_n + \lambda I)^{-\frac{1}{2}} (T + \lambda I)^{-\frac{1}{2}} \|_{\text{op}} \| (T + \lambda I)^{-\frac{1}{2}} L^{\frac{1}{2}}_{K,M} L^{\frac{1}{2}}_{K,M} (\hat{d} - d) \|
\leq \| T^2(T_n + \lambda I)^{-\frac{1}{2}} \|_{\text{op}} \| (T + \lambda I)^{-\frac{1}{2}} L^{\frac{1}{2}}_{K,M} (\hat{d} - d) \|
\]
where in the last inequality we have used \(\|A^\top B\|_\text{op} \leq \|AB\|_\text{op}^\gamma\) for any \(0 < \gamma < 1\) (Blanchard & Krämer, 2010). We can then bound the first term thanks to Theorem C.3. We therefore turn to the second term.

First observe that

\[
\mathbb{E}[\hat{d}] = d + o(1), \quad \mathbb{E}[(\hat{d} - d) \otimes (\hat{d} - d)] = \frac{1}{n} (C - \pi_1 \mu_1 \otimes \mu_1 - \pi_2 \mu_2 \otimes \mu_2) + o(1),
\]

and, therefore,

\[
\langle L_{K,M}^{\frac{1}{2}} (\hat{d} - d), \eta_k \rangle^2 = \langle L_{K,M}^{\frac{1}{2}} L_{\hat{d} - d \otimes \hat{d} - d}^{\frac{1}{2}} \rangle L_{K,M}^{\frac{1}{2}} \eta_k, \eta_k \rangle
\]

\[
= \frac{1}{n} \left( T \eta_k, \eta_k \right) - \frac{\pi_1}{n} \left( L_{K,M}^{\frac{1}{2}} L_{\mu_1 \otimes \mu_1} L_{K,M}^{\frac{1}{2}} \eta_k, \eta_k \right) - \frac{\pi_2}{n} \left( L_{K,M}^{\frac{1}{2}} L_{\mu_2 \otimes \mu_2} L_{K,M}^{\frac{1}{2}} \eta_k, \eta_k \right)
\]

\[
= \frac{1}{n} \left( \tau_k - \pi_1 \| L_{K,M}^{\frac{1}{2}} \mu_1 \|^2 - \pi_2 \| L_{K,M}^{\frac{1}{2}} \mu_2 \|^2 \right),
\]

with \(\{\tau_k\}\) and \(\{\eta_k\}\) denoting the eigenvalues and eigenfunctions of \(T\), respectively. As in Gaynanova and Kolar (2015), we ignore the bias term \(o(1)\).

We then have

\[
\mathbb{E}\left\| (T + \lambda I)^{-\frac{1}{2}} L_{K,M}^{\frac{1}{2}} (\hat{d} - d) \right\|^2 = \mathbb{E}\left\| \sum_k (T + \lambda I)^{-\frac{1}{2}} \eta_k (L_{K,M}^{\frac{1}{2}} (\hat{d} - d), \eta_k) \right\|^2
\]

\[
\leq \frac{1}{n} \sum_k \left( \frac{\tau_k - \pi_1 \| L_{K,M}^{\frac{1}{2}} \mu_1 \|^2 - \pi_2 \| L_{K,M}^{\frac{1}{2}} \mu_2 \|^2}{\lambda + \tau_k} \right) \leq \frac{1}{n} D(\lambda).
\]

Moreover, by appealing to the Markov inequality, we have that with confidence at least \(1 - \delta/2\)

\[
\left\| (T + \lambda I)^{-\frac{1}{2}} L_{K,M}^{\frac{1}{2}} (\hat{d} - d) \right\| \leq \frac{2}{\delta} \sqrt{\frac{D(\lambda)}{n}} \leq \frac{1}{\kappa \delta} B_{n,\lambda},
\]

where \(B_{n,\lambda} = \frac{2\pi}{\sqrt{n}} \left( \frac{\lambda}{\lambda + \kappa} + \sqrt{D(\lambda)} \right)\) and \(\kappa\) is defined such that \(\kappa^2 = \text{ess sup} \| L_{K,M}^{1/2} X \|^2\).

Thanks to the inequality in equation (28) and Theorem C.3, with probability at least \(1 - \delta\), we have

\[
I_1 = \left\| T^\frac{1}{2} (T_n + \lambda I)^{-1} L_{K,M}^{\frac{1}{2}} \left( \frac{1}{n} \sum_{i=1}^n y_i x_i - d \right) \right\| \leq \left( \frac{B_{n,\lambda} \log(2/\delta)}{\lambda} + 1 \right)^2 \left( \frac{1}{2\kappa \delta} B_{n,\lambda} \right).
\]

We now turn to the bias term \(I_2\) in equation (27).

