TEST (2016) 25:1–22
DOI 10.1007/s11749-015-0470-2

1 Introduction

In this paper, we consider the classification of sparse functional data by coupling tools from effective dimension reduction and support vector machine. For ease of exposition, we begin with a brief review on these topics. Along the review, we also lay down the motivation for the proposed methodology and describe the connection to various relevant topics.

1.1 Functional Data Analysis

Functional data analysis (FDA) has attracted substantial research interest over the past decades, owing to vastly emerging complex data consisting of curves or other infinite-dimensional objects. Originally functional data refer to a sample of fully observed random trajectories, such as those obtained in automatic monitoring or recording systems; for a general introduction, see Ramsay and Silverman (2002, 2005). Some recent contributions for systematically documenting the advances in this area include the monographs by Ferraty and Vieu (2006) and Horváth and Kokoszka (2012). The former focused on nonparametric modelling of functional data coupled with suitable semi-metrics and kernel estimation methods, whilst the latter emphasized inference developments on the structures and regression relationship for both independent and

differential equations, whereas Jones and Rice (1992) emphasized applications. Various theoretical properties have been studied by Silverman (1996), Boente and Fraiman (2000), Kneip and Utikal (2001) and Hall and Hosseini-Nasab (2006), amongst others. A relevant topic in FDA is to model the variation in time in addition to amplitude, caused by different developmental pace or biological clock of each experimental unit. This has been addressed by time warping or curve alignment (Gasser and Kneip 1995; Ramsay and Li 1998; Gervini and Gasser 2004; Kneip and Ramsay 2008), and is often considered as preprocessing for further functional data modelling.

Difficulty arises when applying the FDA methodology based on fully observed or smoothed curves in the situation of only a few repeated measures available per experimental unit. To overcome this challenge, Yao et al. (2005a) proposed the principal analysis by conditional expectation (PACE) that introduced a unified framework for sparse/dense designs by borrowing information across the entire sample whilst estimating the population characteristics, such as the mean, covariance and eigenfunctions/values. For recovering individual trajectories, PACE resembles the best linear unbiased prediction (BLUP) of linear mixed-effects (LME) models in the context of FPCA. Yao and Lee (2006) proposed an iterative FPCA procedure to reduce within-subject dependence, whilst Hall et al. (2006) studied the theoretical aspect for sparse functional data. Another line of approaches dealing with sparse functional data stemmed from spline basis representations coupled with LME implementation (Shi et al. 1996; Rice and Wu 2001; Guo 2002), and further extended to wavelet basis (Morris et al. 2003; Morris and Carroll 2006). James et al. (2000) proposed a reduced rank model by expressing the eigenfunctions with splines, and stimulated the penalized approach for modelling spatially correlated functions (Zhou et al. 2010).

1.2

Based on these representation methods for functional data, a great deal of research has been conducted on regression models characterized by inclusion of a functional predictor or a functional response or both. In view of the close connection to functional classification, we focus on reviewing the case that linearly associates a scalar response with a functional predictor,

$$E(Y|X) = \alpha + \int_T \{X(t) - \mu_X(t)\}\beta(t)\mathrm{d}t, \tag{1}$$

where $Y \in \mathbb{R}$ is the scalar response, X is the predictor process that is assumed to reside in $L^2(T)$ with smooth trajectories and the mean function $\mu_X(t)$, $\alpha \in \mathbb{R}$ is the intercept and $\beta \in L^2(T)$ is the squared integrable regression parameter function. The domain T is often assumed to be a compact interval that denotes time, and may also correspond to other index variables. A variety of implementations and asymptotic results have been developed for this functional linear model (Faraway 1997; Cuevas et al. 2002; Cardot et al. 2003a, b; Yuan and Cai 2010), including optimality considerations (Cai and Hall 2006; Hall and Horowitz 2007; Cai and Yuan 2012).

An important extension of (1) is the generalized functional linear model (GFLM) with the response often to be a discrete outcome, such as binomial or Poisson (James 2002; Escabias et al. 2004; Müller and Stadtmüller 2005). With a monotonic and invertible link function g and a variance function $\mathrm{var}(Y|X) = V\{E(Y|X)\}$, the GFLM may be written as

$$E(Y|X) = g\left(\alpha + \int_T \{X(t) - \mu_X(t)\}\beta(t)\mathrm{d}t\right), \tag{2}$$

and has been studied for both known and unknown link/variance functions in Müller and Stadtmüller (2005). This model framework can be immediately applied to classification problems with a binary or multi-level response.

Classification is an important problem in FDA, where the data consist of trajectories $X_i \in L^2(T)$ and the labels Y_i take values on $\{-1, 1\}$, $i = 1, \ldots, n$. The aim is to decide which class a new observation X will be assigned to based on the collected data. This topic is under rapid development and has been extensively studied for completely or densely observed functional data. For instance, Ferraty and Vieu (2003) and Biau et al. (2005) studied the classification from a nonparametric perspective, treating the fully observed random functions a random variable in Hilbert space without dimension reduction. Leng and Müller (2006) considered a FPC-based method for gene expression profiles, whilst Cuevas et al. (2007) explored the robust aspect via projection-based depth notions. Delaigle and Hall (2012) proposed a centroid classifier based on a reduced representation via FPC scores or partial least squares.

By contrast, the research on classification for sparse functional data is relatively scanty, when only a few measurements are available for some, even all, individuals. James and Hastie (2001) suggested an extension of linear discriminant analysis for such data using a finite-dimensional spline approximation. Müller (2005) conducted functional principal component analysis coupled with a logistic regression. Wu and Liu (2013) applied support vector machines after performing functional principal component analysis. Unfortunately, none of these methods utilized the relationship between the observed trajectories and the associated labels whilst conducting dimension reduction. In this paper, we will tackle classification of sparsely observed functional data, not only utilizing the information of the predictor process but also considering the response variable to improve the classification accuracy.

1.3 ff

Our goal is to properly make use of the joint information for classifying sparse functional data. It is known that the effective dimension reduction (EDR) methods, originated from multivariate regression, have been proven useful in this regard. It is attractive to find the most effective directions $\beta_1, \ldots, \beta_K$ based on both the covariate X and the response Y, where K is the unknown structural dimension. Such direction functions $\beta_1, \ldots, \beta_K$ are also called index functions, and the model is referred to as

functional index model, with an unknown link function g and the model error ϵ,

$$Y = g\left(\langle \beta_1, X\rangle, \ldots, \langle \beta_K, X\rangle; \epsilon\right), \tag{3}$$

whilst $K = 1$ corresponds to a single index model. This is the general form of functional regression without structure assumption, whilst the functional linear model is a special case when $K = 1$ and g is linear.

