Published in final edited form as:
J Comput Graph Stat. 2013 April 1; 22(2): 379–395. doi:10.1080/10618600.2012.680823.

Functional robust support vector machines for sparse and irregular longitudinal data

Yichao Wu [Assistant Professor] and
Department of Statistics, North Carolina State University, Raleigh, NC 27695 (wu@stat.ncsu.edu)

Yufeng Liu [Associate Professor]
Department of Statistics and Operations Research, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599 (yfliu@email.unc.edu)

Abstract

Functional and longitudinal data are becoming more and more common in practice. This paper focuses on sparse and irregular longitudinal data with a multicategory response. The predictor consists of sparse and irregular observations, potentially contaminated with measurement errors, on the predictor trajectory. To deal with this type of complicated predictors, we borrow the strength of large margin classifiers in statistical learning for classification of sparse and irregular longitudinal data. In particular, we propose functional robust truncated-hinge-loss support vector machines to perform multicategory classification with the aid of functional principal component analysis.

Key Words and Phrases

Classification; functional principal component analysis; longitudinal data; multicategory; reproducing kernel Hilbert space; sparse and irregular; SVM; truncated-hinge-loss SVM

1 Introduction

Recent technology advance has enriched us with lots of data involving a functional predictor, which refers to a smooth random trajectory. The longitudinal study is a typical example, in which data consist of repeated measurements made on each individual patient or, more generally, subject. These measurements are typically sparse and made at irregular time points over the time horizon of the study. In addition these sparse and irregular measurements are possibly contaminated with measurement errors. Interested readers may consult Diggle, Heagerty, Liang and Zeger (2002) for an introduction to longitudinal data analysis. Through out this paper, we call this type of data sparse and irregular functional data or, interchangeably, sparse and irregular longitudinal data. Instead of treating these observations as a vector, we think these observations coming from a smooth trajectory over the time horizon of the study. In this way, theoretically we are handling statistical problems with a functional predictor instead of a vector predictor.

For regression of a continuous response on such sparse and irregular functional data, Yao, Müller and Wang (2005b) studied the functional linear regression by using the principal

Correspondence to: Yufeng Liu.

Supplementary materials
Computer Code: The Matlab code used in Section 6 is contained in the zip file FRSVM.zip available online. Please refer to the readme file for a description on the code.

component analysis through conditional expectation (PACE, Yao, Müller and Wang, 2005a). Along this direction, there has been a number of extensions to deal with more complicated models such as functional additive models (Müller and Yao, 2008) and functional quadratic regression (Yao and Müller, 2010). See references therein for other extensions.

Different from regression, classification deals with a categorical response. Its goal is to estimate a classification rule, which will be used to predict the categorical response. In the traditional setting with a multivariate predictor, many methods have been proposed for classification. They include, but are not limited to, logistic regression, linear/quadratic discriminant analysis, classification trees, random forests, support vector machines (SVMs), and boosting. Interested readers may consult Hastie, Tibshirani and Friedman (2009); Cristianini and Shawe-Taylor (2000). Albeit so many methods in the traditional setting with a multivariate predictor, most of these methods cannot be readily applied to functional data directly and very limited techniques are available to deal with a sparse and irregular functional predictor. James and Hastie (2001) extended linear discriminant analysis to deal with irregularly sampled curves. Leng and Müller (2006) proposed an extension of binary logistic regression to deal with sparse and irregular functional data and more generally Müller and Stadtmüller (2005) considered functional generalized linear models. When the predictor trajectory is either fully observable or sampled on a fine grid, Li and Yu (2008) proposed functional segment discriminant analysis and Lee (2004) and Rossi and Villa (2006) extended the SVM. Hall et al. (2001) treated signals as smooth curves and proposed a functional data-analytic approach for signal discrimination. See references therein for other relevant extensions.

Despite the aforementioned development in classification of functional data, more new techniques are needed for sparse and irregular functional data. In this paper, we consider the general problem of multicategory classification with a sparse and irregular functional predictor and the response being the discrete class membership of multiple classes. In particular, we borrow the strength of large margin classifiers in multivariate data for functional data. In the machine learning literature, the binary SVM has been very popular and enjoyed great success in a wide range of application areas (Cristianini and Shawe-Taylor, 2000). Partially due to its great success, the binary SVM has been generalized in several different ways to handle multicategory classification problems. Its multicategory extensions include Weston and Watkins (1999); Bredensteiner and Bennett (1999); Crammer and Singer (2001); Lee et al. (2004); Liu and Shen (2006); Liu and Yuan (2010), and many others. In this paper, we extend the robust truncated-hinge-loss SVM (RSVM) due to its robustness to outliers and nice interpretation in terms of support vectors as demonstrated in Wu and Liu (2007). The proposed functional RSVM can maintain robustness as the regular RSVM does. In particular, when we have mislabeled data such as a functional curve in one class mislabeled as in another class, the RSVM can reduce the impact of those outliers. Using the similar idea, parallel extensions to other multicategory classification methods can be made.

The remaining of the paper is organized as follows. Section 2 gives a brief review on large margin classifiers with multivariate predictors. Section 3 presents the functional linear RSVM, the extension of the linear RSVM. The corresponding nonlinear extension, functional nonlinear RSVM, is given in Section 4. Several implementation issues are discussed in Section 5. Simulation studies and two real application examples are presented in Sections 6 and 7, respectively. We conclude the paper with some discussion in Section 8.

