Biometrika (2019), **106**, 1, *pp.* 161–180
Printed in Great Britain
doi: 10.1093/biomet/asy060
Advance Access publication 17 December 2018

Classification of functional fragments by regularized linear classifiers with domain selection

By DAVID KRAUS

Department of Mathematics and Statistics, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic
david.kraus@mail.muni.cz

and MARCO STEFANUCCI

Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Roma, Italy
marco.stefanucci@uniroma1.it

Summary

We consider classification of functional data into two groups by linear classifiers based on one-dimensional projections of functions. We reformulate the task of finding the best classifier as an optimization problem and solve it by the conjugate gradient method with early stopping, the principal component method, and the ridge method. We study the empirical version with finite training samples consisting of incomplete functions observed on different subsets of the domain and show that the optimal, possibly zero, misclassification probability can be achieved in the limit along a possibly nonconvergent empirical regularization path. We propose a domain extension and selection procedure that finds the best domain beyond the common observation domain of all curves. In a simulation study we compare the different regularization methods and investigate the performance of domain selection. Our method is illustrated on a medical dataset, where we observe a substantial improvement of classification accuracy due to domain extension.

Some key words: Classification; Conjugate gradient; Domain selection; Functional data; Partial observation; Regularization; Ridge method.

1. Introduction

We consider classification of a functional observation into one of two groups. Classification of functional data is a rich, longstanding topic and is comprehensively surveyed in Baíllo et al. (2011b). Delaigle & Hall (2012a) showed that depending on the relative geometric positions of the difference of the group means, representing the signal, and the covariance operator, summarizing the structure of the noise, certain classifiers can have zero misclassification probability. This remarkable phenomenon, called perfect classification, is a special property of the infinite-dimensional setting and cannot occur in the multivariate context, except in degenerate cases. Delaigle & Hall (2012a) showed that a particularly simple class of linear classifiers, based on a carefully chosen one-dimensional projection of the function to be classified, can achieve this optimal error rate either exactly or in the limit along a sequence of approximations. Berrendero et al. (2018) further elucidated the perfect classification phenomenon from the point

of view of the Feldman–Hájek dichotomy between mutual singularity and absolute continuity of two Gaussian measures on abstract spaces with respect to each other.

Motivated by these findings, we reformulate the problem of determining the best classifier as a quadratic optimization problem on a function space or, equivalently, a linear inverse problem. These problems are ill-posed; however, unlike with most inverse problems, this is not a complication but rather an advantage in the sense that the more ill-posed the problem is, the better the optimal misclassification probability. We use regularization techniques, such as the method of conjugate gradients with early stopping and ridge regularization, to solve the optimization problem, obtaining a class of regularized linear classifiers. The optimal misclassification rate is the limit along the regularization path of solutions which themselves may not converge.

We study the empirical version of the problem, where the objective function in the constrained minimization must be estimated from finite training data, and make two contributions. First, we show that it is possible to construct an empirical regularization path towards the possibly nonexistent unconstrained solution such that the classification error converges to its best value, possibly zero. We do this for conjugate gradient, principal component and ridge classification in a truly infinite-dimensional manner, in the sense that the convergence takes place along a path with decreasing regularization and holds without restrictions on the mean difference between classes. Second, all our methods and theory are developed in the setting of partially observed functional data, where trajectories are observed only on subsets of the domain. This type of incomplete data, also called functional fragments, is increasingly common in applications; see, for example, Bugni (2012), Delaigle & Hall (2013), Liebl (2013), Goldberg et al. (2014), Kraus (2015), Delaigle & Hall (2016) and Gromenko et al. (2017). The principal difficulty for inference with fragments is that temporal averaging is precluded by the incompleteness of the observed functions. Our formulation as an optimization problem enables us to overcome this issue under certain assumptions, because only averaging across individuals in the training data is needed, and not individual curves.

Since the observation domains may vary in the training sample and the new curve to be classified may be observed on a different subset, it is natural to ask which domain should be used. We propose a domain selection strategy that looks for the best classifier with domain ranging from a minimum common domain to the entire domain of the function to be classified. For various methods of selecting the best observation points, see Ferraty et al. (2010), Delaigle et al. (2012), Pini & Vantini (2016), Berrendero et al. (2018) and Stefanucci et al. (2018).

Our simulation study confirms that domain selection can considerably reduce the misclassification rate. Further simulations compare the performances of the three types of regularization. Among other findings, this study shows that the principal component and conjugate gradient classifiers often achieve comparable error rates but that the latter usually needs a lower dimension of the regularization subspace, in agreement with a theoretical result we provide.

Application to a dataset on the geometric features of the internal carotid artery in patients with and without aneurysm demonstrates the utility of our proposed approach. These data consist of trajectories observed on intervals of different lengths. Previous analyses of the data used the common domain of all curves in classification. With our results we can include information beyond this minimum domain, which leads to a substantial drop in the error rate of discrimination between risk groups.

General references on functional data analysis include Ramsay & Silverman (2005) and Horváth & Kokoszka (2012). Further relevant references are Cuesta-Albertos et al. (2007) for other methods based on one-dimensional projections, Berrendero et al. (2016) for variable selection in classification, Bongiorno & Goia (2016) and Dai et al. (2017) for classification beyond the Gaussian setting, and Cuevas (2014) for an overview.

2. Regularized linear classification

2.1. *Projection classifiers*

We regard functional observations as random elements of the separable Hilbert space $L^2(\mathcal{I})$ of square-integrable functions on a compact domain $\mathcal{I}$ equipped with inner product $\langle f,g\rangle = \int_{\mathcal{I}} f(t)g(t)\,\mathrm{d}t$ and norm $\|f\| = \langle f,f\rangle^{1/2}$. In most applications $\mathcal{I}$ is an interval and the observations are curves, but our results can be extended to other objects, such as surfaces or images. We consider classification of a Gaussian random function, X, into one of two groups of Gaussian random functions: group 0 has mean μ_0; group 1 has mean μ_1. Both groups have covariance operator $\mathscr{R}$ defined as the integral operator

$$(\mathscr{R}f)(\cdot) = \int_{\mathcal{I}} \rho(\cdot\,,t)f(t)\,\mathrm{d}t$$

with kernel $\rho(s,t) = \mathrm{cov}\{X(s),X(t)\}$. In this section we assume that μ_0, μ_1 and $\mathscr{R}$ are known, which corresponds to the asymptotic situation with an infinite training sample. To simplify the presentation we assume throughout the paper that the new observation to be classified may come from either of the two classes with equal prior probability. The general case is treated in the Supplementary Material.

Like Delaigle & Hall (2012a) we consider the class of centroid classifiers that are based on one-dimensional projections of the form $\langle X,\psi\rangle$, where ψ is a function in $L^2(\mathcal{I})$. If X belongs to group j ($j = 0,1$), the distribution of $\langle X,\psi\rangle$ is normal with mean $\langle \mu_j,\psi\rangle$ and variance $\langle \psi,\mathscr{R}\psi\rangle$. Denote the corresponding Gaussian densities by $f_{\psi,j}$. The optimal classifier based on $\langle X,\psi\rangle$ assigns X to the class $C_\psi(X)$ given by

$$C_\psi(X) = 1_{\{f_{\psi,1}(\langle X,\psi\rangle)/f_{\psi,0}(\langle X,\psi\rangle)>1\}} = 1_{\{\langle X-\mu_0,\psi\rangle^2-\langle X-\mu_1,\psi\rangle^2>0\}} = 1_{\{T_\psi(X)>0\}},$$

where $T_\psi(X) = \langle X-\bar\mu,\psi\rangle\langle \mu,\psi\rangle$ with $\bar\mu = (\mu_0+\mu_1)/2$ and $\mu = \mu_1-\mu_0$. The misclassification probability of this classifier is

$$\begin{aligned} D(\psi) &= P_0\{C_\psi(X)=1\}/2 + P_1\{C_\psi(X)=0\}/2 = P_0(\langle X-\bar\mu,\psi\rangle\langle \mu,\psi\rangle > 0) \\ &= P_0(\langle X-\mu_0,\psi\rangle > |\langle \mu,\psi\rangle|/2) = 1-\Phi\left(\frac{|\langle \mu,\psi\rangle|}{2\langle \psi,\mathscr{R}\psi\rangle^{1/2}}\right), \end{aligned} \tag{1}$$

where P_j is the distribution of curves in group j and Φ is the standard normal cumulative distribution function.

To find the best function ψ, one would ideally like to maximize $|Z(\psi)|$, where

$$Z(\psi) = \frac{\langle \mu,\psi\rangle}{\langle \psi,\mathscr{R}\psi\rangle^{1/2}}.$$

Similarly to Delaigle & Hall (2012a) and Berrendero et al. (2018), we see that if $\|\mathscr{R}^{-1/2}\mu\| < \infty$, then by the Cauchy–Schwarz inequality,

$$\frac{|\langle \mu,\psi\rangle|}{\langle \psi,\mathscr{R}\psi\rangle^{1/2}} = \frac{|\langle \mathscr{R}^{-1/2}\mu,\mathscr{R}^{1/2}\psi\rangle|}{\langle \psi,\mathscr{R}\psi\rangle^{1/2}} \leqslant \frac{\|\mathscr{R}^{-1/2}\mu\|\|\mathscr{R}^{1/2}\psi\|}{\langle \psi,\mathscr{R}\psi\rangle^{1/2}} = \|\mathscr{R}^{-1/2}\mu\|. \tag{2}$$

If, moreover, $\|\mathscr{R}^{-1}\mu\| < \infty$, then the equality is achieved for $\psi = \mathscr{R}^{-1}\mu$. For this choice of ψ, or any multiple of it, the probability of misclassification is $1-\Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$, which is positive due

to the finiteness of $\|\mathscr{R}^{-1/2}\mu\|$, which can be seen as the signal-to-noise ratio. If $\|\mathscr{R}^{-1/2}\mu\| < \infty$, then regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$ or not, two Gaussian measures with mean difference μ and covariances $\mathscr{R}$ are mutually absolutely continuous and $1-\Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ is the Bayes error for distinguishing them, i.e., the lowest possible misclassification probability for this problem among all possible classifiers (Berrendero et al., 2018). If $\|\mathscr{R}^{-1/2}\mu\| < \infty$ but $\|\mathscr{R}^{-1}\mu\| = \infty$, then the Bayes risk cannot be achieved by a projection classifier based on a bounded linear functional of the form $\langle X, \psi\rangle$ for some $\psi \in L^2(\mathcal{I})$. One can, however, use the theory of reproducing kernel Hilbert spaces to define a linear classifier that achieves the Bayes risk. We do not pursue this line of development here because, as will be seen in § 2.2, approximations in the form of projections can asymptotically achieve the Bayes risk.