One may conduct direct statistical estimation of the index and link functions under the general form (3), for example, Xia et al. (2002) proposed the minimum average variance estimation approach based on kernel methods for multivariate data. Instead of estimating $\beta_1, \ldots, \beta_K$ and g altogether, often it is of more interest to learn the low-dimensional projections $\langle \beta_1, X\rangle, \ldots \langle \beta_K, X\rangle$ that form a sufficient statistic, which is the aim of EDR-based methods, i.e., characterizing the so-called EDR space $S_{Y|X} = \text{span}(\beta_1, \ldots, \beta_K)$, also known as the central subspace (Cook 1998). Thus, one major advantage of EDR methods is "link-free" (Duan and Li 1991). Pioneered by Li (1991) that proposed the sliced inverse regression (SIR) using the information concerning the inverse conditional mean $E(X|Y)$, Cook and Weisberg (1991) considered the inverse variance estimation utilizing the information of $\text{var}(X|Y)$, Li (1992) dealt with the Hessian matrix of the regression curve, Chiaromonte et al. (2002) modified sliced inverse regression for categorical predictors, Li and Wang (2007) worked with empirical directions, and Zhu et al. (2010) proposed cumulative slicing estimation to improve upon SIR.

Since dimension reduction is particularly useful for modelling functional data that reside in an infinite-dimensional space, Ferré and Yao (2003) applied the SIR to complete or dense functional data, based on which Li and Hsing (2010) proposed a test to decide the dimensionality of EDR space. Besides EDR methods, James and Silverman (2005) and Chen et al. (2011) estimated the index functions $\beta_1, \ldots, \beta_K$ and additive link function $g_1, \ldots, g_K$ jointly under the form

$$g(\langle \beta_1, X\rangle, \ldots \langle \beta_K, X\rangle; \epsilon) = \beta_0 + \sum_{k=1}^{K} g_k(\langle \beta_k, X\rangle) + \epsilon.$$

However, these existing methods are not applicable to sparse functional data. Recently, Jiang et al. (2014) proposed an inverse regression method for sparse functional data by estimating the inverse conditional mean functions with a two-dimensional smoother that requires considerable computation. Inspired by cumulative slicing estimation (CUME) for multivariate data (Zhu et al. 2010), Yao et al. (2015) proposed to borrow information across subjects via a one-dimensional smoother, named the functional cumulative slicing (FCS), which is closely related to the proposed method in this paper. The EDR methods are intended for continuous response and have been rarely used for classification problems due to few distinct response values. In other words, homogeneity in partitioning Y values fails to capture sufficient variability to estimate the EDR space. In this paper, we investigate this problem by exploring the idea of the support vector machine (SVM) that finds a separation boundary between two classes.

1.4

The SVM originates from the classical Perceptron algorithm (Rosenblatt 1958, 1962) that is one of the first binary linear classifiers and is capable of performing online learning. Along this line, Vapnik and Lerner (1963) introduced the generalized portrait algorithm. With the kernel trick (Aronszajn 1950), the SVM is a nonlinear generalization of the generalized portrait algorithm. Aizerman et al. (1964) introduced the geometric interpretation of the kernel as inner product in a feature space implicitly defined by the kernel. Cover (1965) introduced the concept of large-margin hyperplanes. Close to its current form, the SVM was first introduced by Boser et al. (1992). Its soft margin version was introduced by Cortes and Vapnik (1995). For a detailed exposition of the SVM, interested readers may read Vapnik (1995, 1998), Cristianini and Shawe-Taylor (2000) and references therein.

As an auxiliary tool, the SVM paradigm provides a geometric interpretation of differentiating two classes by a hyperplane with maximal separation margin in the input space. Owing to its flexibility and capability in dealing with high-dimensional data, SVM classifier received considerable attention and was also generalized from binary to multicategory settings, see Weston and Watkins (1999), Bredensteiner and Bennett (1999), Crammer and Singer (2001), Lee et al. (2004), Wang and Shen (2006, 2007), Liu and Shen (2006), Wu and Liu (2007), Liu and Yuan (2011), Chang et al. (2011), He et al. (2012) and references therein.

In the standard SVM, observations from different classes are weighted equally whilst training the classifier. However, this may not be optimal especially in the unbalanced case with one class dominating the other. Thus motivated, Lin et al. (2004) proposed weighted SVM (WSVM) by weighting observations from different classes with different weights in the training process and established its Fisher consistency. Based on the WSVM's Fisher consistency, Wang et al. (2008) proposed a probability estimation scheme to estimate the conditional probability for each new observation belonging to each class. Particularly inspired from this probability estimation scheme using the WSVM to address the aforementioned homogeneity issue in partitioning a binary response, Shin et al. (2014) proposed a probability-enhanced dimension reduction method for multivariate data.

In this paper, we target at functional data and propose an integrated classification procedure called probability-enhanced functional cumulative slicing (PEFCS) for sparse functional data. The key idea is, instead of directly partitioning the response values, to conduct functional cumulative slicing based on ranking the underlying probabilities $p(X) = p\{Y = 1|X(\cdot)\}$ that are obtainable from the WSVM. The resultant slices would have sufficient heterogeneity to capture the variability and recover the EDR space. As we illustrate later, the PEFCS adopts a pooling strategy combining information from all individuals to handle sparse functional data.

The rest of article is organized as follows. In Sect. 2, we describe the proposed probability-enhanced method coupled with the functional cumulative slicing, whilst Sect. 3 presents an estimation procedure using sparse functional data. Section 4 provides a simulation study, and Sect. 5 offers three real data examples. Concluding remarks are given in Sect. 6.

2 Functional data

The data considered consist of random trajectories that are independent realizations of a smooth process $X(t)$ that is defined on a compact interval T and belongs to a separable Hilbert space $H \equiv L_2(T)$

choice of cumulative slicing is thus critical to ensure sufficient number of observations. Yao et al. (2015) extended this principle based on a two-slice scheme,

$$m(\cdot, \tilde{y}) = E\{X(\cdot)I(Y \leq \tilde{y})\},$$

across all possible values $\tilde{y}$ in the range of Y to maximize the utility of the data. As a consequence, the EDR space is characterized by

$$\Lambda(s,t) = E\{m(s,\tilde{Y})m(t,\tilde{Y})\},$$

where $\tilde{Y}$ is an independent copy of Y. Unfortunately, for functional classification, homogeneity exists owing to only two possible slices based on a binary response. Thus, at most one direction of $S_{Y|X}$ can be recovered. To overcome this difficulty for multivariate data, Shin et al. (2014) proposed to construct slices based on the conditional probabilities $p(\ ) = p(Y = 1|\ )$, where denotes a p-dimensional vector of multivariate covariates. This introduced an equivalent central subspace, denoted by $S_{p(\ )|\ }$, which is the intersection of the spaces spanned by all $K \times p$ matrices satisfying $p(\ ) \perp \ |\ ^{\top}$.