2 Background on large margin classifiers with multivariate predictors

In this section, we briefly review large margin classifiers in the standard setting with multivariate predictors. Their extensions to the case with longitudinal data are presented in Sections 3 and 4. Consider a K-class classification problem. Let $\{(\mathbf{x}_i, y_i); i = 1, \ldots, n\}$ denote a training dataset. The n pairs of observations $(\mathbf{x}_i, y_i)$'s are assumed to be independent realizations of a random pair $(\mathbf{X}, Y)$, which has an unknown probability distribution $P(\mathbf{x}, y)$. Here $\mathbf{x} \in S \subset \mathbb{R}^d$ denotes an input vector and $y \in \{1, \ldots, K\}$ represents an output (class) variable. We use $\mathbf{X}$ and Y to denote random variables and $\mathbf{x}$ and y to represent corresponding observations.

Define $\mathbf{f} = (f_1, \ldots, f_K)$, each f_j being a mapping from S to $\mathbb{R}$, as a decision function vector. These K functions represent K different classes with f_j corresponding to class j; $j = 1, \ldots, K$. Once $\mathbf{f}$ is obtained from the training dataset, a classifier $\hat{y} = \operatorname{argmax}_{j=1,\ldots,K} f_j(\mathbf{x})$ is employed to predict the class of any input vector $\mathbf{x} \in S$. In other words, $f_{\hat{y}}(\mathbf{x})$ is the maximum among K values of $\mathbf{f}(\mathbf{x})$. One important goal of multicategory classification is to find a classifier which minimizes the probability of misclassifying a new input vector $\mathbf{X}$, namely the generalization error (GE), $\mathrm{Err}(\mathbf{f}) = P[Y \neq \operatorname{argmax}_j f_j(\mathbf{X})]$. Denote the multiple comparison vector of class y versus the rest as $\mathbf{g}(\mathbf{f}(\mathbf{x}), y) = (f_y(\mathbf{x}) - f_1(\mathbf{x}), \ldots, f_y(\mathbf{x}) - f_{y-1}(\mathbf{x}), f_y(\mathbf{x}) - f_{y+1}(\mathbf{x}), \ldots, f_y(\mathbf{x}) - f_K(\mathbf{x}))$. Then $\mathbf{f}$ produces correct classification for $(\mathbf{x}, y)$ if $\min(\mathbf{g}(\mathbf{f}(\mathbf{x}), y)) > 0$. Using the notation of generalized functional margin $\min(\mathbf{g}(\mathbf{f}(\mathbf{x}), y))$, we can rewrite the classification error rate on the training dataset as $(1/n) \sum_{i=1}^{n} I(\min(\mathbf{g}(f(\mathbf{x}_i), y_i)) \leq 0)$, where $I(\cdot)$ is an indicator function (Liu and Shen, 2006).

After replacing the indictor function, also known as the 0–1 loss, by a surrogate loss function, a large margin classifier solves the following minimization problem:

$$\min_{\mathbf{f}} \left(\lambda \sum_{j=1}^{K} J(f_j) + \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{f}(\mathbf{x}_i), y_i) \right) \text{ subject to } \sum_{j=1}^{K} f_j(\mathbf{x}) = 0 \forall \mathbf{x} \in S, \quad (1)$$

where $\ell(u)$ is a large margin loss function. The first term $\sum_{j=1}^{K} J(f_j)$ in the objective function in (1) can be viewed as a roughness penalty of $\mathbf{f}$. More information on large margin classifiers can be found in Shen et al. (2003); Lin (2004); Bartlett et al. (2006); Liu and Shen (2006); Liu (2007). As we will discuss in Sections 3 and 4, the key for our extension is to convert the sparse and irregular functional predictor into multivariate predictors and then utilize the large margin classifiers. We will focus on the RSVM by Wu and Liu (2007) using the truncated hinge loss on the generalized functional margin $\min(\mathbf{g}(\mathbf{f}(\mathbf{x}), y))$.

3 Functional linear robust SVM

In this paper we focus on classification problems with a functional predictor. We assume that the predictor process $X(t)$, defined on a finite domain $\mathcal{T}$ is square integrable, namely $X(\cdot) \in L_2(\mathcal{T}$, where $L_2(\mathcal{T}$ denotes all square integrable functions defined on the domain $\mathcal{T}$ As discussed in Section 2, the categorical response is denoted by $Y \in \{1, 2, \cdots, K\}$, where K denotes the number of classes. Our goal is to estimate functionals $f_1(\cdot), f_2(\cdot), \cdots, f_K(\cdot)$ and use the argmax rule $\operatorname{argmax}_k f_k(X(\cdot))$ to make future class prediction.

3.1 Karhunen-Loève expansion

Define $\mu_X(t) = EX(t)$ and $G(s, t) = \operatorname{cov}(X(s), X(t))$ for $s, t \in \mathcal{T}$ as the mean and covariance functions, respectively. We use $X^c(t) = X(t) - \mu_X(t)$ to denote the centered predictor process. Denote λ_m and $\phi_m(\cdot)$; $m = 1, 2, \cdots$, to be the eigenvalues and eigenfunctions of the

autocovariance operator of X, where eigenvalues are sorted in a non-increasing order satisfying $\lambda_1 \geq \lambda_2 \geq \cdots$. Then the covariance function $G(s, t)$ can be represented as $G(s,t)=\sum_{m=1}^{\infty}\lambda_m\phi_m(s)\phi_m(t)$. In addition, with the aid of these eigenfunctions, the predictor process can be represented by the Karhunen-Loève representation

$$X(t)=\mu_X(t)+\sum_{m=1}^{\infty}\xi_m\phi_m(t) \quad (2)$$

for $t \in \mathcal{T}$ Here the functional principal component (FPC) scores are given by $\xi_m = \int_{\mathcal{T}} X^c(t)\phi_m(t)dt$; $m = 1, 2, \cdots$, and they are uncorrelated with $E\xi_m = 0$ and $\text{var}(\xi_m) = \lambda_m$.