The maximization of $|Z(\psi)|$ can be solved as the task of maximizing $\langle \mu, \psi\rangle$ subject to $\langle \psi, \mathscr{R}\psi\rangle = 1$. Using Lagrange multipliers $\langle \mu, \psi\rangle + \lambda(1 - \langle \psi, \mathscr{R}\psi\rangle)$ and taking the Fréchet derivative with respect to ψ, one obtains the equation $2\lambda\mathscr{R}\psi = \mu$. Solutions for all $\lambda > 0$, if they exist, i.e., if $\|\mathscr{R}^{-1}\mu\| < \infty$, yield the same optimal misclassification probability. Without loss of generality we take $\lambda = 1/2$. Thus, minimizing the error rate translates into the unconstrained quadratic optimization problem to maximize $\langle \mu, \psi\rangle - \langle \psi, \mathscr{R}\psi\rangle/2$, or

$$\text{minimize } \langle \psi, \mathscr{R}\psi\rangle/2 - \langle \mu, \psi\rangle, \tag{3}$$

i.e., into the linear problem $\mathscr{R}\psi = \mu$.

2.2. *Regularization*

If $\psi = \mathscr{R}^{-1}\mu$ does not exist in $L^2(\mathcal{I})$, i.e., $\|\mathscr{R}^{-1}\mu\| = \infty$, there is no maximizer of $|Z(\psi)|$. One can instead consider an approximating, regularized problem that can be solved. Regularization is typically used to solve, in a stable way, ill-posed inverse problems for which a solution exists. In such contexts, the path of regularized solutions converges to the solution to the problem of interest. Here it may be that no solution exists, but paths of regularized solutions towards the possibly nonexistent solution still turn out to be useful, since the misclassification probability converges to the optimal value along these paths.

If a solution exists, one can approximate it by an iterative numerical method. This approach can also be used when no solution exists. The idea is to construct a sequence of iterations of an appropriate numerical optimization method. The number of steps taken along this divergent sequence towards the nonexistent solution can be seen as a regularization parameter. The conjugate gradient method is particularly suitable for this situation.

The first m steps of the conjugate gradient method applied to the linear inverse problem $\mathscr{R}\psi = \mu$, or equivalently to the minimization of the quadratic functional $\langle \psi, \mathscr{R}\psi\rangle/2 - \langle \mu, \psi\rangle$, are described in Algorithm 1. This formulation is based on the multivariate version in Phatak & de Hoog (2002, § 5), where one can find further references and details on how applying the conjugate gradient method to the normal equations in linear regression leads to partial least squares regression. The functions v_j are conjugate directions in the sense that $\langle v_j, \mathscr{R}v_k\rangle = 0$ for $j \neq k$, and the functions ζ_j are called residuals in numerical analysis and are orthogonal, i.e., $\langle \zeta_j, \zeta_k\rangle = 0$ for $j \neq k$. In step j, the algorithm moves from the current approximate solution $\hat{\psi}_j^{\mathrm{CG}}$ along the conjugate direction v_j with step length h_j that minimizes the quadratic objective. The residual is then updated to ζ_{j+1}. The new conjugate direction v_{j+1} is obtained by projecting the residual ζ_{j+1} onto the orthogonal complement of the span of the previous conjugate directions, where orthogonality is in the sense of the inner product $\langle \cdot, \mathscr{R}(\cdot)\rangle$.

Algorithm 1. Conjugate gradient regularized classification direction.

Initialize $\psi_0^{\mathrm{CG}} = 0$, $\nu_0 = \zeta_0 = \mu$
Repeat for $j = 0, \ldots, m-1$
$h_j = \langle \nu_j, \zeta_j \rangle / \langle \nu_j, \mathscr{R}\nu_j \rangle$
$\psi_{j+1}^{\mathrm{CG}} = \psi_j^{\mathrm{CG}} + h_j \nu_j$
$\zeta_{j+1} = \mu - \mathscr{R}\psi_{j+1}^{\mathrm{CG}}$ $(= \zeta_j - h_j \mathscr{R}\nu_j)$
$g_j = -\langle \zeta_{j+1}, \mathscr{R}\nu_j \rangle / \langle \nu_j, \mathscr{R}\nu_j \rangle$
$\nu_{j+1} = \zeta_{j+1} + g_j \nu_j$
Output ψ_m^{CG}

The conjugate gradient approach is an example of dimension reduction regularization. The method solves the minimization problem (3) with ψ restricted to the Krylov subspace $K_m(\mathscr{R}, \mu)$ spanned by $\mu, \mathscr{R}\mu, \ldots, \mathscr{R}^{m-1}\mu$ and also by the first m conjugate directions ν_j or the first m residuals ζ_j; that is, it seeks to minimize $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle$ subject to $\psi \in K_m(\mathscr{R}, \mu)$. The projection direction that solves this minimization is ψ_m^{CG}.

Another popular choice is to minimize $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle$ subject to $\psi \in E_m(\mathscr{R})$, where $E_m(\mathscr{R})$ is the subspace spanned by the first m eigenfunctions, $\varphi_1, \ldots, \varphi_m$, of $\mathscr{R}$ in the spectral decomposition

$$\mathscr{R} = \sum_{j=1}^{\infty} \lambda_j \varphi_j \otimes \varphi_j,$$

with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots > 0$ being the eigenvalues. The solution $\psi_m^{\mathrm{PC}} = \sum_{j=1}^{m} \lambda_j^{-1} \langle \mu, \varphi_j \rangle \varphi_j$ gives the principal component classifier of Delaigle & Hall (2012a).

In general one can minimize $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle$ subject to $\psi \in S_m$, where S_m is the m-dimensional subspace generated by some functions $s_1, \ldots, s_m$ such that the s_j $(j = 1, 2, \ldots)$ generate the range of $\mathscr{R}$. Let $\mathscr{P}_m$ be the projection operator that projects onto S_m, and let $\mathscr{R}_m = \mathscr{P}_m \mathscr{R} \mathscr{P}_m$ and $\mathscr{R}_m^- = \mathscr{P}_m \mathscr{R}^{-1} \mathscr{P}_m$. Then the solution of the regularized minimization problem is $\psi_m = \mathscr{R}_m^- \mu$. More explicitly, considering solutions of the form $\psi_m = \sum_{j=1}^{m} c_j s_j$ leads to the m-variate minimization of $c^{\mathrm{T}} Q c / 2 - u^{\mathrm{T}} c$ where the matrix Q is such that $Q_{jk} = \langle s_j, \mathscr{R} s_k \rangle$ and the vector u has components $u_j = \langle \mu, s_j \rangle$, i.e., to the solution with coefficients $c = Q^{-1} u$. In the case of the Krylov subspace, the iterative conjugate gradient method given in Algorithm 1 is, however, preferred because the matrix Q is ill-conditioned.

We can also take another approach to regularization, based on ridge regression. Optimizing the misclassification probability in a ball with radius $\theta^{1/2}$ leads to the task of minimizing $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle$ subject to $\|\psi\|^2 \leqslant \theta$ or, equivalently, minimizing $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle + \alpha \|\psi\|^2 / 2$, where $\alpha \geqslant 0$ is a regularization parameter. The solution is $\psi_\alpha^{\mathrm{R}} = \mathscr{R}_\alpha^{-1} \mu$, where $\mathscr{R}_\alpha = \mathscr{R} + \alpha \mathscr{I}$ and $\mathscr{I}$ denotes the identity operator. Despite its practical performance and amenability to theoretical analysis, the functional ridge classifier does not seem to have been considered before.

There is an important difference between the conjugate gradient method and the other approaches. While the principal component and ridge methods regularize the problem without the main goal in mind, the conjugate gradient approach greedily follows the goal of optimal classification. Indeed, the conjugate gradient method as an iterative optimization procedure constructs the regularization path focusing on the minimization of the misclassification probability, whereas the other approaches regularize by modifying the operator to be inverted regardless of the goal.

From a computational point of view the conjugate gradient method is simplest because it does not require inversion or eigendecomposition.

2.3. *Properties of regularization paths*

While ψ_m, the solution regularized by a subspace constraint, in general need not converge as $m \to \infty$ since a solution to the unconstrained minimization problem may not exist, the misclassification probability associated with the linear classifier given by ψ_m converges along the regularization path. The following and all other results are proved in the Appendix.

PROPOSITION 1. *The misclassification probability of the regularized linear classifier based on* $\psi_m = \mathscr{R}_m^- \mu$ *converges to* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ *as* $m \to \infty$.

This result holds regardless of whether the unconstrained minimization problem (3) has a solution, i.e., regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$. The limiting misclassification probability is positive if $\|\mathscr{R}^{-1/2}\mu\| < \infty$ or zero if $\|\mathscr{R}^{-1/2}\mu\| = \infty$. As discussed earlier, the optimal error is achieved exactly by the one-dimensional projection onto $\psi = \mathscr{R}^{-1}\mu$, when $\|\mathscr{R}^{-1}\mu\| < \infty$. Even when $\|\mathscr{R}^{-1}\mu\| = \infty$, both of the dimension reduction techniques, namely the conjugate gradient and principal component methods, and also ridge regularization as we will soon see, achieve the optimal limiting error rate along a possibly nonconvergent path of one-dimensional projection directions.

It is natural to investigate and compare how quickly the misclassification rate approaches the limit for the two main types of subspace regularization. It turns out that the conjugate gradient classifier, being a greedy, goal-oriented procedure, performs as well as or better than the principal component classifier with the same dimension.

PROPOSITION 2. *Regardless of whether the optimal misclassification probability can be achieved exactly or along a regularization path, i.e., whether* $\|\mathscr{R}^{-1}\mu\| < \infty$ *or* $\|\mathscr{R}^{-1}\mu\| = \infty$, *and regardless of whether the optimal misclassification probability is zero or positive, i.e., whether* $\|\mathscr{R}^{-1/2}\mu\| = \infty$ *or* $\|\mathscr{R}^{-1/2}\mu\| < \infty$, *the misclassification probability of the principal component classifier using m components is higher than or equal to the misclassification probability of the m-step conjugate gradient classifier.*

Phatak & de Hoog (2002, § 6.2) showed in the multivariate setting that 'PLS fits closer than PCR'. In infinite dimensions, in the context of kernel partial least squares, Blanchard & Krämer (2010, Theorem 1) showed that the partial least squares solution is closer to the true solution of the inverse problem than is the principal component solution with the same number of components. Unlike these results, our Proposition 2 does not assume the existence of a solution and instead focuses on the values of the misclassification probability.