Our goal is to benefit from data pooling with the functional cumulative slicing and enhance the EDR estimation based on the underlying conditional probability for classifying sparsely observed functional data. Similar to the multivariate case, we define the central subspace $S_{p(X)|X}$ based on $p(X) = p\{Y = 1|X(\cdot)\}$ as the intersection of spaces spanned by all sets of $\{\beta_1, \ldots, \beta_K\}$ satisfying $p(X) \perp X|\langle\beta_1, X\rangle, \ldots, \langle\beta_K, X\rangle$. We then establish the following equivalence between the central subspaces based on the conditional probability and that based on the binary response. Given $X(t) = x(t)$, the binary response Y can be equivalently expressed as a function satisfying

$$Y\{p(x), \varepsilon^*\} = \begin{cases} 1, & \varepsilon^* \leq p(x) \\ -1, & \text{otherwise,} \end{cases}$$

where ε^* is a random noise $\varepsilon^* \sim U(0,1)$ and independent of $X(t)$.

1 *$p(X)$ contains the same information as X to predict Y, i.e., $S_{Y|X} = S_{p(X)|X}$.*

This proposition is a functional version of Lemma 1 in Shin et al. (2014), and its proof is given in the "Appendix". Hence we may estimate $S_{p(X)|X}$ instead of $S_{Y|X}$ to improve Shin at le 2014

To recover the EDR space, it is necessary to run π across the support of $p(\tilde{X})$, with $\tilde{X}$ being an independent copy of X

A , 3

$$\sum_{j=1}^{\infty}\sum_{l=1}^{\infty}\alpha_j^{-2}\alpha_l^{-1}E^2(E[\xi_j I\{p(X)\le p(\tilde{X})\}|p(\tilde{X})]E[\xi_l I\{p(X)\le p(\tilde{X})\}|p(\tilde{X})])<\infty.$$

2 *Under Assumptions 1–3, the eigenspace associated with the K nonzero eigenvalues of $\Sigma^{-1}\Lambda$ is well defined in H.*

This proposition guarantees the validity of the proposed PEFCS, which is a direct analogue to Theorem 4.8 in He et al. (2003) and Theorem 2.1 in Ferré and Yao (2005).

3 **f** **f**

In light of the equivalence $S_{Y|X}$=$S_{p(X)|X}$, it suffices to know whether $p(X_i) \le \pi$ for an arbitrary $0 < \pi < 1$, $i = 1, \ldots, n$. Lin et al. (2004) showed that the solution to the WSVM has an important property, the Fisher consistency. Hence, by training a sequence of WSVMs for different values of π, we are able to consistently estimate $I\{p(X_i) \le \pi\}$. The trajectories X_i are observed intermittently with noise, and collected in the form of $(t_{ij}, Uij$

$$\min_{g_\pi \in \mathcal{F}_K} (1-\pi) \sum_{i:Y_i=1} H_1\{Y_i g_\pi(\hat{X}_i)\} + \pi \sum_{i:Y_i=-1} H_1\{Y_i g_\pi(\hat{X}_i)\} + \frac{\lambda}{2}||g_\pi||^2_{\mathcal{F}_K}, \quad (5)$$

where $H_1(u) = \max\{1-u, 0\}$ is the hinge loss function, $||g_\pi||^2_{\mathcal{F}_K}$ is the penalty term to regulate the complexity of g_π, and $\lambda > 0$ is a tuning parameter which controls the trade-off between data-fit and model complexity. By the representer theorem (Kimeldorf and Wahba 1971), the solution to (5) has a finite representation

$$g_\pi(x) = d_\pi + \lambda^{-1} \sum_{i=1}^{n} d_{i,\pi} Y_i K(x, \hat{X}_i)$$

for $x \in H$, and

$$||g_\pi||^2_{\mathcal{F}_K} = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{i,\pi} d_{j,\pi} Y_i Y_j K(\hat{X}_i, \hat{X}_j).$$

Thus, solving the optimization problem

$$\min_{d_\pi, d_{1,\pi}, \ldots, d_{n,\pi}} (1-\pi) \sum_{i:Y_i=1} H_1\{Y_i g_\pi(\hat{X}_i)\} + \pi \sum_{i:Y_i=-1} H_1\{Y_i g_\pi(\hat{X}_i)\}$$
$$+ \frac{1}{2\lambda} \sum_{i=1}^{n}\sum_{j=1}^{n} d_{i,\pi} d_{j,\pi} Y_i Y_j K(\hat{X}_i, \hat{X}_j) \quad (6)$$

provides an estimate $\text{sign}(\hat{g}_\pi(\cdot))$ for $\text{sign}\{p(x) - \pi\}$, i.e.,

$$\hat{I}\{p(x) < \pi\} = 2^{-1}\{1 - \text{sign}(\hat{g}_\pi(x))\}.$$

More details can be found in Lin et al. (2004) and Shin et al. (2014).

We next estimate the unconditional mean $m(t, \pi) = E[X(t)I\{p(X) \le \pi\}]$ using the strategy of cumulative slicing and borrowing information across subjects. We use a local linear estimator $\hat{m}(t, \pi) = \hat{a}_0$ (Fan and Gijbels 1996), given by

$$\min_{a_0, a_1} \sum_{i=1}^{n}\sum_{j=1}^{n_i} [U_{ij} \hat{I}\{p(\hat{X}_i) \le \pi\} - a_0 - a_1(t_{ij} - t)]^2 K_1\left(\frac{t_{ij} - t}{h_1}\right), \quad (7)$$

where $\hat{X}_i$ is the PACE estimate of X_i, K_1 is a non-negative and symmetric univariate kernel density, and $h_1 = h_1(n)$ is the bandwidth to control the amount of smoothing. We follow the suggestion of ignoring the dependency amongst the data from the same individual (Lin and Carroll 2000), and use leave-one-curve-out cross-validation to select h_1 (Rice and Silverman 1991). To estimate $\Lambda(s, t)$ in (4) by running π across $\{p(\hat{X}_i) : i = 1, \ldots, n\}$, it only requires the ranking of such values. One may simply use a dense set $0 < \pi_1 < \cdots < \pi_m < 1$, and take the centre of the nearest interval

containing $p(\hat{X}_i)$ as a surrogate, denoted by $\tilde{p}(\hat{X}_i)$. Then, an estimator of the kernel function $\Lambda(s,t)$ is

$$\hat{\Lambda}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}\{s, \tilde{p}(\hat{X}_i)\}\hat{m}\{t, \tilde{p}(\hat{X}_i)\}w\{\tilde{p}(\hat{X}_i)\}. \tag{8}$$