C.2 Bias

By simple calculations, we have

\[
I_2 = \left\| T^\frac{1}{2} \left[ (T_n + \lambda I)^{-1} L_{K,M}^{\frac{1}{2}} d - T^{-1} L_{K,M}^{\frac{1}{2}} d \right] \right\|
\]

\[
= \left\| T^\frac{1}{2} (T_n + \lambda I)^{-1} [T - T_n + \lambda I] \right\|_\text{op} \| f_0 \|
\]

\[
\leq \left\| (T + \lambda I)^{-\frac{1}{2}} (T_n + \lambda I)^{-\frac{1}{2}} \right\|_\text{op} \left\| (T_n + \lambda I)^{-\frac{1}{2}} [T - T_n + \lambda I] \right\|_\text{op} \| f_0 \|
\]

\[
\leq \left\| (T + \lambda I)[T_n + \lambda I]^{-1} \right\|_\text{op} \left\| (T + \lambda I)^{-\frac{1}{2}} [T - T_n] \right\|_\text{op} \| f_0 \|
\]

\[
\leq \left\| (T + \lambda I)[T_n + \lambda I]^{-1} \right\|_\text{op} \left\| (T + \lambda I)^{-\frac{1}{2}} [T - T_n] \right\|_\text{op} \| f_0 \|
\]

\[
+ \left\| (T + \lambda I)[T_n + \lambda I]^{-1} \right\|_\text{op} \left\| (T + \lambda I)^{-\frac{1}{2}} \right\|_\text{op} \| f_0 \|.
\]
Moreover, by using Theorems C.2–C.3, and the inequality
\[
\lambda \|(T + \lambda I)^{-\frac{1}{2}}\|_{op} \leq \frac{\lambda}{\sqrt{\lambda}} = \sqrt{\lambda},
\]  
we have that with probability at least 1 − δ
\[
\|T^2[(T_n + \lambda I)^{-1} L_{K,M}^1 d - T^{-1} L_{K,M}^1 d] \| \leq \left( \frac{B_{n,\lambda} \log (4/\delta)}{\sqrt{\lambda}} + 1 \right)^2 (B_{n,\lambda} \log (4/\delta) + \sqrt{\lambda}) \|f_0\|. 
\]  

C.3 Final rates
Using the variance and bias bounds in equations (29) and (31), we get the following bound for the out-of-sample risk:
\[
\|T^2(\hat{f}_\lambda - f_0)\|^2
\]  
\[
\leq 2 \|T^2(T_n + \lambda I)^{-1} L_{K,M}^1 (\hat{d} - d)\|^2 + 2 \|T^2[(T_n + \lambda I)^{-1} L_{K,M}^1 d - T^{-1} L_{K,M}^1 d] \|^2 
\]  
\[
\leq 2 \left( \frac{B_{n,\lambda} \log (2/\delta)}{\sqrt{\lambda}} + 1 \right)^4 \left( \frac{1}{2\kappa \delta} B_{n,\lambda} \right)^2 
\]  
\[
+ 2 \left( \frac{B_{n,\lambda} \log (4/\delta)}{\sqrt{\lambda}} + 1 \right)^4 \left( B_{n,\lambda} \log (4/\delta) + \sqrt{\lambda} \right)^2 \|f_0\|^2 
\]  
\[
\leq 2 \frac{\lambda}{\delta^2} \left( \frac{B_{n,\lambda} \log (2/\delta)}{\sqrt{\lambda}} + 1 \right)^4 \left( \frac{1}{2\kappa \sqrt{\lambda}} B_{n,\lambda} \right)^2 
\]  
\[
+ 2\lambda \left( \frac{B_{n,\lambda} \log (4/\delta)}{\sqrt{\lambda}} + 1 \right)^4 \left( \frac{B_{n,\lambda} \log (4/\delta) + 1}{\sqrt{\lambda}} \right)^2 \|f_0\|^2 
\]  
\[
\leq C \frac{(\log (4/\delta))^6}{\delta^2} n^{-\frac{1}{m}},
\]  
where in the last inequality we have chosen \( \lambda = n^{-\frac{1}{m}} \) and used the inequality \( B_{n,\lambda} \leq 2\kappa(\kappa + \sqrt{\lambda}) \sqrt{\lambda} \).

The constant \( c \) here is from Assumption 3.3.