Although Proposition 2 suggests that the conjugate gradient method will typically use fewer components than the principal component method to achieve the best result, the resulting misclassification probability with the best number of components need not be better. We address this in the simulation study. A similar phenomenon was previously studied in the literature on partial least squares in finite dimensions and in the functional setting by Febrero-Bande et al. (2017).

As in the case of subspace regularization, below we obtain the convergence of the error probability of the ridge classifier, whether or not the unconstrained minimization problem (3) has a solution, i.e., regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$. The limiting misclassification probability is positive if $\|\mathscr{R}^{-1/2}\mu\| < \infty$ or zero if $\|\mathscr{R}^{-1/2}\mu\| = \infty$.

PROPOSITION 3. *The misclassification probability of the regularized linear classifier based on* $\psi_\alpha^{\mathrm{R}} = \mathscr{R}_\alpha^{-1}\mu$ *converges to* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ *as* $\alpha \to 0+$.

3. EMPIRICAL CLASSIFIERS FOR FRAGMENTARY FUNCTIONS

3.1. *Construction of classifiers with incomplete training samples*

So far we have assumed that the parameters of each group are known. We now present the empirical version with a finite training dataset, and show that under regularity conditions such classifiers can achieve asymptotically the same optimal error rate as if there were infinite training data. We aim to do this not only in the case of fully observed functions but also in the case of incomplete curves. Incompleteness can occur in the training data, with each curve possibly observed on a different domain, as well as in the new curve that we wish to classify. One strategy would be to consider all curves on the intersection of their observation domains, if it is nonempty. However, such a restriction can be too severe and is unnecessary. We will construct classifiers that use the observed new curve on a set $\mathcal{I}$, which may be its entire observation set or a subset thereof, without requiring that all training curves be completely observed on $\mathcal{I}$.

For group j let there be a training sample consisting of n_j curves, $X_{j1}, \ldots, X_{jn_j}$. The training data are assumed to be mutually independent. Curves may be observed incompletely, with values known only on a subset O_{ji} of the domain and with no information about the values on the complement. The observation domains are assumed to be independent of the curves and consist of a finite union of intervals. We let $O_{ji}(t)$ denote the indicator of the curve X_{ji} being observed at time t. Similarly, let $U_{ji}(s,t)$ indicate observation at times s and t, i.e., $U_{ji}(s,t) = O_{ji}(s)O_{ji}(t)$.

The mean μ_j of group j can be estimated by the cross-sectional average

$$\hat{\mu}_j(t) = \frac{1_{\{N_j(t)>0\}}}{N_j(t)} \sum_{i=1}^{n_j} O_{ji}(t)X_{ji}(t) \quad (j = 0, 1),$$

where $N_j(t) = \sum_{i=1}^{n_j} O_{ji}(t)$ is the total number of observed curves in group j at time t. The covariance kernel $\rho(s,t)$ can be estimated by the empirical covariance using pairwise complete observations of groupwise centred curves. Formally, the estimator is

$$\hat{\rho}(s,t) = \frac{M_1(s,t)\hat{\rho}_1(s,t) + M_2(s,t)\hat{\rho}_2(s,t)}{M_1(s,t) + M_2(s,t)},$$

where $M_j(s,t) = \sum_{i=1}^{n_j} U_{ji}(s,t)$ and

$$\hat{\rho}_j(s,t) = \frac{1_{\{M_j(s,t)>0\}}}{M_j(s,t)} \sum_{i=1}^{n_j} U_{ji}(s,t)\{X_{ji}(s) - \hat{\mu}_{jst}(s)\}\{X_{ji}(t) - \hat{\mu}_{jst}(t)\}$$

with $\hat{\mu}_{jst}(s) = 1_{\{M_j(s,t)>0\}} M_j(s,t)^{-1} \sum_{i=1}^{n_j} U_{ji}(s,t)X_{ji}(s)$. If $N_j(t) = 0$ or $M_j(s,t) = 0$, the estimators are defined as $\hat{\mu}_j(t) = 0$ or $\hat{\rho}_j(s,t) = 0$, respectively. This happens with asymptotically vanishing probability under Assumption 1 below.

Suppose that the new independent curve to be classified, X_{new}, is observed on the domain O_{new}. Let us fix the target domain $\mathcal{I} \subseteq O_{\mathrm{new}}$ on which we aim to apply the classifier to X_{new}. The empirical classifier $\hat{C}_{\hat{\psi}}$ trained on partially observed curves is defined like the theoretical one, with unknown quantities replaced by their estimators. It assigns X_{new} restricted to $\mathcal{I}$ to the class

$\hat{C}_{\hat{\psi}}(X_{\text{new}}) = 1_{\{\hat{T}_{\hat{\psi}}(X_{\text{new}})>0\}}$, where $\hat{T}_{\hat{\psi}}(X_{\text{new}}) = \langle X_{\text{new}} - \tilde{\mu}, \hat{\psi}\rangle\langle \hat{\mu}, \hat{\psi}\rangle$. Here $\tilde{\mu} = (\hat{\mu}_0 + \hat{\mu}_1)/2$ and $\hat{\mu} = \hat{\mu}_1 - \hat{\mu}_0$, with $\hat{\mu}_j$ being the estimators defined above restricted to $\mathcal{I}$. The projection direction $\hat{\psi}$ is one of $\hat{\psi}_m^{\text{CG}}$, $\hat{\psi}_m^{\text{PC}}$ or $\hat{\psi}_\alpha^{\text{R}}$, constructed respectively by conjugate gradient, principal component or ridge regularization applied to $\hat{\mu}$ and $\hat{\mathscr{R}}$, where $\hat{\mathscr{R}}$ is the integral operator with kernel $\hat{\rho}(s,t)$ introduced above, restricted to $\mathcal{I} \times \mathcal{I}$.

All methods discussed in the previous section can be formulated in terms of the population parameters, i.e., the mean difference and covariance operator, and not in terms of individual observations in the training set. The population parameters can be consistently estimated by averaging individual observations, whereas temporal averaging of individual curves, for example in inner products, is impossible due the incompleteness of the observed functions. In particular, the conjugate gradient method can be applied to fragmentary training data, whereas the usual algorithms for multivariate or functional partial least squares, such as those in De Jong (1993), Hastie et al. (2009, Algorithm 3.3) and Delaigle & Hall (2012b, § 4.2 and Appendix A.2), involve the computation of certain scores, i.e., inner products, for individual curves.

3.2. *Asymptotic behaviour along the empirical regularization path*

We aim to study the behaviour of classifiers on incomplete training samples of increasing size with decreasing amounts of regularization. Previous asymptotic results in related settings include those of Delaigle & Hall (2013), who established the consistency of empirical principal component classifiers based on partially observed training data. In the setting of complete curves, Berrendero et al. (2018) used dimension reduction regularization by evaluation of curves at a finite set of arguments; they proved consistency of the empirical version but did not study the asymptotics for decreasing amounts of regularization, i.e., they did not consider letting the dimension grow. Baíllo et al. (2011a) studied optimal classifiers for Gaussian measures based on Radon–Nikodym derivatives and investigated the performance of their empirical version in the special class of processes with triangular covariance functions. In contrast, all of our methods, including the ridge approach not considered previously, have been developed for fragmentary training samples and shown to achieve the Bayes error rate for general Gaussian processes along the empirical regularization path, as we now explain.

The following assumptions will be needed for the derivation of asymptotic properties of empirically trained regularized linear classifiers.

Assumption 1. The distributions in groups $j = 0, 1$ satisfy $E_{P_j}(\|X\|^4) < \infty$.

Assumption 2. For a domain $\mathcal{I}$, there exists $\delta > 0$ such that the observation patterns in training samples $j = 0, 1$ satisfy, as $n_j \to \infty$,

$$\sup_{(s,t)\in\mathcal{I}\times\mathcal{I}} \text{pr}\{n_j^{-1} M_j(s,t) > \delta\} = O(n_j^{-2}).$$

Assumption 1 guarantees the consistency of the empirical mean and covariance operator for samples of completely observed curves; see, for example, Bosq (2000) or Horváth & Kokoszka (2012). Kraus (2015, Proposition 1) showed, under the additional Assumption 2 with $\mathcal{I}$ equal to the entire domain of the curves, that the root-n consistency of the sample mean and covariance restricted to $\mathcal{I}$ continues to hold in the fragmentary setting. In particular, it follows that $\|\hat{\mu}_j - \mu_j\| = O_{\text{p}}(n_j^{-1/2})$ and hence $\|\hat{\mu} - \mu\| = O_{\text{p}}(n^{-1/2})$ for $n = \min(n_0, n_1) \to \infty$, and also that $\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = O_{\text{p}}\{(n_0 + n_1)^{-1/2}\}$, where $\|\cdot\|_\infty$ is the operator norm. When $\mathcal{I}$ is a subset of

the domain, analogous results hold for the restrictions of the functions and integral kernels to $\mathcal{I}$. Assumption 2 means that at all pairs of time-points there is an asymptotically nonnegligible fraction of observed values. Assumption 2 is less restrictive than the requirement that there be complete curves in the sample. It can be satisfied, for example, in situations where the observed curves consist of several shorter fragments. If the assumption is not satisfied because the data contain only one short fragment per curve, other estimation methods can be used; see, for example, Delaigle & Hall (2016) and Descary & Panaretos (2019).

We now study the asymptotic behaviour of the empirical classifier when the number m_n of steps of the conjugate gradient algorithm grows as the training sample size grows. Under certain conditions on the regularization path, we establish the convergence of the misclassification probability of the conjugate gradient classifier trained on collections of functional fragments to the same optimal limit as for the theoretical conjugate gradient classifier with an infinite training sample, regardless of whether the limiting error rate is zero or positive and regardless of whether the limit can be theoretically achieved exactly or along the path.