3.2 Functional linear classification model

As aforementioned, we are interested in estimating functionals $f_k : L_2(\mathcal{T} \to \mathbb{R}; k = 1, 2, \cdots, K$, and they will be plugged into the argmax rule to make future class prediction. Since we are focusing on functional linear classification in this section, we take the motivation of functional linear regression and assume the functional $f_k(\cdot)$ in our functional classification to take the form of $b_k+\int_{\mathcal{T}} \beta_k(t)X^c(t)dt$, where $b_k \in \mathbb{R}$ and $\beta_k(\cdot) \in L_2(\mathcal{T}$ denote the unknown parameters, $k = 1, 2, \cdots, K$. Our goal of functional linear classification is to estimate $b_1, b_2, \cdots, b_K$ and $\beta_1(\cdot), \beta_2(\cdot), \cdots, \beta_K(\cdot)$.

As we assume that $\beta_k(\cdot)$ is square integrable, we can expand the parameter function $\beta_k(t)$ in terms of the eigenfunctions $\phi_m(\cdot)$ similarly as in the Karhunen-Loève expansion, namely assume $\beta_k(t)=\sum_{m=1}^{\infty}\beta_{mk}\phi_m(t)$, $k = 1, 2, \cdots, K$. Then the estimation of $\beta_k(\cdot)$ is converted to the estimation of β_{mk}, $m = 1, 2, \cdots, \infty$, $k = 1, 2, \cdots, K$. Note that $\beta_{mk}; m = 1, 2, \cdots, \infty$, is an infinite sequence and thus it is difficult, if possible at all, to estimate all of them. To solve this difficulty, we note that $f_k(X(\cdot))=b_k+\int_{\mathcal{T}}\beta_k(t)X^c(t)dt=b_k+\sum_{m=1}^{\infty}\xi_m\beta_{mk}$. The tail term $\sum_{m=M+1}^{\infty}\xi_m\beta_{mk}$ in this summation does not play a very important role when M is large enough by noting that $\text{var}(\xi_m) = \lambda_m$ is a decreasing sequence. Consequently the regularization-via-truncation technique of Yao, Müller and Wang (2005b) can be borrowed here.

3.3 Estimation

As aforementioned in the introduction, instead of observing the whole predictor trajectory $X(\cdot)$, typically we only have sparse and irregular randomly-spaced repeated measurements of the predictor trajectory in longitudinal studies. These repeated measurements are most likely contaminated with additional random errors. In rare cases, we may have dense and regularly-spaced repeated measurements. While presenting our proposed estimation scheme, we focus on sparse and irregular data. However the proposed method can be applied to dense and regular longitudinal data as well.

For the ith subject with trajectory $X_i(\cdot)$, a random number N_i of repeated observations are made on $X_i(\cdot)$ at irregular and random time-points T_{ij}, $j = 1, 2, \cdots, N_i$. The repeated measurements are denoted by $U_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$; $j = 1, 2, \cdots, N_i$, where ε_{ij} is assume to be *i.i.d.* with mean zero and variance σ^2. We assume that $X_i(\cdot)$s are *i.i.d.* copies of $X(\cdot)$ with mean $\mu_X(\cdot)$ and covariance $G(\cdot, \cdot)$. Using the Karhunen-Loève representation, we have

$$U_{ij}=\mu_X(T_{ij})+\sum_{m=1}^{\infty}\xi_{im}\phi_m(T_{ij})+\varepsilon_{ij}, T_{ij} \in \mathcal{T}; j=1,2,\cdots,N_i, \quad (3)$$

where ξ_{im}, $m = 1, 2, \cdots$, are the FPC scores for trajectory $X_i(\cdot)$. Then our data set is denoted by $\{y_i, (T_{ij}, U_{ij}), j = 1, 2, \cdots, N_i; i = 1, 2, \cdots, n\}$, where $y_i \in \{1, 2, \cdots, K\}$ denotes the class membership of the ith sample.

The first step of our estimation scheme for the functional linear classification is to apply the principal component analysis through conditional expectation (PACE, Yao, Müller and Wang, 2005a) technique to obtain estimates of FPC scores ξ_{im}. The PACE first provides estimates $\hat{\mu}_X(\cdot)$ and $\hat{G}(s, t)$ of the mean and covariance functions $\mu_X(\cdot)$ and $G(\cdot, \cdot)$, respectively, using smoothing based on data $\{(T_{ij}, U_{ij}), j = 1, 2, \cdots, N_i; i = 1, 2, \cdots, n\}$. Then a functional principal component analysis is applied to $\hat{G}(\cdot, \cdot)$ to obtain estimates $\hat{\lambda}_m$ and $\hat{\phi}_m(\cdot)$ of eigenvalues λ_m and eigenfunctions $\phi_m(\cdot)$. Finally, the PACE gives estimate $\hat{\xi}_{im}$ of the FPC scores ξ_{im} by treating $U_{ij}, j = 1, 2, \cdots, N_i$ as joint-normally distributed from the trajectory $X_i(\cdot)$ and using conditional expectation. To save space, we skip all the details. Interested readers may read Yao et al. (2005a) for a detailed exposition.