For the covariance operator Σ, following Yao et al. (2005b), define $C_i(t_{ij}, t_{il}) = U_{ij}U_{il}$ that satisfies $E\{C_i(t_{ij}, t_{il})\} = \Sigma(t_{ij}, t_{il}) + \sigma_\varepsilon^2 I_{ij}$, where $I_{ij} = 1$ if $i = j$ and 0 otherwise. Thus, we remove the diagonal raw covariances and the local linear estimator is given by $\hat{\Sigma}(s,t) = \hat{b}_0$, solving

$$\min_{b_0,b_1,b_2} \sum_{i=1}^{n}\sum_{j\neq l}^{n_i}\{C_i(t_{ij}, t_{il}) - b_0 - b_1(t_{ij} - s) - b_2(t_{il} - t)\}^2 \times K_2\left(\frac{t_{ij} - s}{h_2}, \frac{t_{il} - t}{h_2}\right), \tag{9}$$

where K_2 is a non-negative and symmetric bivariate kernel density and h_2 is the bandwidth also chosen by the leave-one-curve-out cross-validation. Then, we use a sequence of finite rank operator estimators $\Sigma_{s_n}^{-1} = \sum_{j=1}^{s_n} \alpha_j^{-1}\phi_j \otimes \phi_j$ (respectively $\hat{\Sigma}_{s_n}^{-1} = \sum_{j=1}^{s_n} \hat{\alpha}_j^{-1}\hat{\phi}_j \otimes \hat{\phi}_j$) to approximate the unbounded Σ^{-1}. Hence, the eigenfunctions associated with the K largest non-zero eigenvalues of $\hat{\Sigma}_{s_n}^{-1}\hat{\Lambda}$ are obtained as the estimates of $\{\beta_k\}_{k=1,\ldots,K}$.

For completely observed X_i, the estimation procedure is simpler, and the quantities are estimated by their sample moments,

$$\hat{m}(t,\pi) = n^{-1}\sum_{i=1}^{n} X_i(t)\hat{I}\{p(X_i) \leq \pi\},$$

$$\hat{\Sigma}(s,t) = n^{-1}\sum_{i=1}^{n} X_i(s)X_i(t),$$

whilst $\hat{\Lambda}$ remains the same as (8). For densely observed X_i, the error introduced by individual smoothing has been shown asymptotically negligible, thus is equivalent to the case of completely observed X_i (Hall et al. 2006). To select the tuning parameter λ in WSVM, the structural dimension K and the truncation s_n, we choose them together by minimizing the (cross-validated) classification error in our simulated and real examples since classification is the primary goal.

4

As a dimension reduction tool, the proposed PEFCS projects the functional data onto a feature space, in a similar spirit of FPCA. In this section, we apply different classifiers to the reduced data in such feature spaces obtained from either PEFCS or FPCA . To be

specific, we consider the linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), the logistic additive regression, and the centroid method proposed by Delaigle and Hall (2012). These classifiers are estimated from a training sample based on projected data $\langle \hat{\psi}, X_i \rangle$, where ψ denotes the EDR direction β_k or the eigenfunction ϕ_k. For the dense setting, the projections are well approximated by integrals on the observed times. For the sparse setting, we substitute with the PACE estimate $\hat{X}_i$. To assess the classification error, we generate a validation sample and compute the predicted class of Y^* based on $\langle \hat{\psi}, X^* \rangle$, where X^* is the underlying trajectory in the validation sample. We report the classification error that is minimized jointly over the tuning parameter λ in WSVM, the structural dimension K and the truncation s_n. Specifically the FPCA is implemented using the software package PACE available at http://www.stat.ucdavis.edu/PACE/, and the truncation is also chosen by minimizing the classification error.

We generate a training sample of $n = 200$ in each Monte Carlo run from the process $X_i(t) = \sum_{j=1}^{50} \xi_{ij}\phi_j(t)$ for $t \in [0, 10]$, where $\phi_j(t) = 5^{-1/2}\cos(\pi tj/5)$ for odd j and $\phi_j(t) = 5^{-1/2}\sin(\pi tj/5)$ for even j and ξ_{ij} is independently distributed as $N(0, j^{-1.5})$. In the sparse setting, the number of observations n_i is i.i.d. and uniform over $\{10, 11, \ldots, 20\}$, the observation times T_{ij} are i.i.d. from $U[0, 10]$, and the measurement error ε_{ij} is i.i.d. from $N(0, 0.1)$. In the densely observed functional data, $T_{ij} = 0.1(j-1)$ for $j = 1, \ldots, 101$. The EDR directions used are $\beta_1(t) = \sum_{j=1}^{50} b_j\phi_j(t)$ for $b_j = 1$ if $j = 1, 2$ and $b_j = (j-2)^{-3}$ for $j = 3, \ldots, 50$, and $\beta_2(t) = \sqrt{3/10}(t/5 - 1)$. We consider the following single and multiple index models:

$$\begin{aligned}
&\text{Model I:} \quad f(X) = \sin(\pi\langle\beta_1, X\rangle/4),\\
&\text{Model II:} \quad f(X) = \exp(\langle\beta_1, X\rangle/2) - 1,\\
&\text{Model III:} \quad f(X) = \sin(\pi\langle\beta_1, X\rangle/3) + \exp(\langle\beta_2, X\rangle/3) - 1,\\
&\text{Model IV:} \quad f(X) = \arctan(\pi\langle\beta_1, X\rangle) + \exp(\langle\beta_2, X\rangle/3) - 1.
\end{aligned}$$

Then, the class labels are generated by $Y = \text{sign}\{f(X) + \epsilon\}$, where the model error ϵ is i.i.d. from $N(0, 0.1)$ for all models. In the implementation of our PEFCS, we choose 100 equally distanced points between 0 and 1 as the probability candidate set, i.e., $\{\pi_k = k/100 : k = 1, \ldots, 100\}$. A validation sample of $N = 500$ is generated from the same setting, and 100 Monte Carlo runs are used to assess the expected classification error. We report the classification errors obtained from different classifiers based on projections onto the EDR space or the eigenspace for both sparse and dense settings in Table 1. We can see that all four classifiers based on the proposed PEFCS improve or are comparable to the classification results over those based on the FPCA, amongst which the centroid method based on EDR projections has the most improvements and appears to outperform other classifiers. It is also noted that the structural dimension K in all cases has been correctly identified when the average classification error is minimized.