This implies the result stated in Theorem 3.1.

C.4 Auxiliary results

**Theorem C.1**  Let \( \mathcal{H} \) be a Hilbert space endowed with a norm \( \|\cdot\|_\mathcal{H} \) and let \( X \) be a random variable taking values in \( \mathcal{H} \). Let \( \zeta_1, \ldots, \zeta_n \) be a sequence of \( n \) independent
copies of $X$. Assume that $\|\xi\|_H \leq M$ (a.s.), then for $0 < \delta < 1$
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mathbb{E}[\xi_i]) \right\|_H^2 \leq \frac{2M \log (2/\delta)}{n} + \sqrt{\frac{2\mathbb{E}[\|\xi\|_H^2] \log (2/\delta)}{n}}
\]
with probability at least $1 - \delta$.

**Proof.** A proof can be found in Pinelis (2007).

Using Theorem C.1, it can be shown that the following two theorems hold.

**Theorem C.2** Under Assumption 3.1, for any $0 < \delta < 1$, the inequality
\[
\| (T + \lambda I)^{-\frac{1}{2}}(T - T_n) \|_{op} \leq B_{n,\lambda} \log (2/\delta)
\]
holds with confidence at least $1 - \delta$.

**Proof.** A proof can be found in Tong and Ng (2018).

**Theorem C.3** Under Assumption 3.1, for any $0 < \delta < 1$ with confidence at least $1 - \delta$,
\[
\| (T + \lambda I)(T_n + \lambda I)^{-1} \|_{op} \leq \left( \frac{B_{n,\lambda} \log (2/\delta)}{\sqrt{\lambda}} + 1 \right)^2.
\]
Moreover, the confidence set is the same as the one in Theorem C.2.

**Proof.** A proof can be found in Tong and Ng (2018).

C.5 Multivariate model

Thanks to the multivariate notation introduced in Section 4, we can reformulate the multivariate model in equation (15) as
\[
\text{minimize } \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \langle x_i, L_{K} f \rangle_H \right)^2 + \lambda \| f \|_{H}^2,
\]
with $x_i := (v_i, x_i)$. Therefore, the proof of Theorem 4.1 follows the same lines as that of Theorem 3.1 and is therefore omitted.

**References**

Arguillé S., Miller M. I., & Younes L. (2016). Diffeomorphic surface registration with atrophy constraints. *SIAM Journal on Imaging Sciences*, 9(3), 975–1003. https://doi.org/10.1137/15M104431X

Berliner A., & Thomas-Agnan C (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer US. https://doi.org/10.1007/978-1-4419-9096-9

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

Biffi C., De Marvao A., Attard M. I., Dawes T. J., Whiffin N., Bai W., Shi W., Francis C., Meyer H., Buchan R., Cook S. A., Rueckert D., & O’Regan D. P. (2018). Three-dimensional cardiovascular imaging-genetics: A mass univariate framework. *Bioinformatics*, 34(1), 97–103. https://doi.org/10.1093/bioinformatics/btx352

Blanchard, G., & Krämer, N. (2010). Optimal learning rates for kernel conjugate gradient regression. In *Proceedings of the 23rd international conference on neural information processing systems – volume 1 (NIPS’10)* (pp. 226–234). Curran Associates Inc., Red Hook, NY, USA. https://doi.org/10.5555/2997189.2997215
Bondareff W., Mountjoy C. Q., Roth M., & Hauser D. L. (1989). Neurofibrillar degeneration and neuronal loss in Alzheimer’s disease. *Neurobiology of Aging*, 10(6), 709–715. https://doi.org/10.1016/0197-4580(89)90007-9

Braak H., Alafuzoff I., Arzberger T., Kretzschmar H., & Tredici K. (2006). Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathologica*, 112(4), 389–404. https://doi.org/10.1007/s00401-006-0127-z

Breiman L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiz H (2011). *Functional analysis, Sobolev spaces and partial differential equations*. Springer New York. https://doi.org/10.1007/978-0-387-37914-7

Bruveris M., Risser L., & Vialard F. X. (2012). Mixture of kernels and iterated semidirect product of diffeomorphisms groups. *Multiscale Modeling and Simulation*, 10(4), 1344–1368. https://doi.org/10.1137/110846324

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

Charlier B., Charon N., & Trouvé A. (2017). The fshape framework for the variability analysis of functional shapes. *Foundations of Computational Mathematics*, 17(2), 287–357. https://doi.org/10.1007/s10208-015-9288-2