THEOREM 1. *Suppose that Assumption 1 holds. Assume that* $n = \min(n_0, n_1) \to \infty$ *and* $m_n \to \infty$ *in such a way that* $m_n \leqslant Cn^{1/2}$ *for some* $C > 0$ *and*

$$n^{-1/2}\omega_{m_n}^{-1}\|\gamma^{(m_n)}\| + n^{-1}\omega_{m_n}^{-3} \to 0, \tag{4}$$

where ω_{m_n} *is the smallest eigenvalue of the* $m_n \times m_n$ *matrix* H *with entries* $h_{jk} = \langle \kappa_j, \mathscr{R}\kappa_k \rangle$ *for* $\kappa_j = \mathscr{R}^{j-1}\mu$ *and the* m_n*-vector* $\gamma^{(m_n)}$ *is defined as* $\gamma^{(m_n)} = H^{-1}d$ *with* d *being the* m_n*-vector having components* $d_j = \langle \mu, \kappa_j \rangle$*. Then the misclassification probability of the empirical regularized linear classifier based on* $\hat{\psi}_{m_n}^{\mathrm{CG}}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

Condition (4) guarantees that the number of components does not grow too fast in relation to the growing number of training observations and to the increased ill-conditioning of the theoretical problem. Condition (4) is analogous to (5.10) in Delaigle & Hall (2012b) for partial least squares. The vector $\gamma^{(m_n)}$ contains the coefficients of the theoretical regularized solution $\psi_{m_n}^{\mathrm{CG}}$ with respect to the non-orthogonal basis $\kappa_1, \ldots, \kappa_{m_n}$ of the Krylov subspace $K_{m_n}(\mathscr{R}, \mu)$, i.e., $\psi_{m_n} = \sum_{j=1}^{m_n} \gamma_j^{(m_n)} \kappa_j$. The eigenvalues of H are called the Ritz values in numerical analysis. For details on connections with partial least squares see Lingjærde & Christophersen (2000).

In the proof given in the Appendix we use the results of Delaigle & Hall (2012b) on the consistency of partial least squares regression for functional data. These results were obtained for situations that differ from our setting in several ways. In particular, we work with functional fragments instead of complete curves, the conjugate gradient path differs from partial least squares regression, e.g., in the group centring in the estimation of the covariance, and we do not require that the population inverse problem, $\mathscr{R}\psi = \mu$ in our context, have a solution. However, inspection of the underlying technical arguments in Delaigle & Hall (2012b) shows that appropriate analogous results can be obtained and used in our setting, as we explain in the proof.

Next, we show that the empirically trained principal component classifier with an increasing number of components asymptotically achieves the optimal misclassification probability.

THEOREM 2. *Suppose that Assumption 1 holds. Assume that* $n = \min(n_0, n_1) \to \infty$ *and* $m_n \to \infty$ *in such a way that* $\lambda_{m_n}^4 n \to \infty$ *and* $\lambda_{m_n}^2 n (\sum_{j=1}^{m_n} a_j)^{-2} \to \infty$*, where* $a_1 = 2^{3/2}(\lambda_1 - \lambda_2)^{-1}$ *and* $a_j = 2^{3/2}\max\{(\lambda_{j-1} - \lambda_j)^{-1}, (\lambda_j - \lambda_{j+1})^{-1}\}$ *for* $j = 2, 3, \ldots$*. Then the misclassification*

probability of the empirical regularized linear classifier based on $\hat{\psi}^{\mathrm{PC}}_{m_n}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

The conditions on the principal component regularization path are the same as in the case of functional principal component regression (Cardot et al., 1999). Unlike in the functional linear model, it is not assumed that the inverse problem has a solution, since the goal is not to estimate the possibly nonexistent bounded linear regression functional.

Finally, the empirical ridge classifier with finite training data asymptotically attains the same optimal error rate as its theoretical counterpart. Unlike for the conjugate gradient and principal component classifiers, the conditions on the ridge path classifier do not involve parameters of the distributions because no subspace is constructed.

THEOREM 3. *Suppose that Assumption 1 holds. Assume that* $n = \min(n_0, n_1) \to \infty$ *and* $\alpha_n \to 0+$ *in such a way that* $\alpha_n^4 n \to \infty$. *Then the misclassification probability of the empirical regularized linear classifier based on* $\hat{\psi}^{\mathrm{R}}_{\alpha_n}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

3.3. *Selection of the regularization parameter*

The regularization parameter can be selected by minimizing an estimate of the misclassification probability. We use leave-one-out crossvalidation. The Supplementary Material provides details of crossvalidation in the presence of incomplete curves. The best value of the regularization parameter is searched for over a grid of values, such as the values corresponding to integer degrees of freedom up to some maximum value. The number of degrees of freedom for the subspace methods is the dimension of the subspace, and for the ridge method it is defined as the trace of $(\hat{\mathscr{R}} + \alpha\mathscr{I})^{-1}\hat{\mathscr{R}}$, i.e., $\sum_{j=1}^{n_0+n_1} \hat{\lambda}_j/(\hat{\lambda}_j + \alpha)$ where $\hat{\lambda}_j$ are the eigenvalues of $\hat{\mathscr{R}}$. The maximum number of degrees of freedom we use is one fifth of the number of curves.

4. DOMAIN SELECTION

To classify the new curve X_{new} observed on O_{new}, we apply the classifier on the target domain $\mathcal{I} \subseteq O_{\mathrm{new}}$, the choice of which we now consider. One possibility would be to restrict attention to the intersection of the observation domains of all curves, say $\mathcal{I}_0$, if it is nonempty. An obvious drawback of this approach is that one can lose discriminatory power because any differences between the classes may be more pronounced outside $\mathcal{I}_0$. An advantage of our approach is its capability of working with incomplete curves, since the empirical construction of the projection direction requires only the estimation of μ and $\mathscr{R}$ on the target domain. Hence one can look at a domain larger than $\mathcal{I}_0$. A natural choice is the largest subset of O_{new} that contains enough data for estimation of the classifier, i.e., satisfies Assumption 2, and contains enough functions for validation in the crossvalidation procedure, i.e., has a sufficiently large set V. In this way one hopes to capture the widest range of shapes of the group difference. On the other hand, it could be that not even this maximal domain, $\mathcal{I}^{\max}$, will lead to the best classification accuracy, because one includes more uncertainty in the estimation due to the missing values and because the mean difference may not be important in the added part of the domain. Therefore, it seems reasonable to also consider intermediate choices between $\mathcal{I}_0$ and $\mathcal{I}^{\max}$.

Here we present a domain selection strategy for the most common case of interval observation sets. The idea, worked out in detail in Stefanucci et al. (2018), is to construct the classifier on a series of intervals that range from the common domain $\mathcal{I}_0$ to the maximal domain $\mathcal{I}^{\max}$, extending the working interval by a fixed percentage at each step. More formally, we consider a sequence

of nested intervals $\mathcal{I}_0 \subset \mathcal{I}_1 \subset \cdots \subset \mathcal{I}_k \subset \cdots \subset \mathcal{I}_K = \mathcal{I}^{\max}$, starting from $\mathcal{I}_0$ and ending in $\mathcal{I}_K = \mathcal{I}^{\max}$, and build the classifier on each interval. The regularization parameter for the kth domain is selected by crossvalidation as described in the Supplementary Material. Among these $K + 1$ candidates we select the one that minimizes the crossvalidation estimate of error.

The search strategy can be extended by considering larger systems of candidate domains; for example, one could vary the two endpoints independently. The idea can be generalized to other situations, such as non-interval observation sets, multivariate functional data or functions indexed by multivariate arguments. In each situation one needs to define a meaningful system of domains and optimize the crossvalidation score over the system.

5. Simulations

5.1. *Behaviour of regularized classifiers on complete data*

In this section we illustrate the behaviour of the three estimators of ψ in different settings. We consider Gaussian processes on $[0, 1]$ with covariance kernel $\rho(s, t) = \exp(-|s - t|^2/0.01)$ and mean function depending on the group label. Group 0 has mean $\mu_0(t) = 0$ in each setting. Group 1 has mean $\mu_1(t) = \mu(t)$, for which we consider eight different forms: (i) ct, (ii) $c(t-0.5)^2$, (iii) $c(t-0.5)^3$, (iv) $c\sin(20t)$, (v) $c\varphi_1(t)$, (vi) $c\varphi_{10}(t)$, (vii) $cb(t; 5, 5)$, and (viii) $cb(t; 2, 6)$, where φ_j is the jth eigenfunction of the kernel ρ and $b(t; \alpha, \beta) = t^{\alpha-1}(1 - t)^{\beta-1}$ is the beta density. In each case the parameter c is selected to yield a reasonable misclassification rate.

In each of 5000 repetitions we generated 50 curves from each group and evaluated them on a grid of 100 equispaced points in $[0, 1]$. We also generated a new observation that could arise from group 0 or group 1 with equal probability. Then we constructed the regularized classification direction by the principal component, conjugate gradient and ridge methods with m degrees of freedom and predicted the label of the new observation. We considered $m = 1, \ldots, 20$, corresponding to a reasonable minimum of five observations per degree of freedom.

Figure 1 shows the misclassification proportion over the 5000 repetitions as a function of m for the eight different choices of $\mu(t)$. As expected, the conjugate gradient method performs well in all settings and is not much affected by the shape of $\mu(t)$. By contrast, the performance of the principal component classifier depends strongly on $\mu(t)$. To see this, consider the two extreme situations in settings (v) and (vi). The classification error of the principal component approach is close to that of the conjugate gradient method in case (v), where $\mu(t)$ is the first eigenfunction, but is much higher at lower dimensions in case (vi), where $\mu(t)$ is the tenth eigenfunction. In the latter case, the principal component method reaches the same level of error as the conjugate gradient method only when $m = 10$ or more. These findings agree with Proposition 2 and with the conclusions of Delaigle & Hall (2012a) and Febrero-Bande et al. (2017), who pointed out that principal components need more degrees of freedom than partial least squares to achieve good performance. In this regard ridge regularization seems to lie between the two subspace methods, but is more similar to the conjugate gradient method in most cases. In particular, it does not completely fail at low degrees of freedom in case (vi), because it does not construct a subspace that could miss the important information; however, it also suffers in this situation, where $\mu(t)$ is on the tail of the spectrum, because ridge penalization shrinks higher-index spectral components more than lower-index components. Nevertheless, with sufficiently many degrees of freedom, the three methods behave similarly.

Additional simulation results, reported in the Supplementary Material, show that similar conclusions can be drawn when functions have nonsmooth trajectories and that the capability to discriminate between two groups with different means is robust with respect to the assumption of equal covariances. Results for increased training sample size are also provided in the Supplementary Material.