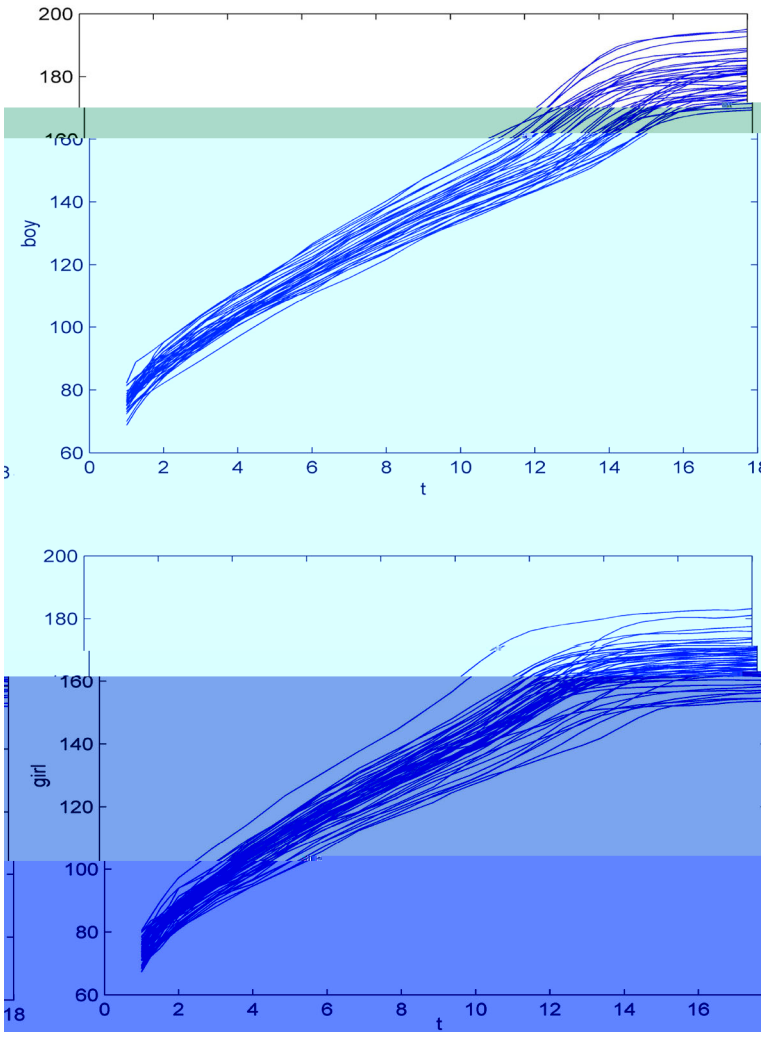
1 The average classification error with the standard error (in parenthesis) in percentage (%) obtained from 100 Monte Carlo repetitions

Model	Method	LDA	QDA	Centroid	Logistic
Sparse					
I	PEFCS	13.5 (0.18)	13.5 (0.20)	12.9 (0.19)	13.6 (0.19)
	FPCA	13.4 (0.21)	14.1 (0.25)	18.0 (0.54)	14.3 (0.24)
II	PEFCS	17.7 (0.22)	17.8 (0.23)	17.0 (0.22)	17.7 (0.23)
	FPCA	17.6 (0.23)	18.2 (0.21)	22.1 (0.52)	17.6 (0.29)
III	PEFCS	15.8 (0.30)	15.9 (0.33)	15.3 (0.24)	15.6 (0.32)
	FPCA	16.1 (0.29)	17.5 (0.30)	21.1 (0.78)	16.5 (0.50)
IV	PEFCS	8.03 (0.26)	8.59 (0.24)	7.41 (0.23)	8.81 (0.25)
	FPCA	7.92 (0.20)	8.62 (0.21)	14.5 (0.49)	8.92 (0.29)
Dense					
I	PEFCS	12.6 (0.18)	12.4 (0.18)	12.1 (0.16)	12.4 (0.17)
	FPCA	12.7 (0.19)	12.9 (0.19)	15.2 (0.26)	12.8 (0.19)
II	PEFCS	17.2 (0.20)	16.7 (0.18)	16.3 (0.18)	16.9 (0.18)
	FPCA	17.1 (0.18)	17.2 (0.17)	18.6 (0.22)	17.0 (0.17)
III	PEFCS	14.9 (0.23)	14.6 (0.22)	14.3 (0.21)	14.2 (0.22)
	FPCA	14.8 (0.20)	15.0 (0.21)	17.1 (0.22)	16.1 (0.18)
IV	PEFCS	7.26 (0.17)	7.02 (0.14)	6.46 (0.13)	6.98 (0.14)
	FPCA	7.10 (0.16)	7.31 (0.16)	11.7 (0.28)	6.83 (0.15)

5

5.1 2 2

Studies of human growth dynamics are an important topic in biological and medical applications that have profound impact for many years. This example concerns the Berkeley growth study originally published in Tuddenham and Snyder (1954) and analyzed in (Ramsay and Silverman 2005). The dataset contains 93 children's height trajectories, of which 54 are girls and 39 are boys. The height from each child was measured at quarterly from ages 1 to 2, annually from 2 to 8, and semiannually from 8 till 18, yielding 31 measurements per child. Gender serves as a natural class label. The interpolated trajectories for each gender are shown in Fig. 1. For illustration purpose, besides analyzing the original data, we also randomly sample the number of observations n_i from $\{12, \ldots, 15\}$ with the times t_{ij} randomly chosen from the original measurement times with equal probability to construct the sparsely observed data. To assess the classification error, we randomly split dataset into training and validation sets of sizes 75 and 18, respectively, and report in Table 2 the minimized average classification error computed from 20 random partitions. It is seen that the all four types of PEFCS-based classifiers consistently outperform those based on the FPCA for both the original and sparse settings, though the QDA is clearly suboptimal.