As argued at the end of Section 3.2, it is difficult and also not necessary to estimate the infinite sequences β_{mk}, $m = 1, 2, \cdots, \infty$, $k = 1, 2, \cdots, K$. So we borrow the regularization-via-truncation technique of Yao et al. (2005b) while estimating functionals $\beta_k(\cdot)$, $k = 1, 2, \cdots, K$. More explicitly, we apply regularization by truncating the infinite sequence $\hat{\xi}_{im}$ to $\hat{\xi}_{i1}$, $\hat{\xi}_{i2}$, $\cdots$, $\hat{\xi}_{iM}$ for some large M such that the first M eigenvalues $\hat{\lambda}_m$, $m = 1, 2, \cdots, M$ contribute a big proportion to the sum of all eigenvalues. Denote

$$\hat{X}_i(t) = \hat{\mu}_X(t) + \sum_{m=1}^{M} \hat{\xi}_{im} \hat{\phi}_m(t), \quad (4)$$

which then estimates $X_i(t)$ based on sparse and irregular observations $\{(T_{ij}, U_{ij}), j = 1, 2, \cdots, N_i\}$, for $i = 1, 2, \cdots, n$. We can proceed using $\hat{X}_i(\cdot)$ as the predictor for the ith observation. Yet we want to point out that it is equivalent to use $(\hat{\xi}_{i1}, \cdots, \hat{\xi}_{iM})^T$ as the corresponding predictor.

To estimate b_k and β_{mk}, we treat $\hat{\xi}_{i1}, \hat{\xi}_{i2}, \cdots, \hat{\xi}_{iM}$ as observations and use the linear RSVM. Denote $\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i1}, \hat{\xi}_{i2}, \cdots, \hat{\xi}_{iM})^T$, $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \cdots, \beta_{Mk})^T$, and $h_k(\hat{\xi}_i) = b_0 + \hat{\xi}_i^T \beta_k$. Denote $\mathbf{h}(\hat{\boldsymbol{\xi}}_i) = (h_1(\hat{\boldsymbol{\xi}}_i), h_2(\hat{\boldsymbol{\xi}}_i), \cdots, h_K(\hat{\boldsymbol{\xi}}_i))^T$. There are several different types of multicategory extensions of binary large-margin classification methods. In this paper, we will adopt the extension studied by Liu and Shen (2006) and define comparison vector $\mathbf{g}(\mathbf{h}(\hat{\boldsymbol{\xi}}_i), y_i) = (h_1(\hat{\boldsymbol{\xi}}_i) - h_{y_i}(\hat{\boldsymbol{\xi}}_i), \cdots, h_{y_i-1}(\hat{\boldsymbol{\xi}}_i) - h_{y_i}(\hat{\boldsymbol{\xi}}_i), h_{y_i+1}(\hat{\boldsymbol{\xi}}_i) - h_{y_i}(\hat{\boldsymbol{\xi}}_i), \cdots, h_K(\hat{\boldsymbol{\xi}}_i) - h_{y_i}(\hat{\boldsymbol{\xi}}_i))^T$ accordingly. The ith sample is misclassified if $\min \mathbf{g}(\mathbf{h}(\hat{\boldsymbol{\xi}}_i), y_i) \le 0$ as we are using the argmax rule. By replacing the 0–1 loss with the truncated hinge loss $H_{T_s}(\cdot)$, we can estimate β_{mk} by solving

$$\min_{\beta_1, \beta_2, \cdots, \beta_K} \sum_{i=1}^{n} H_{T_s}(\min \mathbf{g}(\mathbf{h}(\hat{\xi}_i), y_i)) + \lambda \sum_{k=1}^{K} \|\beta_k\|^2 \quad (5)$$

$$\text{subject to } \sum_{k=1}^{K} \beta_{mk} = 0; m = 1, 2, \cdots, M, \sum_{k=1}^{K} b_k = 0,$$

where $\|\beta_k\| = \sqrt{\sum_{m=1}^{M} \beta_{mk}^2}$ and constrains are used to ensure identifiability. Here we choose to use the truncated hinge loss $H_{T_s}(u) = \min(H_1(s), H_1(u))$ due to its nice performance as demonstrated in Wu and Liu (2007), where $H_1(u) = \max(1 - u, 0)$ is the hinge loss used in

the standard SVM. In particular, the truncated hinge loss is Fisher-consistent for multicategory classification as long as the truncation location s satisfies $s \in [-\frac{1}{K-1}, 0]$ and they recommended to use the least truncation of $s = -1/(K-1)$. Through out this paper, we will follow their recommendation and set truncation location $s = -1/(K-1)$.