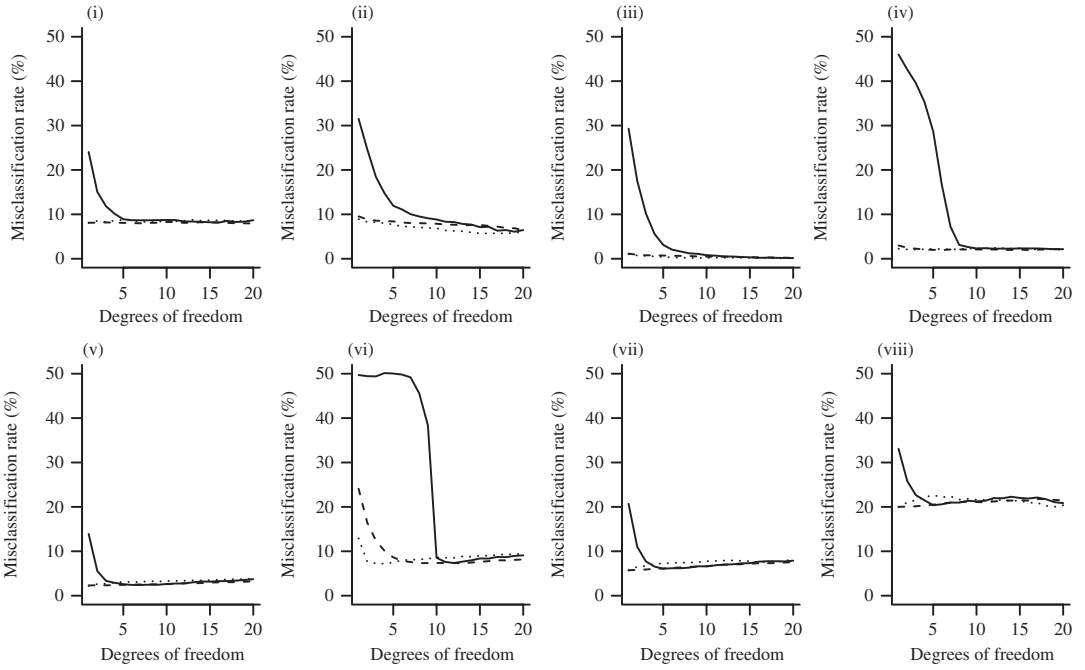

Fig. 1. Misclassification rate (%) versus degrees of freedom for different forms of $\mu(t)$: (i) linear, (ii) quadratic, (iii) cubic, (iv) sinusoidal, (v) first eigenfunction, (vi) tenth eigenfunction, (vii) symmetric beta, and (viii) asymmetric beta. The different curves represent the principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers.

Table 1. *Misclassification rates* (%)*, with standard errors in parentheses, achieved by classifiers with degrees of freedom selected by crossvalidation in the different settings; for each classifier the numbers in the second row are the minimum misclassification rates*

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
PC	13.0 (0.34)	8.3 (0.28)	1.3 (0.11)	2.5 (0.16)	7.2 (0.26)	7.6 (0.27)	10.7 (0.31)	26.2 (0.44)
	8.1	6.1	0.1	2.2	2.4	7.4	6.1	20.4
CG	8.6 (0.28)	6.5 (0.25)	0.7 (0.09)	2.1 (0.14)	2.6 (0.16)	7.8 (0.27)	6.1 (0.24)	20.9 (0.41)
	8.1	5.7	0.1	2.1	2.2	7.2	5.7	19.9
R	8.4 (0.28)	7.7 (0.27)	0.7 (0.09)	2.2 (0.15)	2.4 (0.15)	7.9 (0.27)	6.1 (0.24)	20.8 (0.41)
	7.9	6.5	0.2	2.0	2.3	7.3	5.7	20.0

PC, principal component classifier; CG, conjugate gradient classifier; R, ridge classifier.

5.2. *Performance of crossvalidation for selection of degrees of freedom*

We used simulation to investigate the performance of leave-one-out crossvalidation in choosing the correct level of regularization. The settings were the same as in § 5.1, but classification was done using the number of degrees of freedom selected by leave-one-out crossvalidation. We summarize the classification errors in Table 1. Crossvalidation performs well as a selector of the best level of regularization since the misclassification rate in Table 1 is in each case close to the corresponding minimum error in Fig. 1. The principal component method appears to perform worst, while the conjugate gradient and ridge methods have comparable performance. The latter two methods nearly achieve the respective minimum errors. Table 2 reports the mean and median selected degrees of freedom. The principal component method often uses considerably more degrees of freedom than the other methods. This is particularly interesting in case (v), where the

Table 2. *Mean and median (in parentheses) degrees of freedom selected by crossvalidation*

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
PC	8.2 (7)	14.3 (15)	9.9 (9)	10.9 (10)	4.6 (4)	11.9 (11)	5.3 (4)	8.6 (6)
CG	5.4 (3)	10.7 (11)	3.4 (2)	4.5 (2)	2.4 (1)	4.9 (3)	2.7 (1)	8.6 (7)
R	6.4 (3)	11.6 (13)	6.0 (3)	6.1 (4)	2.7 (1)	9.3 (8)	3.4 (1)	6.7 (3)

PC, principal component classifier; CG, conjugate gradient classifier; R, ridge classifier.

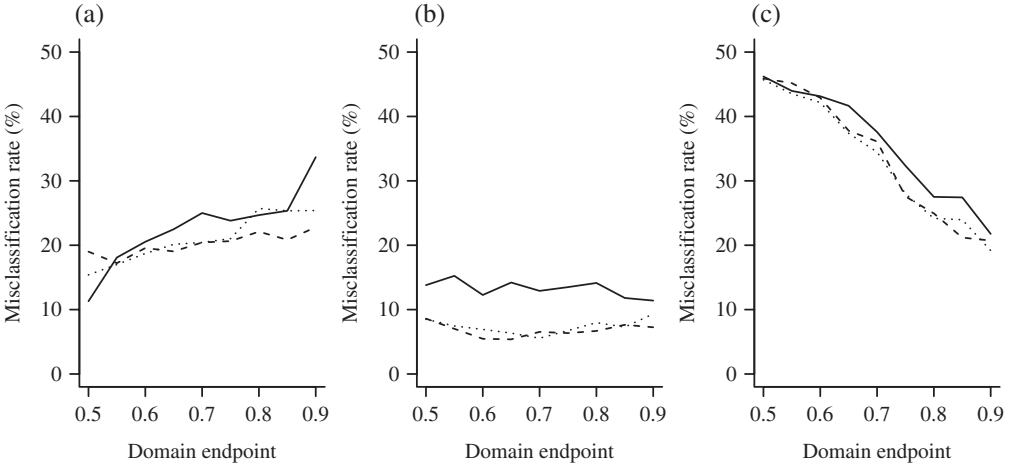

Fig. 2. Misclassification rate (%) plotted as a function of the domain extension, for $\mu(t)$ being the (a) Be(2, 6), (b) Be(5, 5) or (c) Be(6, 2) density for the principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers with selected degrees of freedom. Classification is performed on the domains $[0, u]$ with $u \in [0.5, 0.9]$, and the error values are plotted against u.

mean difference equals the first eigenfunction and so one component should be the best choice in theory. These results again illustrate the general phenomenon that the principal component approach is inappropriate for inference about means due to the possible lack of informativeness of the principal components about the mean and the extra uncertainty associated with their estimation.

5.3. *Missing data and domain extension*

We now demonstrate the usefulness of the domain extension approach presented in § 4, using Gaussian processes on [0, 1] with the same covariance as in § 5.1 and considering three scenarios for the mean difference in the form of a multiple of a beta density, (a) $b(t; 2, 6)$, (b) $b(t; 5, 5)$ and (c) $b(t; 6, 2)$, which reflect situations where discrimination due to a peak is in the left, central and right parts of the domain, respectively. We sampled 50 curves from each group on a sequence of 100 equispaced points in [0, 1]. Then we generated endpoints of the observation interval for each curve from the uniform distribution on (0.5, 1); that is, each curve was observed between 0 and the endpoint and treated as missing beyond the endpoint. The new observation had an endpoint sampled between 0.5 and 1. So the first half of [0, 1], $\mathcal{I}_0 = [0, 0.5]$, was the common observation domain of all curves. We considered extensions of $\mathcal{I}_0$ to $\mathcal{I}_k = [0, 0.5 + 0.05k]$ $(k = 0, \ldots, 8)$. For each interval of this form that was contained in the observation domain of the curve to be classified, we estimated the classifiers, choosing the best degrees of freedom via crossvalidation, and classified the new curve. This procedure was repeated 1000 times. We plot the behaviour of the resulting classification error as a function of the endpoint of the extended domain in Fig. 2.

When the peak of the mean difference is in the left part of [0, 1], extending the domain does not lead to better classification. In this case the interval where the means mainly differ corresponds to the part of the domain where all the data are available, and inflating the domain only increases

Table 3. *Misclassification rates* (%)*, with standard errors in parentheses, achieved by classifiers with domain and degrees of freedom selected by crossvalidation in the different settings; the minimum and maximum misclassification rates are given in square brackets*

	(a)	(b)	(c)
PC	18.1 (0.38) [11.3, 33.7]	11.9 (0.32) [11.4, 15.2]	31.1 (0.46) [21.8, 46.0]
CG	19.6 (0.39) [15.4, 25.7]	7.4 (0.26) [5.6, 9.3]	30.4 (0.46) [19.2, 45.7]
R	22.4 (0.42) [17.2, 22.8]	6.9 (0.25) [5.4, 8.6]	28.4 (0.45) [20.7, 45.9]

PC, principal component classifier; CG, conjugate gradient classifier; R, ridge classifier.

the uncertainty due to missing data. In the second case, the peak of the mean difference is exactly at 0.5, and extending the domain leads to little improvement. The third scenario is the opposite of the first, as the discrimination is mainly in the right part of $[0, 1]$. In this case, extending the domain reduces the error considerably because good classification is only possible by employing the right part of the domain. The classification error is about 45% when using only $\mathcal{I}_0$, but drops to about 20% when using also the part of the interval where the data are partially observed.

5.4. *Performance with selected domain*

Domain extension may or may not improve the performance of classifiers, depending on the interplay between the form of the mean difference, the covariance structure and the missingness pattern. In practice, the user is not an oracle with access to misclassification errors for candidate subsets whose estimates are plotted in Fig. 2, and hence would select the best domain by crossvalidation. In Table 3 we report simulation results for classifiers with both domain and degrees of freedom selected by crossvalidation, for the same configurations as in § 5.3. Selection of the domain leads to a considerable improvement of the error rate compared with the worst-performing domain. On the other hand, this improvement has some limitations and a gap remains between the achieved value and the best value; this can be explained by the fact that crossvalidation provides only an estimate of the error, not the true value.