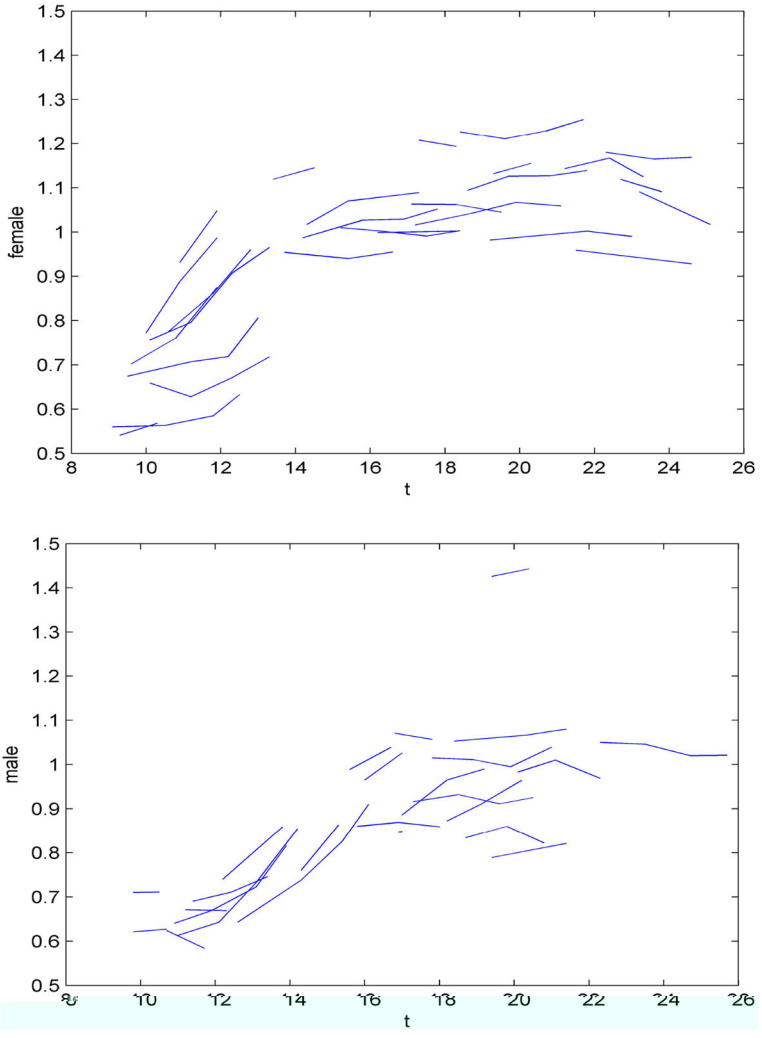
. 1 Height trajectories of 39 boys (*top*) and 54 girls (*bottom*) from the Berkeley growth data

2 The average classification error (×100 %), with its standard error in parenthesis obtained from 20 random partitions of the Berkeley growth data

Data	Method	LDA	QDA	Centroid	Logistic
Original	PEFCS	3.06 (0.75)	3.56 (0.74)	3.33 (0.85)	3.50 (0.83)
	FPCA	5.58 (1.53)	5.72 (1.29)	5.56 (1.03)	5.50 (0.75)
Sparse	PEFCS	4.33 (0.85)	4.34 (0.94)	4.39 (1.00)	4.37 (0.87)
	FPCA	8.61 (1.37)	9.83 (1.47)	9.33 (1.11)	8.72 (1.61)

5.2 [illegible]

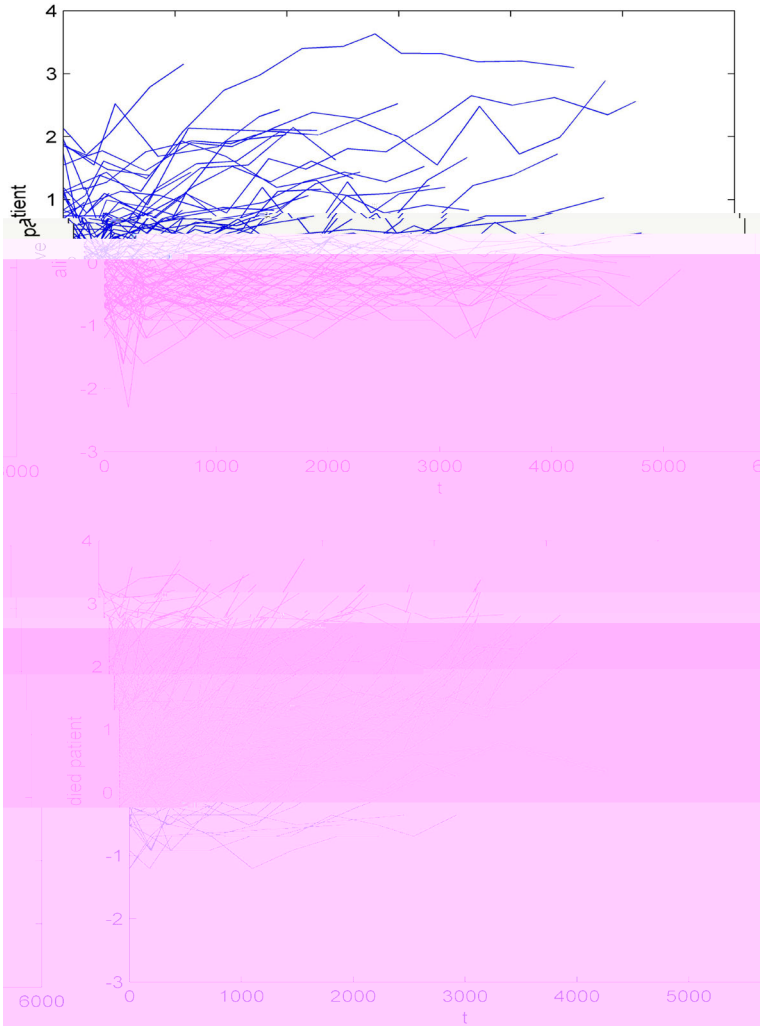
We next study the bone density data investigated by James and Hastie (2001), where only 2–4 measurements of the bone density are available at widely different times for 280 individuals. It also contains the gender and ethnicity information, such as Asian, Black, Hispanic or White, with each individual belonging to one of these ethnicity groups. To remove the confounding effect between gender and ethnicity, we consider the gender classification for 52 Hispanic individuals, of which 27 are female and 25 male. The sparsely observed data are shown in Fig. 2. We again randomly split the data into training and validation sets of sizes 42 and 10 for assessing the classification.



[illegible] 2 Spinal bone density data for Hispanic female (*top*) and male (*bottom*)

3 The average classification error (×100 %), with its standard error in parenthesis obtained from 20 random partitions of the spinal bone density data

Method	LDA	QDA	Centroid	Logistic
PEFCS	30.0 (4.10)	30.5 (4.00)	22.5 (2.80)	30.0 (4.35)
FPCA	39.5 (4.00)	38.0 (3.60)	31.5 (3.27)	38.5 (3.93)



3 Logarithm-transformed measurements of serum bilirubin for the patients that are alive (*top*) or dead (*bottom*) beyond 10 years from the primary biliary cirrhosis data

Results over 20 random partitions are reported in Table 3. From the minimized average classification error over 20 random partitions, we see that the proposed method is superior to the FPCA across all four classifiers.