As discussed above, truncating the hinge loss helps to achieve Fisher-consistency. However this truncation leads (5) to be a non-convex optimization problem, which is challenging to solve. Note that the truncated hinge loss can be decomposed as the difference of two convex functions, namely $H_{T_s}(u) = H_1(u) - H_s(u)$ as shown in Figure 1, where $H_s(u) = \max(s - u, 0)$. Based on this decomposition, Wu and Liu (2007) proposed to use the difference convex algorithm (DCA) (An and Tao, 1997) to optimize (5). See Wu and Liu (2007) for more details on the corresponding algorithmic development. Denote the optimizer of (5) by $\hat{b}_k$ and $\hat{\beta}_{mk}$, $m = 1, 2, \cdots, M$, $k = 1, 2, \cdots, K$. Then our estimated parameter function is given by $\hat{\beta}_k(t) = \sum_{m=1}^{M} \hat{\beta}_{mk}\hat{\phi}_m(t)$. Correspondingly the estimated functional is given by $\hat{f}_k(X(\cdot)) = \hat{b}_k + \int_{\mathcal{T}} (X(t) - \hat{\mu}_X(t))\, \hat{\beta}_k(t)dt$ and class membership can be predicted by $\operatorname{argmax}_k \hat{f}_k(X(\cdot))$ for any $X(\cdot) \in L(\mathcal{T})$.

3.4 Class prediction

The above estimated functional linear classification rule is defined for the case that the whole predictor trajectory $X(\cdot)$ is fully observed. However our focus of the current paper is on sparse and irregular longitudinal data. Thus it is necessary to discuss the prediction for this case as well. Suppose instead of observing the whole trajectory $X(\cdot)$ while making prediction, we make sparse and irregular measurements U_j, potentially contaminated with measurement errors, on $X(\cdot)$ at T_j for $j = 1, 2, \cdots, N$ and a random number N. More explicitly $U_j = X(T_j) + \varepsilon_j$, for $j = 1, 2, \cdots, N$. We can use the PACE to estimate FPC scores ξ_m of $X(\cdot)$ based on estimates $\hat{\mu}_X(\cdot)$, $\hat{\lambda}_m$, and $\hat{\phi}_m(\cdot)$. Denote the estimated FPC scores by $\hat{\xi}_m$ for $m = 1, 2, \cdots, M$ and the estimated trajectory by $\hat{X}(t) = \hat{\mu}_X(t) + \sum_{m=1}^{M} \hat{\xi}_m \hat{\phi}_m(t)$. Then the corresponding class membership prediction is given by $\operatorname{argmax}_k \hat{f}_k(\hat{X}(\cdot))$.

4 Functional nonlinear RSVM

In the previous section, we focus on functional linear classification, in which we assume the functional $f_k(\cdot)$ to be parametric. In general this parametric assumption may be restrictive in some situations and functional nonparametric classification can be consequently desirable. For classification with a multivariate predictor, there are many existing techniques for nonparametric classification such as basis expansion. They may be extended to deal with classification with a functional predictor. In the literature, there are some previous applications of kernel methods for functional data analysis. For example, Canu et al. (2002) studied some general properties of kernel learning, Preda (2007) applied kernel learning for functional data with a binary response using logistic regression. In this section, we choose the approach of using reproducing kernel Hilbert space (RKHS) due to its great flexibility and propose RKHS-based functional nonlinear RSVM.

In functional nonlinear classification, we do not make any parametric assumption on the functionals $f_k(\cdot)$, $k = 1, 2, \cdots, K$. We assume for the moment that our data are given by the complete trajectory $X_i(t)$ and the categorical response y_i, $i = 1, 2, \cdots, n$. The discussion on how to deal with sparse and irregular functional data when trajectories $X_i(\cdot)$ are not fully observable will be presented shortly. Denote $\mathbf{f}(X(\cdot)) = (f_1(X(\cdot)), f_2(X(\cdot)), \cdots, f_K(X(\cdot)))^T$.

Using the RSVM, we then need to solve

$$\min_{f_1(\cdot),f_2(\cdot),\cdots,f_K(\cdot)\in\mathscr{F}} \sum_{i=1}^{n} H_{T_s}(\min \mathbf{g}(\mathbf{f}(X_i(\cdot)), y_i)) + \lambda \sum_{k=1}^{K} J(f_k(\cdot)), \quad (6)$$

$$\text{subject to } \sum_{k=1}^{K} f_k(\cdot) = 0,$$

where $J(f_k(\cdot))$ denotes some roughness penalty of the functional $f_k(\cdot)$ and $\mathscr{F}$ denotes some functional space. For a normed space $\mathscr{F}$, one example of $J(f_k(\cdot))$ is the norm of $f_k(\cdot)$. More details on the estimation are given in Section 4.1. Here a similar sum-to-zero constraint is included to ensure identifiability.