6. ANEURISK DATA EXAMPLE

We apply the proposed method to the AneuRisk dataset from an interdisciplinary project aimed at investigating the effects of blood vessel morphology, blood fluid dynamics and biomechanical properties of the vascular wall on the pathogenesis of cerebral aneurysms. An introduction to the data can be found in Sangalli et al. (2014b). This dataset has previously been analysed in several works that focused on different methodological aspects, such as function and derivative estimation (Sangalli et al., 2009b), exploratory analysis and classification (Sangalli et al., 2009a), and alignment and clustering (Sangalli et al., 2014a), among others.

The data consist of measurements of the radius and curvature of the internal carotid artery in a sample of 65 patients, 33 of which have an aneurysm at the bifurcation of the vessel or after it, while the other 32 either have an aneurysm before the bifurcation, which is much less dangerous, or are healthy. The goal is to classify the patients based on the morphology of their internal carotid artery. In this example we work with only one of the observed variables, the radius. The data have previously been pre-processed, registered and smoothed, and are observed on a grid of 2000 points in the interval $[-100.3, 5.1]$, where the argument represents the distance between the observation point and the terminal bifurcation of the internal carotid artery, with positive values indicating points inside the skull. As we can see in Fig. 3, the data are partially observed because

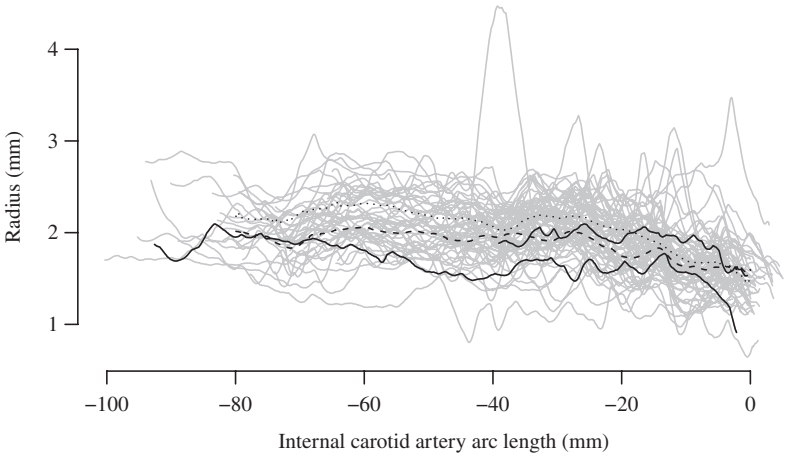

Fig. 3. Radius along the carotid artery from the AneuRisk dataset, along with the mean of the group of subjects with an aneurysm after the bifurcation (dotted) and the mean of the group of subjects with an aneurysm before the bifurcation or without an aneurysm (dashed). Curves for two example subjects are highlighted as solid lines. Note the different start and end points for different subjects in the study.

the start and end points are different from subject to subject. All subjects are observed on the subset $\mathcal{I}_0 = [-32.9, -7.4]$, which corresponds to 24.3% of the whole domain.

We first apply the regularized linear classifiers to curves restricted to the common domain $\mathcal{I}_0$. The classification error estimated by crossvalidation is 29.2% for the principal component method, 29.2% for the conjugate gradient method, and 32.3% for ridge regularized classification.

We compare the above procedure with a different approach consisting of a multivariate classification method applied to principal component scores. The covariance kernel is estimated from observations centred to their respective group means, its eigenfunctions are computed, and quadratic discriminant analysis is applied to the inner products of the uncentred curves with the eigenfunctions. This procedure is similar to that in Sangalli et al. (2009a). The best classifier of this type turns out to exhibit a misclassification error of 32.3%, obtained with two eigenfunctions.

These values show that in this dataset, when attention is restricted to the common domain $\mathcal{I}_0$, our proposed method is comparable to the more standard multivariate technique.

Next, we consider classification on extended domains including observed values outside the common domain $\mathcal{I}_0$. We build the sequence of domains $\mathcal{I}_0, \ldots, \mathcal{I}_K$ by enlarging the domain at each step by 1.25% of the complement of $\mathcal{I}_0$. This step size is a compromise between the fineness of the grid and the computational cost. We consider extended domains up to $K = 40$, corresponding to $\mathcal{I}_{40} = [-66.6, -1.2]$, because not enough subjects have observed values outside this interval for reliable estimation and crossvalidation. All regularized linear classification methods benefit from the domain extension; in particular, the error rate for the principal component method drops from 29.2% to 23.2%, for the conjugate gradient method from 29.2% to 25.8%, and for ridge regularization from 32.3% to 25%. The best domain is $\mathcal{I}_{10} = [-41.3, -5.8]$ for the conjugate gradient method and $\mathcal{I}_{11} = [-42.2, -5.7]$ for the other two methods.

The alternative method based on multivariate classification of scores cannot be applied on extended domains since the individual scores of incomplete curves cannot be computed, although they can be predicted (Kraus, 2015). By contrast, the proposed methods are entirely formulated in terms of distributional parameters, which can be consistently estimated from incomplete data, unlike individual quantities.

ACKNOWLEDGEMENT

The AneuRisk data and useful comments were kindly provided by Laura Sangalli. The work of David Kraus was supported by the Czech Science Foundation. We are grateful to two referees, an associate editor and the editor for helpful suggestions and corrections.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the derivation of classifiers under unequal prior class probabilities, algorithmic details of crossvalidation, and additional simulation and real-data results.

APPENDIX

Proof of Proposition 1

The misclassification probability for ψ_m is $D(\psi_m)$ given in (1). Since $\psi_m \in S_m$, we compute

$$\frac{|\langle \mu, \psi_m \rangle|}{\langle \psi_m, \mathscr{R}\psi_m \rangle^{1/2}} = \frac{\langle \mu, \mathscr{R}_m^- \mu \rangle}{\langle \mu, \mathscr{R}_m^- \mathscr{R} \mathscr{R}_m^- \mu \rangle^{1/2}} = \|(\mathscr{R}_m^-)^{1/2}\mu\|.$$

By Lebesgue's monotone convergence theorem, the right-hand side converges to $\|\mathscr{R}^{-1/2}\mu\|$, finite or infinite, and therefore the limiting misclassification probability that is attained along the regularization path ψ_m, as $m \to \infty$, is $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

Proof of Proposition 2

The conjugate gradient method minimizes the quadratic objective function in the Krylov subspace $K_m(\mathscr{R}, \mu)$ whose elements are in the form $\eta = \sum_{k=0}^{m-1} c_k \mathscr{R}^k \mu = p(\mathscr{R})\mu$, where p is a polynomial of order lower than m. Then $\eta \in K_m(\mathscr{R}, \mu)$ can be written as $\eta = \sum_{j=1}^{\infty} p(\lambda_j) b_j \varphi_j$ with $b_j = \langle \mu, \varphi_j \rangle$. The objective function at η equals

$$\begin{aligned} \langle \eta, \mathscr{R}\eta \rangle/2 - \langle \mu, \eta \rangle &= \langle p(\mathscr{R})\mu, \mathscr{R}p(\mathscr{R})\mu \rangle/2 - \langle \mu, p(\mathscr{R})\mu \rangle \\ &= \sum_{j=1}^{\infty} b_j^2 \{p(\lambda_j)^2 \lambda_j/2 - p(\lambda_j)\} \\ &= \sum_{j=1}^{\infty} \frac{b_j^2}{2\lambda_j} q(\lambda_j)\{q(\lambda_j) - 2\}, \end{aligned} \tag{A1}$$

where $q(\lambda) = p(\lambda)\lambda$ is a polynomial of degree at most m such that $q(0) = 0$. The conjugate gradient method seeks the polynomial with these properties that minimizes the objective function. To prove the proposition we shall find a polynomial q with the required properties such that the objective function above is smaller than or equal to the objective function for the principal component classifier. The principal component classifier uses $\psi_m^{\mathrm{PC}} = \sum_{j=1}^{m} \lambda_j^{-1} b_j \varphi_j$, and the objective function at ψ_m^{PC} is

$$\langle \psi_m^{\mathrm{PC}}, \mathscr{R}\psi_m^{\mathrm{PC}} \rangle/2 - \langle \mu, \psi_m^{\mathrm{PC}} \rangle = -\sum_{j=1}^{m} \frac{b_j^2}{2\lambda_j}. \tag{A2}$$

Consider the polynomial of degree m,

$$q(\lambda) = 1 - (-1)^m \frac{\lambda - \lambda_1}{\lambda_1} \cdots \frac{\lambda - \lambda_m}{\lambda_m},$$

with $q(0) = 0$. We see that $q(\lambda_j) = 1$ for $j = 1, \ldots, m$, so the first m summands in the series (A1) and (A2) are equal. For $j > m$ we have that $0 \leqslant q(\lambda_j) \leqslant 2$ due to the properties of the eigenvalue sequence; so $q(\lambda_j)\{q(\lambda_j) - 2\} \leqslant 0$ and therefore the corresponding summands in the series (A1) are negative, whereas they are zero in the series (A2). Hence, for this polynomial,

$$\sum_{j=1}^{\infty} \frac{b_j^2}{2\lambda_j} q(\lambda_i)\{q(\lambda_i) - 2\} \leqslant -\sum_{j=1}^{m} \frac{b_j^2}{2\lambda_j},$$

and so the objective at the conjugate gradient solution must be smaller than or equal to the objective at the principal component solution. The inequality between the minima of the quadratic objective function implies the inequality between the misclassification probabilities stated in the proposition.

Proof of Proposition 3

Proceeding as in the proof of Proposition 1, we need to show that

$$\frac{\langle \mu, \mathscr{R}_\alpha^{-1}\mu\rangle}{\langle \mu, \mathscr{R}_\alpha^{-1}\mathscr{R}\mathscr{R}_\alpha^{-1}\mu\rangle^{1/2}} = \frac{\sum_{j=1}^{\infty} \frac{b_j^2}{\lambda_j+\alpha}}{\{\sum_{j=1}^{\infty} \frac{\lambda_j b_j^2}{(\lambda_j+\alpha)^2}\}^{1/2}} \xrightarrow[\alpha\to 0+]{} \left(\sum_{j=1}^{\infty} \frac{b_j^2}{\lambda_j}\right)^{1/2} = \|\mathscr{R}^{-1/2}\mu\|,$$

where $b_j = \langle \mu, \varphi_j\rangle$ is the coefficient of μ in the eigenbasis. If $\sum_{j=1}^{\infty} b_j^2/\lambda_j < \infty$, the convergence follows from Lebesgue's monotone convergence theorem. Otherwise, we use the inequality $\sum_{j=1}^{\infty} \lambda_j b_j^2/(\lambda_j+\alpha)^2 \leqslant \sum_{j=1}^{\infty} b_j^2/(\lambda_j+\alpha)$ to bound the left-hand side expression from below by $\{\sum_{j=1}^{\infty} b_j^2/(\lambda_j+\alpha)\}^{1/2}$, which diverges to infinity again by Lebesgue's theorem.