4 The average classification error (×100 %), with its standard error in parenthesis obtained from 20 random partitions of the primary biliary cirrhosis follow-up data

Method	LDA	QDA	Centroid	Logistic
PEFCS	21.5 (1.31)	22.9 (1.08)	16.6 (11.8)	20.9 (1.15)
FPCA	24.6 (1.17)	24.4 (1.15)	19.9 (1.57)	23.1 (1.02)

5.3

The third example concerns the primary biliary cirrhosis (PBC) follow-up data that were also sparsely observed, see Appendix D in Fleming and Harrington (1991) for description. Different from the original PBC data, the follow-up data contain multiple measurements for 312 patients at sparse and irregular times. Also included is the survival status, of which 143 lived beyond 10 years, 140 died within 10 years and 29 is in the transplantation status. The serum bilirubin has been measured in mg/dl for each patient at different times during the first 9 years, whilst we are interested in distinguishing the death or alive status (thus excluding patients in transplantation status) based on the longitudinally measured bilirubin that are logarithm-transformed and shown in Fig. 3. Similar to previous examples, we assess the classification error using 20 random partitions, each with 227 and 56 patients in training and validation sets. The minimized classification error reported in Table 4 again demonstrates the improvement via the proposed PEFCS method across all four classifiers considered.

6

In this article, we proposed a new method combining the weighted support vector machine and functional cumulative slicing for classifying sparsely observed functional data. The probability-based slicing tackles the lack of heterogeneity for estimating the EDR space when the response in classification problem is binary. For handling the sparsely observed functions, we adopt the cumulative slicing strategy to borrow information across subjects. It is straightforward to apply commonly used classifiers to the reduced data projected onto the resultant EDR space, which has been demonstrated through extensive numerical examples to be superior to those based on FPCA method. The selection of some important parameters, including the tuning parameter in the WSVM, the structural dimension and the truncation for covariance inverse, remains an open problem for further investigation.

A :

Proof of Proposition 1 The proof is similar to that of Lemma 1 in Shin et al. (2014). We first show $S_{p(X)|X} \subseteq S_{Y|X}$, which is equivalent to show that, for any $\{\beta_k\}_{k=1}^K$ such that $X \perp p(X) \,|\, \langle\beta_1, X\rangle, \ldots, \langle\beta_K, X\rangle$, we have $X \perp Y \,|\, \langle\beta_1, X\rangle, \ldots, \langle\beta_K, X\rangle$. Recall $Y\{p(x), \varepsilon^*\}$ is 1 if $\varepsilon^* \le p(x)$ and -1 otherwise. As a consequence, $X \perp Y \,|\, p(X)$ and

$X \perp Y \mid \{p(X), \langle X, \langle \beta_1, X\rangle, \ldots, \langle \beta_K, X\rangle\}$. Since $X \perp p(X) \mid \langle \beta_1, X\rangle, \ldots, \langle \beta_K, X\rangle$, we obtain $X \perp Y \mid \langle \beta_1, X\rangle, \ldots, \langle \beta_K, X\rangle$ owing to Proposition 4.6 of Cook (1998).

To show $S_{Y|X} \subseteq S_{p(X)|X}$ is equivalent to show $Y \perp X \mid \langle \beta_1, X\rangle, \ldots, \langle \beta_K, X\rangle \Rightarrow X \perp p(X) \mid \langle \beta_1, X\rangle, \ldots, \langle \beta_K, X\rangle$ for any $\{\beta_k\}$

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory (COLT'92). ACM, New York, pp 144–152

Bredensteiner EJ, Bennett KP (1999) Multicategory classification by support vector machines. Comput Optim Appl 12:53–79

Cai TT, Hall P (2006) Prediction in functional linear regression. Ann Stat 34(5):2159–2179

Cai TT, Yuan M (2012) Minimax and adaptive prediction for functional linear regression. J Am Stat Assoc 107:1201–1216

Cardot H, Ferraty F, Mas A, Sarda P (2003a) Testing hypotheses in the functional linear model. Scand J Stat Theory Appl 30(1):241–255

Cardot H, Ferraty F, Sarda P (2003b) Spline estimators for the functional linear model. Stat Sin 13(3):571–591

Castro PE, Lawton WH, Sylvestre EA (1986) Principal modes of variation for processes with continuous sample curves. Technometrics 28:329–337

Chang CC, Chien LJ, Lee YJ (2011) A novel framework for multi-class classification via ternary smooth support vector machine. Pattern Recognit 44:1235–1244

Chen D, Hall P, Müller HG (2011) Single and multiple index functional regression models iwth nonparametric link. Annu Stat 39:1720–1747

Chiaromonte F, Cook R, Li B (2002) Sufficient dimensions reduction in regressions with categorical predictors. Ann Stat 30:475–497

Cook RD (1998) Regression graphics ideas for studying regressions through graphics. Wiley, New York

Cook RD, Weisberg S (1991) Comment on sliced inverse regression for dimension reduction.(r)-.14.100017999878(C00