Charlier, B., Nardi, G., & Trouvé, A. (2015). The matching problem between functional shapes via a BV-penalty term: A F-convergence result. arXiv preprint arXiv:1503.07685

Chen L. H., & Jiang C. R. (2018). Sensible functional linear discriminant analysis. *Computational Statistics and Data Analysis*, 126, 39–52. https://doi.org/10.1016/j.csda.2018.04.005

Chung, M. K., Hanson, J. L., & Pollak, S. D. (2016). Statistical analysis on brain surfaces. *Handbook of Neuroimaging Data Analysis*, 233, 46–57. https://doi.org/10.1201/9781315373652-20

Chung, M. K., Hartley, R., Dalton, K. M., & Davidson, R. J. (2008). Encoding cortical surface by spherical harmonics. *Statistica Sinica*, 18(4), 1269–1291. https://www.jstor.org/stable/24308355

Chung M. K., Qiu A., Seo S., & Vorperian H. K. (2015). Unified heat kernel regression for diffusion, kernel smoothing and wavelets on manifolds and its application to mandible growth modeling in CT images. *Medical Image Analysis*, 22(1), 63–76. https://doi.org/10.1016/j.media.2015.02.003

Cortes C., & Vapnik V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cucker F., & Smale S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1), 1–49. https://doi.org/10.1090/S0273-0979-02-00922-2

Cury C., Glaunès J. A., & Colliot O. (2014). Diffeomorphic iterative centroid methods for template estimation on large datasets. In F. Nielsen (Ed), *Geometric Theory of Information* (pp. 273–299). Springer International Publishing. https://doi.org/10.1007/978-3-319-05317-2_10

Dai X., & Müller H. G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Annals of Statistics*, 46(6B), 3334–3361. https://doi.org/10.1214/17-AOS1660

Dai X., Müller H. G., & Yao F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, 104(3), 545–560. https://doi.org/10.1093/biomet/asx024

Dale A. M., Fischl B., & Sereno M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Delailge A., & Hall P. (2012). 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

Dickerson B. C., Bakkour A., Salat D. H., Feczko E., Pacheco J., Greve D. N., Grodstein F., Wright C. L., Blacker D., Rosas H. D., Sperling R. A., Atri A., Growdon J. H., Hyman B. T., Morris J. C., Fischl B., & Buckner R. L. (2009). The cortical signature of Alzheimer’s disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cerebral Cortex*, 19(3), 497–510. https://doi.org/10.1093/cercor/bhn113

Dong Q., Zhang W., Stonnington C. M., Wu J., Gutman B. A., Chen K., Su Y., Baxter L. C., Thompson P. M., Reiman E. M., Caselli R. J., & Wang Y. (2020). Applying surface-based morphometry to study ventricular abnormalities of cognitively unimpaired subjects prior to clinically significant memory decline. *NeuroImage: Clinical*, 27, 102338. https://doi.org/10.1016/j.nicl.2020.102338

Dong Q., Zhang W., Wu J., Li B., Schron E. H., McMahon T., Shi J., Gutman B. A., Chen K., Baxter L. C., Thompson P. M., Reiman E. M., Caselli R. J., & Wang Y. (2019). Applying surface-based hippocampal morphometry to study APOE-E4 allele dose effects in cognitively unimpaired subjects. *NeuroImage: Clinical*, 22, 101744. https://doi.org/10.1016/j.nicl.2019.101744

Dryden I. L., & Mardia K. V. (2016). *Statistical shape analysis, with applications in R* (2nd ed.). John Wiley & Sons, Ltd. https://doi.org/10.1007/9781119072492

Dube P., & Müller H. G. (2020). Functional models for time-varying random objects. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(2), 275–327. https://doi.org/10.1111/rssb.12337
Epifanio I., & Ventura-Campos N. (2014). Hippocampal shape analysis in Alzheimer’s disease using functional data analysis. *Statistics in Medicine*, 33(5), 867–880. https://doi.org/10.1002/sim.v33.5

Fasshauer G. E., & Ye Q. (2013). Reproducing kernels of Sobolev spaces via a green kernel approach with differential operators and boundary operators. *Advances in Computational Mathematics*, 38(4), 891. https://doi.org/10.1007/s10444-011-9264-6

Feng X., Li T., Song X., & Zhu H. (2020). Bayesian scalar on image regression with nonignorable nonresponse. *Journal of the American Statistical Association*, 115(532), 1574–1597. https://doi.org/10.1080/01621459.2019.1686391

Feragen A., Lauze F., & Hauberg S. (2015). Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 07-12-June, pp. 3032–3042), IEEE. https://doi.org/10.1109/CVPR.2015.7298922