Proof of Theorem 1

The probability of misclassifying a new observation using the conjugate gradient classifier based on $\hat{\psi}_{m_n}^{\mathrm{CG}}$ is $D(\hat{\psi}_{m_n}^{\mathrm{CG}}) = 1 - \Phi\{|Z(\hat{\psi}_{m_n}^{\mathrm{CG}})|/2\}$. We need to show that the fraction in $Z(\hat{\psi}_{m_n}^{\mathrm{CG}})$ converges in probability to $\|\mathscr{R}^{-1/2}\mu\|/2$ along the regularization path satisfying the assumptions of the theorem. To deal with the numerator in $Z(\hat{\psi}_{m_n}^{\mathrm{CG}})$, one can show that

$$\langle \mu, \hat{\psi}_{m_n}^{\mathrm{CG}}\rangle - \langle \mu, \psi_{m_n}^{\mathrm{CG}}\rangle = O_{\mathrm{p}}\big(n^{-1/2}\omega_{m_n}^{-1}\|\gamma^{(m_n)}\| + n^{-1}\omega_{m_n}^{-3}\big). \tag{A3}$$

This result follows from an analogue of (5.9) in Theorem 5.3 of Delaigle & Hall (2012b) and intermediate results in the proof of that theorem which can be established in our context. The necessary modifications of the proofs of Theorems 5.1, 5.2 and 5.3 in Delaigle & Hall (2012b) are as follows. All results remain valid for incomplete instead of complete curves, because the proofs depend only on the root-n consistency of the covariance estimators, which holds also for functional fragments (Kraus, 2015, Proposition 1). Moreover, the derivations in Delaigle & Hall (2012b) can be repeated without assuming that the theoretical solution $\psi = \mathscr{R}^{-1}\mu$ exists as an element of $L^2(\mathcal{I})$. Indeed, the proofs in Delaigle & Hall (2012b) are based on stochastic expansions of $\hat{\mathscr{R}}^j\psi = \hat{\mathscr{R}}^j\mathscr{R}^{-1}\mu$, in our notation, about $\mathscr{R}^j\psi = \mathscr{R}^j\mathscr{R}^{-1}\mu = \mathscr{R}^{j-1}\mu$ and derived quantities, but the same steps can be followed for $\hat{\mathscr{R}}^{j-1}\hat{\mu}$ about $\mathscr{R}^{j-1}\mu$ in our setting. In other words, it can be shown that $\hat{\psi}_{m_n}^{\mathrm{CG}}$ and $\psi_{m_n}^{\mathrm{CG}}$ converge to each other without assuming that $\psi_{m_n}^{\mathrm{CG}}$ converges. Similarly, for the denominator in $Z(\hat{\psi}_{m_n}^{\mathrm{CG}})$ we have that

$$\langle \hat{\psi}_{m_n}^{\mathrm{CG}}, \mathscr{R}\hat{\psi}_{m_n}^{\mathrm{CG}}\rangle - \langle \psi_{m_n}^{\mathrm{CG}}, \mathscr{R}\psi_{m_n}^{\mathrm{CG}}\rangle = O_{\mathrm{p}}\big(n^{-1/2}\omega_{m_n}^{-1}\|\gamma^{(m_n)}\| + n^{-1}\omega_{m_n}^{-3}\big). \tag{A4}$$

This last result is analogous to (7.27) of Delaigle & Hall (2012b), whose proof can be repeated with the same modifications for our situation as before. Therefore, regardless of whether $\|\mathscr{R}^{-1}\mu\|$ or $\|\mathscr{R}^{-1/2}\mu\|$ is finite or infinite, the theoretical and empirical regularized quantities approach each other at the rates given in (A3) and (A4). The result on $D(\hat{\psi}_{m_n}^{\mathrm{CG}})$ then follows as in the proof of Proposition 1.

Proof of Theorem 2

We show that $D(\hat{\psi}^{\mathrm{PC}}_{m_n}) = 1-\Phi\{|Z(\hat{\psi}^{\mathrm{PC}}_{m_n})|/2\}$ converges in probability to $1-\Phi(\|\mathscr{R}^{-1/2}\|/2)$. The strategy of the proof is similar to that of Theorem 3.1 in Cardot et al. (1999) for the principal component approach to the functional linear model. The difference lies in the incompleteness of the functional data and in that we do not assume that the underlying theoretical inverse problem has a solution. We write

$$\|\hat{\psi}^{\mathrm{PC}}_{m_n} - \psi^{\mathrm{PC}}_{m_n}\| \leqslant \|\hat{\mathscr{R}}^{-}_{m_n} - \mathscr{R}^{-}_{m_n}\|_\infty \|\hat{\mu}\| + \|\mathscr{R}^{-}_{m_n}\|_\infty \|\hat{\mu} - \mu\|.$$

Proceeding as in the proof of Lemma 5.1 in Cardot et al. (1999), we can show that

$$\|\hat{\mathscr{R}}^{-}_{m_n} - \mathscr{R}^{-}_{m_n}\|_\infty \leqslant \hat{\lambda}^{-1}_{m_n}\lambda^{-1}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty + 2\lambda^{-1}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty \sum_{j=1}^{m_n} a_j.$$

Here $\hat{\lambda}_j$ are the eigenvalues of $\hat{\mathscr{R}}$ in descending order and $\hat{\varphi}_j$ are the corresponding eigenfunctions. In establishing the above inequality one uses the facts that $|\hat{\lambda}_j - \lambda_j| \leqslant \|\hat{\mathscr{R}} - \mathscr{R}\|_\infty$ and $\|\hat{\varphi}_j - \mathrm{sign}\langle \hat{\varphi}_j, \varphi_j\rangle \varphi_j\| \leqslant a_j\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty$, which are known from Bosq (2000, Lemmas 4.2 and 4.3) for the empirical covariance operator from complete curves but hold also for functional fragments; see the proof of Proposition 2 in the supplementary document for Kraus (2015). Since $\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = O_{\mathrm{p}}(n^{-1/2})$, we see that $\hat{\lambda}^{-1}_{m_n}\lambda^{-1}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty 1_{[\hat{\lambda}_{m_n} > \lambda_{m_n}/2]} \leqslant 2\lambda^{-2}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2})$. Since the probability of the event $[\hat{\lambda}_{m_n} < \lambda_{m_n}/2]$ is bounded by $\lambda^{-2}_{m_n} O(n^{-1})$ and hence converges to 0, it follows that $\hat{\lambda}^{-1}_{m_n}\lambda^{-1}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2})$. Combining this with the facts that $\|\hat{\mu}\| = O_{\mathrm{p}}(1)$, $\|\mathscr{R}^{-}_{m_n}\| = \lambda^{-1}_{m_n}$ and $\|\hat{\mu} - \mu\| = O_{\mathrm{p}}(n^{-1/2})$ gives

$$\|\hat{\psi}^{\mathrm{PC}}_{m_n} - \psi^{\mathrm{PC}}_{m_n}\| \leqslant \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2}) + \lambda^{-1}_{m_n} O_{\mathrm{p}}(n^{-1/2}) \sum_{j=1}^{m_n} a_j.$$

Similar arguments can be used in the analysis of the denominator in $Z(\hat{\psi}^{\mathrm{PC}}_{m_n})$. In conclusion, we obtain that the estimation errors for the quantities in the numerator and denominator converge to zero at the rates

$$\langle \mu, \hat{\psi}^{\mathrm{PC}}_{m_n}\rangle - \langle \mu, \psi^{\mathrm{PC}}_{m_n}\rangle = \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2}) + \lambda^{-1}_{m_n} O_{\mathrm{p}}(n^{-1/2}) \sum_{j=1}^{m_n} a_j, \tag{A5}$$

$$\langle \hat{\psi}^{\mathrm{PC}}_{m_n}, \mathscr{R}\hat{\psi}^{\mathrm{PC}}_{m_n}\rangle - \langle \psi^{\mathrm{PC}}_{m_n}, \mathscr{R}\psi^{\mathrm{PC}}_{m_n}\rangle = \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2}) + \lambda^{-1}_{m_n} O_{\mathrm{p}}(n^{-1/2}) \sum_{j=1}^{m_n} a_j. \tag{A6}$$

In light of (A5) and (A6), the asymptotic behaviour of the misclassification probability is driven by the behaviour of the theoretical classifier addressed in Proposition 1.