4.1 Estimation

Let $\mathscr{K}(\cdot, \cdot)$ be a bi-variate kernel function which maps $L_2(\mathscr{T} \times L_2(\mathscr{T}$ to $\mathbb{R}$. One example is the Gaussian kernel $\mathscr{K}(X_i(\cdot), X_j(\cdot)) = \exp(-\| X_i(\cdot) - X_j(\cdot) \|_2^2/\rho^2)$ with $\| X_i(\cdot) - X_j(\cdot) \|_2 = \sqrt{\int_{t\in\mathscr{T}} (X_i(t) - X_j(t))^2 dt}$. Let $\mathscr{F}_{\mathscr{K}}$ be the reproducing kernel Hilbert space generated by kernel $\mathscr{K}(\cdot, \cdot)$ and the corresponding norm by $\|\cdot\|_{\mathscr{F}_{\mathscr{K}}}$. According to the representer theorem (Kimeldorf and Wahba, 1971; Wahba, 1990), the solution of (6) with $\mathscr{F} = \mathscr{F}_{\mathscr{K}}$ and $J(\cdot) = \|\cdot\|_{\mathscr{F}_{\mathscr{K}}}$ takes the form $f_k(X(\cdot)) = c_{0k} + \sum_{i=1}^{n} c_{ik}\mathscr{K}(X(\cdot), X_i(\cdot))$ and the corresponding roughness penalty is given by $J(f_k(\cdot)) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ik}\mathscr{K}(X_i(\cdot), X_j(\cdot))c_{jk}$. Next we may plug $f_k(X(\cdot)) = c_{0k} + \sum_{i=1}^{n} c_{ik}\mathscr{K}(X(\cdot), X_i(\cdot))$ and $J(f_k(\cdot)) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ik}\mathscr{K}(X_i(\cdot), X_j(\cdot))c_{jk}$ into (6) and solve it to get optimizers $\check{c}_{ik}$, $i = 0, 1, \cdots, n$, $k = 1, 2, \cdots, K$. While doing so, the sum-to-zero constraint in (6) can be replaced by $\sum_{k=1}^{K} c_{ik} = 0$ for $i = 0, 1, \cdots, n$. The estimated RKHS-based functional nonlinear classification rule is given by $\operatorname{argmax}_k \check{f}_k(X(\cdot))$, where $\check{f}_k(X(\cdot)) = \check{c}_{0k} + \sum_{i=1}^{n} \check{c}_{ik}\mathscr{K}(X(\cdot), X_i(\cdot))$.

The above presentation of the RKHS-based functional nonlinear RSVM relies on the assumption that the functional predictor $X(\cdot)$ is fully observed. Recall for sparse and irregular functional data, we do not observe the complete trajectory $X_i(\cdot)$ and instead only have sparse and irregular observations (T_{ij}, U_{ij}), $j = 1, 2, \cdots, N_i$. To deal with a sparse and irregular functional predictor, we first apply the PACE as in the functional linear classification and define a natural estimate $\hat{X}_i(\cdot)$ for $X_i(\cdot)$ as in (4). Consequently we may use $\hat{X}_i(\cdot)$ to replace $X_i(\cdot)$ in (6). This completes our whole estimation scheme for RKHS-based functional nonlinear RSVM.

4.2 Class prediction

The corresponding prediction for a trajectory with sparse and irregular observations (T_j, U_j), $j = 1, 2, \cdots, N$ can be defined as well. As in Section 3.4, we may use the PACE to estimate FPC scores and define the estimated trajectory $\hat{X}(\cdot)$ for $X(\cdot)$ based on sparse and irregular observations (T_j, U_j), $j = 1, 2, \cdots, N$. The corresponding class membership can be predicted by $\operatorname{argmax}_k \check{f}_k(\hat{X}(\cdot))$.

The proposed RKHS-based functional nonlinear RSVM is very flexible. In general, one may choose different bi-variate kernels $\mathscr{K}(\cdot, \cdot)$ to get different classifiers. In particular if one uses

the linear kernel $\mathscr{K}X_i(\cdot), X_j(\cdot)) = \int \mathscr{K} X_i(t)X_j(t)dt$, the above RKHS-based functional nonlinear RSVM reduces to the functional linear RSVM of Section 3.

5 Implementation issues

Note that there are two layers of regularization in our whole estimation schemes for proposed functional linear or nonlinear RSVMs. The first one is to choose an M to truncate the functional principal component representation in the Karhunen-Loève expansion and the other one is controlled by the regularization parameter λ in (5) or (6). In the finite sample case, we need to select M and λ appropriately to deliver satisfactory classification performance.

For the truncation regularization, Yao et al. (2005a) proposed to apply either AIC or BIC criterion to all the data to select M. However as one referee pointed out, the AIC or BIC was aiming for regression and may be suboptimal for classification. Here we propose to tune M and λ jointly. For each $M = 1, 2, \cdots$, the PACE returns the estimated FPC scores $\hat{\xi}_{im}$ for $i = 1, 2, \cdots, n$ and $m = 1, 2, \cdots, M$. Then our proposed functional linear RSVM or RKHS-based functional nonlinear RSVM can be coupled with a D-fold cross validation to select an optimal pair of regularization parameters λ and M in (5) or (6). Alternatively we may use an independent tuning set. Here we detail the method of using cross validation for illustration. We use the misclassification error $I(y_i \neq \hat{y}_i)$ as the selection criterion where $I(\cdot)$ is the indicator function and the predicted class membership $\hat{y}_i$ is given by either $\operatorname{argmax}_k \hat{f}_k(\{(U_{ij}, T_{ij}), j = 1, 2, \cdots, N_i\})$ for the functional linear RSVM or $\operatorname{argmax}_k \check{f}_k(\{(U_{ij}, T_{ij}), j = 1, 2, \cdots, N_i\})$ for the RKHS-based functional nonlinear RSVM. In the D-fold cross validation, we randomly split the data into D folds as $\{1, 2, \cdots, n\} = F_1 \cup F_2 \cup \cdots \cup F_D$, where $F_j \cap F_d = \varnothing$ when $1 \leq d \neq j \leq D$. Denote $F_d^c = \{1, 2, \cdots, n\} \backslash F_d$ to be the complement of F_d in $\{1, 2, \cdots, n\}$. For each d, we use all data points with indices in F_d^c and train the proposed functional linear RSVM or RKHS-based functional nonlinear RSVM to obtain an estimate $\hat{f}_k^{F_d^c,\lambda,M}(\cdot)$ or $\check{f}_k^{F_d^c,\lambda,M}(\cdot)$ with the regularization parameters λ and M. Our tuning method is to use