Hall P, Müller HG, Wang JL (2006) Properties of principal component methods for functional and longitudinal data analysis. Ann Stat 34(3):1493–1517
He G, Müller HG, Wang JL (2003) Functional canonical analysis for square integrable stochastic processes. J Multivar Anal 85(1):54–77
He X, Wang Z, Jin C, Zheng Y, Xue X (2012) A simplified multi-class support vector machine with reduced dual optimization. Pattern Recognit Lett 33:71–82
Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer, New York
James GM (2002) Generalized linear models with functional predictors. J R Stat Soc Ser B (Stat Methodol) 64(3):411–432
James GM, Hastie TJ (2001) Functional linear discriminant analysis for irregular sampled curve. J R Stat Soc Ser B 63:533–550
James GM, Silverman BW (2005) Functional additive model estimation. J Am Stat Assoc 100:565–576
James GM, Hastie TJ, Sugar CA (2000) Principal component models for sparse functional data. Biometrika 87(3):587–602
Jank W, Shmueli G (2006) Functional data analysis in electronic commerce research. Stat Sci 21(2):155–166
Jiang CR, Yu W, Wang JL (2014) Inverse regression for longitudinal data. Ann Stat 42(2):563–591
Jones MC, Rice JA (1992) Displaying the important features of large collections of similar curves. Am Stat 46:140–145
Kimeldorf G, Wahba G (1971) Some results on Tchebycheffian spline functions. J Math Anal Appl 33:82–95
Kirkpatrick M, Heckman N (1989) A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. J Math Biol 27(4):429–450
Kneip A, Ramsay JO (2008) Combining registration and fitting for functional models. J Am Stat Assoc 103(483):1155–1165
Kneip A, Utikal KJ (2001) Inference for density families using functional principal component analysis. J Am Stat Assoc 96(454):519–542 (with comments and a rejoinder by the authors)
Lee Y, Lin Y, Wahba G (2004) Multicategory support vector machines theory and application to the classification of microarray data and satellite radiance data. J Am Stat Assoc 99:67–81
Lei E, Yao F, Heckman N, Meyer K (2014) Functional data model for genetically related individuals with application to cow growth. J Comput Graph Stat. doi:10.1080/10618600.2014.948180
Leng X, Müller HG (2006) Classification using functional data analysis for temporal gene expression data. Bioinformatics 22:68–76
Li B, Wang S (2007) On directional regression for dimension reduction. J Am Stat Assoc 102:997–1008
Li KC (1991) Sliced inverse regression for dimension reduction. J Am Stat Assoc 86:316–342
Li KC (1992) On principal hessian directions for data visualization and dimension reduction another application of steins lemma. J Am Stat Assoc 87:1025–1039
Li Y, Hsing T (2010) Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. Annu Stat 38:3028–3062
Lin X, Carroll RJ (2000) Nonparametric function estimation for cluster data when the predictor is measured without/with error. J Am Stat Assoc 95:520–534
Lin Y, Lee Y, Wahba G (2004) Support vector machines for classification in nonstandard situations. Mach Learn 33:191–202
Liu Y, Shen X (2006) Multicategory ψ-learning. J Am Stat Assoc 101:500–509
Liu Y, Yuan M (2011) Reinforced multicategory support vector machines. J Comput Graph Stat 20:901–919
Morris JS, Carroll RJ (2006) Wavelet-based functional mixed models. J R Stat Soc Ser B (Statl Methodol) 68(2):179–199
Morris JS, Vannucci M, Brown PJ, Carroll RJ (2003) Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. J Am Stat Assoc 98(463):573–597 (with comments and a rejoinder by the authors)
Müller HG (2005) Functional modelling and classification of longitudinal data. Scand J Stat Theory Appl 32:223–240
Müller HG (2008) Functional modeling of longitudinal data. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G (eds) Longitudinal data analysis (handbooks of modern statistical methods). Chapman & Hall/CRC, New York, pp 223–252
Müller HG, Stadtmüller U (2005) Generalized functional linear models. Ann Stat 33(2):774–805
Müller HG, Chiou JM, Leng X (2008) Inferring gene expression dynamics via functional regression analysis. BMC Bioinform 9:60

Ramsay J, Silverman B (2002) Applied functional data analysis. Springer series in statistics. Springer, New York
Ramsay JO, Li X (1998) Curve registration. J R Stat Soc Ser B (Stat Methodol) 60(2):351–363
Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York
Ramsay JO, Hooker G, Campbell D, Cao J (2007) Parameter estimation for differential equations: a generalized smoothing approach (with discussion). J R Stat Soc Ser B (Stat Methodol) 69(5):741–796
Rao CR (1958) Some statistical methods for comparison of growth curves. Biometrics 14(1):1–17
Rice AJ, Silverman BW (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. J R Stat Soc Ser B 53:233–243
Rice JA (2004) Functional and longitudinal data analysis: perspectives on smoothing. Stat Sin 14:631–647
Rice JA, Wu CO (2001) Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57(1):253–259
Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–407
Rosenblatt F (1962) Principles of neurodynamics. Spartan, New York
Shi M, Weiss RE, Taylor JMG (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. Ann Stat 45:151–163
Shin SJ, Wu Y, Zhang HH, Liu Y (2014) Probability-enhanced sufficient dimension reduction for binary classification. Biometrics 70:546–555
Silverman BW (1996) Smoothed functional principal components analysis by choice of norm. Ann Stat 24(1):1–24
Tuddenham R, Snyder M (1954) Physical growth of California boys and girls from birth to age 18. Univ Calif Publ Child Dev 1:183–364
Vapnik V (1998) Statistical learning theory. Wiley, New York
Vapnik V, Lerner A (1963) Pattern recognition using generalized portrait method. Autom Remote Control 24:774–780
Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
Wahba G (1990) Spline models for observational data. In: CBMS-NSF regional conference series in applied mathematics, vol 35. SIAM, Philadelphia
Wang J, Shen X, Liu Y (2008) Probability estimation for large-margin classifier. Biometrika 95:149–167
Wang L, Shen X (2006) Multicategory support vector machines, feature selection and solution path. Stat Sin 16:617–634
Wang L, Shen X (2007) On l_1-norm multiclass support vector machines: methodology and theory. J Am Stat Assoc 102:583–594
Weston J, Watkins C (1999) Support vector machines for multiclass pattern recognition. In: European symposium on artificial neural networks, pp 219–224
Wu Y, Liu Y (2007) Robust truncated-hinge-loss support vector machines. J Am Stat Assoc 102:974–983
Wu Y, Liu Y (2013) Functional robust support vector machines for sparse and irregular longitudinal data. J Comput Graph Stat 2:379–395
Xia Y, Tong H, Li W, Zhu LX (2002) An adaptive estimation of dimension reduction space. J R Stat Soc Ser B (Stat Methodol) 64(3):363–410
Yao F, Lee TCM (2006) Penalized spline models for functional principal component analysis. J R Stat Soc Ser B (Stat Methodol) 68(1):3–25
Yao F, Müller HG, Wang JL (2005a) Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100(470):577–590
Yao F, Müller HG, Wang JL (2005b) Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100:577–590
Yao F, Lei E, Wu Y (2015) Effective dimensional reduction for sparse functional data. Biometrika. doi:10.1093/biomet/asv006
Yuan M, Cai TT (2010) A reproducing kernel Hilbert space approach to functional linear regression. Ann Stat 38(6):3412–3444
Zhao X, Marron JS, Wells MT (2004) The functional data analysis view of longitudinal data. Stat Sin 14(3):789–808
Zhou L, Huang JZ, Martinez JW, Maity A, Baladandayuthapani V, Carroll RJ (2010) Reduced rank mixed effects models for spatially correlated hierarchical functional data. J Am Stat Assoc 105:390–400
Zhu L, Zhu L, Feng Z (2010) Dimension reduction in regressions through cumulative slicing estimation. J Am Stat Assoc 105:1455–1466