Proof of Theorem 3

We show that the fraction $|Z(\hat{\psi}^{\mathrm{R}}_{m_n})|$ converges in probability to $\|\mathscr{R}^{-1/2}\mu\|/2$ as $n \to \infty$. For the numerator we write

$$\langle \mu, \hat{\psi}^{\mathrm{R}}_{\alpha_n}\rangle - \langle \mu, \mathscr{R}^{-1}_{\alpha_n}\mu\rangle = \langle \mu, (\hat{\mathscr{R}}^{-1}_{\alpha_n} - \mathscr{R}^{-1}_{\alpha_n})\hat{\mu}\rangle + \langle \mu, \mathscr{R}^{-1}_{\alpha_n}(\hat{\mu} - \mu)\rangle. \tag{A7}$$

For the first term on the right we find that

$$\begin{aligned} |\langle \mu, (\hat{\mathscr{R}}^{-1}_{\alpha_n} - \mathscr{R}^{-1}_{\alpha_n})\hat{\mu}\rangle| &\leqslant \|\mu\|\|\hat{\mathscr{R}}^{-1}_{\alpha_n} - \mathscr{R}^{-1}_{\alpha_n}\|_\infty\|\hat{\mu}\| \\ &= \|\mu\|\|\hat{\mathscr{R}}^{-1}_{\alpha_n}(\hat{\mathscr{R}}_{\alpha_n} - \mathscr{R}_{\alpha_n})\mathscr{R}^{-1}_{\alpha_n}\|_\infty\|\hat{\mu}\| \end{aligned}$$

$$\leqslant \|\mu\|\|\hat{\mathscr{R}}_{\alpha_n}^{-1}\|_\infty \|\hat{\mathscr{R}}_{\alpha_n} - \mathscr{R}_{\alpha_n}\|_\infty \|\mathscr{R}_{\alpha_n}^{-1}\|_\infty \|\hat{\mu}\|$$
$$\leqslant \alpha_n^{-2} O_{\mathrm{p}}(n^{-1/2}),$$

since $\|\hat{\mathscr{R}}_{\alpha_n}^{-1}\|_\infty \leqslant \alpha_n^{-1}$, $\|\mathscr{R}_{\alpha_n}^{-1}\|_\infty \leqslant \alpha_n^{-1}$, $\|\hat{\mu}\| = O_{\mathrm{p}}(1)$ and $\|\hat{\mathscr{R}}_{\alpha_n} - \mathscr{R}_{\alpha_n}\|_\infty = \|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = O_{\mathrm{p}}\{(n_0 + n_1)^{-1/2}\}$ (Kraus, 2015, Proposition 1). For the second term on the right-hand side of (A7), we obtain

$$|\langle \mu, \mathscr{R}_{\alpha_n}^{-1}(\hat{\mu} - \mu)\rangle| \leqslant \|\mu\|\|\mathscr{R}_{\alpha_n}^{-1}\|_\infty \|\hat{\mu} - \mu\| \leqslant \alpha_n^{-1} O_{\mathrm{p}}(n^{-1/2}).$$

The quantity in the denominator of $Z(\hat{\psi}_{m_n}^{\mathrm{R}})$ can be rewritten as

$$\langle \hat{\psi}_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle - \langle \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\psi_{\alpha_n}^{\mathrm{R}}\rangle = \langle \hat{\psi}_{\alpha_n}^{\mathrm{R}} - \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle + \langle \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}(\hat{\psi}_{\alpha_n}^{\mathrm{R}} - \psi_{\alpha_n}^{\mathrm{R}})\rangle. \tag{A8}$$

The first term on the right is

$$\begin{aligned}\langle \hat{\psi}_{\alpha_n}^{\mathrm{R}} - \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle &= \langle \hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu} - \mathscr{R}_{\alpha_n}^{-1}\mu, \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle \\ &= \langle \mathscr{R}_{\alpha_n}^{-1}(\mathscr{R}_{\alpha_n} - \hat{\mathscr{R}}_{\alpha_n})\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}, \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle + \langle \mathscr{R}_{\alpha_n}^{-1}(\hat{\mu} - \mu), \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle. \end{aligned} \tag{A9}$$

For the first summand in (A9) we have

$$\begin{aligned}|\langle \mathscr{R}_{\alpha_n}^{-1}(\mathscr{R}_{\alpha_n} - \hat{\mathscr{R}}_{\alpha_n})\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}, \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle| &\leqslant \|\hat{\mu}\|^2 \|\hat{\mathscr{R}}_{\alpha_n}^{-1}\|_\infty^2 \|\mathscr{R}\mathscr{R}_{\alpha_n}^{-1}\|_\infty \|\hat{\mathscr{R}} - \mathscr{R}\|_\infty \\ &\leqslant \alpha_n^{-2} O_{\mathrm{p}}(n^{-1/2}),\end{aligned}$$

using properties mentioned previously and the fact that $\|\mathscr{R}\mathscr{R}_{\alpha_n}^{-1}\|_\infty \leqslant 1$, and for the second summand we have

$$|\langle \mathscr{R}_{\alpha_n}^{-1}(\hat{\mu} - \mu), \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle| \leqslant \|\mathscr{R}\mathscr{R}_{\alpha_n}^{-1}\|_\infty \|\hat{\mathscr{R}}_{\alpha_n}^{-1}\|_\infty \|\hat{\mu} - \mu\| \leqslant \alpha_n^{-1} O_{\mathrm{p}}(n^{-1/2}).$$

Putting these results together, we see that the absolute value of the first term on the right-hand side of (A8) is dominated by $\alpha_n^{-2} O_{\mathrm{p}}(n^{-1/2})$. The second term on the right-hand side of (A8) can be analysed in a similar way to the first two terms on the right-hand side of (A7) with $\mathscr{R}\mathscr{R}_{\alpha_n}^{-1}\mu$ in place of μ. Thus we bound the absolute value from above by $\alpha_n^{-2} O_{\mathrm{p}}(n^{-1/2})$. These results imply that the estimation errors vanish at rates

$$\langle \mu, \hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle - \langle \mu, \psi_{\alpha_n}^{\mathrm{R}}\rangle = \alpha_n^{-2} O_{\mathrm{p}}(n^{-1/2}),$$
$$\langle \hat{\psi}_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle - \langle \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\psi_{\alpha_n}^{\mathrm{R}}\rangle = \alpha_n^{-2} O_{\mathrm{p}}(n^{-1/2}).$$

Hence the empirical classifier has the same limiting error as the theoretical one addressed in Proposition 3.

References

Baíllo, A., Cuevas, A. & Cuesta-Albertos, J. A. (2011a). Supervised classification for a family of Gaussian functional models. *Scand. J. Statist.* **38**, 480–98.

Baíllo, A., Cuevas, A. & Fraiman, R. (2011b). Classification methods for functional data. In *The Oxford Handbook of Functional Data Analysis*, F. Ferraty & Y. Romain, eds. Oxford: Oxford University Press, pp. 259–97.

Berrendero, J. R., Cuevas, A. & Torrecilla, J. L. (2016). Variable selection in functional data classification: A maxima-hunting proposal. *Statist. Sinica* **26**, 619–38.

Berrendero, J. R., Cuevas, A. & Torrecilla, J. L. (2018). On the use of reproducing kernel Hilbert spaces in functional classification. *J. Am. Statist. Assoc.* **113**, 1210–8.

Blanchard, G. & Krämer, N. (2010). Kernel partial least squares is universally consistent. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh & M. Titterington, eds., vol. 9 of *Proceedings of Machine Learning Research*. International Joint Conferences on Artificial Intelligence (IJCAI) Organization, pp. 57–64.

BONGIORNO, E. G. & GOIA, A. (2016). Classification methods for Hilbert data based on surrogate density. *Comp. Statist. Data Anal.* **99**, 204–22.
BOSQ, D. (2000). *Linear Processes in Function Spaces*. New York: Springer.
BUGNI, F. A. (2012). Specification test for missing functional data. *Economet. Theory* **28**, 959–1002.
CARDOT, H., FERRATY, F. & SARDA, P. (1999). Functional linear model. *Statist. Prob. Lett.* **45**, 11–22.
CUESTA-ALBERTOS, J. A., DEL BARRIO, E., FRAIMAN, R. & MATRÁN, C. (2007). The random projection method in goodness of fit for functional data. *Comp. Statist. Data Anal.* **51**, 4814–31.
CUEVAS, A. (2014). A partial overview of the theory of statistics with functional data. *J. Statist. Plan. Infer.* **147**, 1–23.
DAI, X., MÜLLER, H.-G. & YAO, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika* **104**, 545–60.
DE JONG, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemomet. Intel. Lab. Syst.* **18**, 251–63.
DELAIGLE, A. & HALL, P. (2012a). Achieving near perfect classification for functional data. *J. R. Statist. Soc.* B **74**, 267–86.
DELAIGLE, A. & HALL, P. (2012b). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40**, 322–52.
DELAIGLE, A. & HALL, P. (2013). Classification using censored functional data. *J. Am. Statist. Assoc.* **108**, 1269–83.
DELAIGLE, A. & HALL, P. (2016). Approximating fragmented functional data by segments of Markov chains. *Biometrika* **103**, 779–99.
DELAIGLE, A., HALL, P. & BATHIA, N. (2012). Componentwise classification and clustering of functional data. *Biometrika* **99**, 299–313.
DESCARY, M.-H. & PANARETOS, V. M. (2019). Recovering covariance from functional fragments. *Biometrika* **106**, 145–60.
FEBRERO-BANDE, M., GALEANO, P. & GONZÁLEZ-MANTEIGA, W. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *Int. Statist. Rev.* **85**, 61–83.
FERRATY, F., HALL, P. & VIEU, P. (2010). Most-predictive design points for functional data predictors. *Biometrika* **97**, 807–24.
GOLDBERG, Y., RITOV, Y. & MANDELBAUM, A. (2014). Predicting the continuation of a function with applications to call center data. *J. Statist. Plan. Infer.* **147**, 53–65.
GROMENKO, O., KOKOSZKA, P. & SOJKA, J. (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Ann. Appl. Statist.* **11**, 898–918.
HASTIE, T. J., TIBSHIRANI, R. J. & FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning*. New York: Springer, 2nd ed.
HORVÁTH, L. & KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. New York: Springer.
KRAUS, D. (2015). Components and completion of partially observed functional data. *J. R. Statist. Soc.* B **77**, 777–801.
LIEBL, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Statist.* **7**, 1562–92.
LINGJÆRDE, O. C. & CHRISTOPHERSEN, N. (2000). Shrinkage structure of partial least squares. *Scand. J. Statist.* **27**, 459–73.
PHATAK, A. & DE HOOG, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: Alternative proofs of some properties of PLS. *J. Chemomet.* **16**, 361–7.
PINI, A. & VANTINI, S. (2016). The interval testing procedure: A general framework for inference in functional data analysis. *Biometrics* **72**, 835–45.
RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis*. New York: Springer, 2nd ed.
SANGALLI, L. M., SECCHI, P. & VANTINI, S. (2014a). Analysis of AneuRisk65 data: k-mean alignment. *Electron. J. Statist.* **8**, 1891–904.
SANGALLI, L. M., SECCHI, P. & VANTINI, S. (2014b). AneuRisk65: A dataset of three-dimensional cerebral vascular geometries. *Electron. J. Statist.* **8**, 1879–90.
SANGALLI, L. M., SECCHI, P., VANTINI, S. & VENEZIANI, A. (2009a). A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *J. Am. Statist. Assoc.* **104**, 37–48.
SANGALLI, L. M., SECCHI, P., VANTINI, S. & VENEZIANI, A. (2009b). Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines. *J. R. Statist. Soc.* C **58**, 285–306.
STEFANUCCI, M., SANGALLI, L. M. & BRUTTI, P. (2018). PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set. *Statist. Neer.* **72**, 246–64.

[*Received on* 22 *August* 2017. *Editorial decision on* 2 *August* 2018]