$$\text{Error}_1(\lambda, M) = \sum_{d=1}^{D} \sum_{i \in F_d} I(y_i \neq \operatorname{argmax}_k \hat{f}_k^{F_d^c,\lambda,M}(\{(U_{ij}, T_{ij}), j = 1, 2, \cdots, N_i\}))$$

and

$$\text{Error}_2(\lambda, M) = \sum_{d=1}^{D} \sum_{i \in F_d} I(y_i \neq \operatorname{argmax}_k \check{f}_k^{F_d^c,\lambda,M}(\{(U_{ij}, T_{ij}), j = 1, 2, \cdots, N_i\})).$$

We select an optimal pair of λ and M by minimizing $\text{Error}_1(\lambda, M)$ (resp. $\text{Error}_2(\lambda, M)$) via a grid search over λ and M for the functional linear RSVM (resp. KHS-based functional nonlinear RSVM).

As a remark, in our numerical examples we use the above joint tuning. However the joint tuning is more time consuming than an alternative tuning of using BIC or AIC to select M first. According to our limited numerical experience, the BIC-based tuning works fairly well but does slightly worse than the joint tuning. Thus the BIC-based tuning can be used in practice when the computational time is a concern.

For given M and λ, functional RSVMs need to solve (5) or (6). As discussed earlier, the corresponding optimization problem is nonconvex. We can use the DC algorithm as in Wu and Liu (2007) to implement it via solving iterative convex optimization problems.

6 Monte Carlo simulation

Predictor trajectories in our simulation examples are generated as $X(t) = \mu_X(t) + \sum_{m=1}^{3} \xi_m \phi_m(t)$ for $t \in \mathcal{T} = [0, 10]$. The mean predictor trajectory is $\mu_X(t) = t + \sin(t)$. The three

eigenfunctions are $\phi_1(t) = -\sqrt{1/5}\cos(\pi t/5), \phi_2(t) = \sqrt{1/5}\sin(\pi t/5),$ and $\phi_3(t) = -\sqrt{1/5}\cos(2\pi t/5)$ and the corresponding FPC scores are independently distributed as $\xi_1 \sim N(0, 2^2)$, $\xi_2 \sim N(0, \sqrt{2}^2)$, and $\xi_3 \sim N(0, 1^2)$. Instead of observing the complete predictor trajectory, we make sparse and irregular contaminated observations. For each trajectory, the random number N of observations is uniformly generated from the discrete set {5, 6, ⋯, 10}. Given N, we generate the observation times T_j; $j = 1, 2, \cdots, N$, from the uniform distribution over $\mathcal{T} = [0, 10]$. Next these sparse and irregular data are generated by contaminating a random measurement error, namely $U_j = X(T_j) + \varepsilon_j$, where $\varepsilon_j \sim N(0, \sigma^2)$; $j = 1, 2, \cdots, N$, are independent and σ^2 will be specified later for each example.

We consider the following two examples: one with a true functional linear classification rule and the other with a true functional nonlinear classification rule. For both examples, we use the joint tuning proposed in the previous section to select λ and M. To reduce the computational load, we generate an independent tuning set, of the same size as the training set, to select the regularization parameters by minimizing the classification error over the tuning set. In order to evaluate the classification performance of each method, we use the classification error over an additional independent test set. We compare our proposed methods with the functional linear discriminant analysis of James and Hastie (2001). The average of testing errors over 100 repetitions and the corresponding standard deviations are reported for each method. As we consider more general multicategory classification and Leng and Müller (2006) can only handle a binary response, we do not compare with their method even though it can be coupled with the one-versus-the-rest technique to deal with a multicategory response.

Example 6.1 True functional linear classification rule

In this example the independent contamination error is generated from $N(0, 0.5^2)$. Conditional on the true curve $X(t)$ with ξ_1, ξ_2, and ξ_3, the categorical response is generated by $\text{argmax}_k\,[(\cos(2k\pi/3)\xi_1 + \sqrt{2}\sin(2k\pi/3)\xi_2 + \varepsilon_k]$, where $\varepsilon_k \sim N(0, 1)$, $k = 1, 2, 3$, are independent of $X(\cdot)$. In this way, the functional linear classification leads to a correct model specification. Written in the function form, given predictor $X(\cdot)$, the response is generated by

$$\underset{k}{\text{argmax}}\left[\int_0^{10}(X(t) - \mu_X(t))\beta_k(t)dt + \varepsilon_k\right],$$

where $\beta_k(t) = (\cos(2k\pi/3)\phi_1(t) + \sqrt{2}\sin(2k\pi/3)\phi_2(t)$ for $k = 1, 2, 3$. Thus the corresponding Bayes classification rule is given by $\text{argmax}_k \int_0^{10}(X(t) - \mu_X(t))\beta_k(t)dt$ since $\varepsilon_k \sim N(0, 1)$; $k = 1, 2, 3$, are independent of each other. The sample sizes of the training, tuning, and testing sets are 200, 200, and 1000, respectively.