Ferrand 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

Ferraty F., & Vieu P. (2006). *Nonparametric functional data analysis*. Springer New York. https://doi.org/10.1007/10.1007-3-87662-20-2

Feydy J., Charlier B., Vialard F. X., & Peyré G. (2017). Optimal transport for diffeomorphic registration. In M. Ferrard, L. Maier-Hein, A. Franz, P. Jannin, D. Collins, & S. Duchesne (Eds.), *Lecture Notes in Computer Science* (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, LNCS) (Vol. 10433, pp. 291–299), Springer. https://doi.org/10.1007/978-3-319-66182-7_34

Fischl B., & Dale A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20), 11050–11055. https://doi.org/10.1073/pnas.200033797

Fischl B., Sereno M. I., & Dale A. M. (1999). Cortical surface-based analysis: II. Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2), 195–207. https://doi.org/10.1006/nimg.1999.0396

Fried I., & Malkus D. S. (1975). Finite element mass matrix lumping by numerical integration with no convergence rate loss. *International Journal of Solids and Structures*, 11(4), 461–466. https://doi.org/10.1016/0020-7683(75)90081-5

Gaynanova I. (2020). Prediction and estimation consistency of sparse multi-class penalized optimal scoring. *Bernoulli*, 26(1), 286–322. https://doi.org/10.3150/19-BEJ1126

Gaynanova I., & Kolar M. (2015). Optimal variable selection in multi-group sparse discriminant analysis. *Electronic Journal of Statistics*, 9(2), 2007–2034. https://doi.org/10.1214/15-EJS1064

Gaynanova I., & Wang T. (2019). Sparse quadratic classification rules via linear dimension reduction. *Journal of Multivariate Analysis*, 169, 278–299. https://doi.org/10.1016/j.jmva.2018.09.011

Gee A. H., Treece G. M., & Poole K. E. (2018). How does the femoral cortex depend on bone shape? A method for the joint analysis of surface texture and shape. In *Advances in Computational Anatomy* for the joint analysis of surface texture and shape (Vol. LNCS, pp. 3032–3042), IEEE. https://doi.org/10.1109/CVPR.2015.7298922

Ghahramani Z., & Hinton G. E. (2005). Cooperative co-actor learning. *Neurocomputing*, 65(1–2), 5–15. https://doi.org/10.1016/j.neucom.2004.04.051

Goh W., & Lee C. S. (2018). On the dichotomy of variational learning. *Advances in Neural Information Processing Systems*, 31, 213–223. https://doi.org/10.1109/978-1-5386-4651-7.2018.874251

Gorriz J. M., Mora-García J., Criste V., & García-Martínez C. (2017). Sparse quadratic classification rules via linear dimension reduction. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11), 2502–2513. https://doi.org/10.1109/TNNLS.2016.2618262

Graglia R. D., & Otis T. E. (2008). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 103(482), 1255–1270. https://doi.org/10.1198/016214508000000502

Hastie T., Tibshirani R., & Friedman J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-36620-2

Hazlett H. C., Gu H., Munsell B. C., Kim S. H., Styner M., Wolff J. J., Elison J. T., Swanson M. R., Zhu H., Botteron K. N., Collins D. L., Constantino J. N., Dager S. R., Evans A. C., Fonov V. S., Gerig G., Kostopoulos P., McKinstry R. C., …Piven J. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature*, 542(7641), 348–351. https://doi.org/10.1038/nature21369

Hinton E., Rock T., & Zienkiewicz O. C. (1976). A note on mass lumping and related processes in the finite element method. *Earthquake Engineering & Structural Dynamics*, 4(3), 245–249. https://doi.org/10.1002/eqe.v4:3

Hoerl A. E., & Kennard R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Horváth L., & Kokoszka P. (2012). *Inference for functional data with applications* (Vol. Springer Series in Statistics, 200). Springer New York. https://doi.org/10.1007/978-1-4614-3655-3
Zhang, Z., Wu, Y., Xiong, D., Ibrahim, J. G., Srivastava, A., & Zhu, H. (2022). LESA: Longitudinal elastic shape analysis of brain subcortical structures. *Journal of the American Statistical Association, 118*(541), 3–17. https://doi.org/10.1080/01621459.2022.2126363

Zhu H., Li T., & Zhao B. (2023). Statistical learning methods for neuroimaging data analysis with applications. *Annual Review of Biomedical Data Science, 6*(1), 73–104. https://doi.org/10.1146/biodatasci.2023.6.issue-1