As a remark, we note that the Bayes error with both the leading two FPCs ξ_1 and ξ_2 given is 18.08% for this simulation setting. However on each predictor curve, we only make sparse, irregular and measurement error contaminated observations which contain far less information. With sparseness, irregularity, and measurement error taken into account, we calculate the modified Bayes error using a model based method as follows. Note that each curve is represented by sparse and irregular observations $\{(T_i, U_i) : i = 1, 2, \cdots, N\}$. From the data generation setting, we know that $(\xi_1, \xi_2, \xi_3)^T$ is multivariate normal and $(U_1, U_2, \cdots, U_N)^T$ is also multivariate normal conditional on $(\xi_1, \xi_2, \xi_3)^T$ and $T_1, T_2, \cdots, T_N$. Thus we can calculate $E(\xi_j | \{(T_i, U_i) : i = 1, 2, \cdots, N\})$ for $j = 1, 2, 3$ using a joint normal model.

Plugging $E(\xi_j s|\{(T_i, U_i) : i = 1, 2, \cdots, N\})$, $j = 1, 2, 3$ into the Bayes rule, we can make a prediction denoted by $\hat{Y}(\{(T_i, U_i) : i = 1, 2, \cdots, N\})$, which is given by

$$\underset{k}{\operatorname{argmax}} \left[(\cos(2k\pi/3) E(\xi_1 | \{(T_i, U_i)_{i=1}^N\}) + \sqrt{2} \sin(2k\pi/3) E(\xi_2 | \{(T_i, U_i)_{i=1}^N\}) \right] .$$

Then the modified Bayes error is given by $E(Y \neq \hat{Y}(\{(T_i, U_i)_{i=1}^N\}))$, where the expectation is taken with respect to the randomness of sparse and irregular observations $\{(T_i, U_i) : i = 1, 2, \cdots, N\}$. As N is a random variable, it is difficult to calculate the expectation with respect to $\{(T_i, U_i) : i = 1, 2, \cdots, N\}$ directly. To solve this problem, we generate an independent sample of size 100000 and calculate the modified Bayes error using the empirical distribution of this sample. This leads to a modified Bayes error of 25.03%, which is larger than the original Bayes error of 18.08% without taking into account of sparseness, irregularity, and measurement errors, as expected.

The average test error over 100 repetitions is 27.52% with a standard deviation of 0.96% for the proposed functional linear truncated-hinge-loss SVM. The corresponding average test error over 100 repetitions using the functional LDA of James and Hastie (2001) is 29.66% with a standard deviation of 1.32%. The average test errors are quite similar to each other. This shows that the proposed functional linear RSVM performs comparably with the functional LDA when the true classification rule is functional linear. For our functional linear truncated-hinge-loss SVM, 50, 15, 14, 5, 5, 7, 2, 1, and 1 repetitions out of the total 100 repetitions select M to be 2, 3, 4, 5, 6, 7, 8, 9, and 10, respectively. Note that the classification rule only depends on ξ_1 and ξ_2. Thus M for the Bayes classification rule is 2 and it is correctly selected by 50 repetitions out of 100 in our simulation.

Example 6.2 True functional nonlinear classification rule

The second example is devoted to the case when the true classification rule is a functional nonlinear one. Conditional on the true curve $X(t)$ with ξ_1, ξ_2, and ξ_3, the categorical response is given by

$Y=1$ if $\xi_1^2+2\xi_2^2<3.2534$, $Y=2$ if $3.2534 \leq \xi_1^2+2\xi_2^2<8.8119$, and $Y=3$ if $\xi_1^2+2\xi_2^2 \geq 8.8119$. In this way, the true classification rule is not functional linear any more. While generating the sparse and irregular contaminated observations, the independent contamination error is generated from $N(0, 0.2^2)$. Our proposed RKHS-based functional nonlinear RSVM is implemented with the Gaussian kernel $\mathcal{K}(\hat{X}_i(\cdot), \hat{X}_j(\cdot)) = \exp(-\| \hat{X}_i(\cdot) - \hat{X}_j(\cdot) \|_2^2 / \rho^2)$. Borrowing the idea of Brown et al. (2000), the data width parameter ρ is selected as the median pairwise L_2 distance between classes defined as the median of $\{\| \hat{X}_i(\cdot) - \hat{X}_j(\cdot) \|_2 : y_i \neq y_j\}$, where y_i denotes the categorical response of the ith sample and $\hat{X}_i(\cdot)$ denotes the corresponding estimate for $X_i(\cdot)$ based on the sparse and irregular contaminated observations using the PACE as in (4). The sizes of the training, tuning, and testing sets are 150, 150, and 1000, respectively.

Note that the Bayes error for this simulation setting is 0% when both the leading two FPCs ξ_1 and ξ_2 are given. As noted in the previous example, with sparseness, irregularity and measurement error considered, the modified Bayes error is expected to be much higher than 0%. Using a similar calculation scheme as in the previous example, we get the modified Bayes error of 14.26% for this example. Compared with the linear example, this nonlinear example is much more challenging even though the modified Bayes error is relatively small.

Our proposed RKHS-based functional nonlinear RSVM gives an average testing error of 25.05% with a standard deviation of 2.51% among 100 repetitions. The corresponding

average testing error and standard deviation for James and Hastie (2001)'s functional linear discriminant analysis are 65.43% and 0.022%, respectively. This shows a great improvement. Yet we would like to point out that the functional linear discriminant of James and Hastie (2001) corresponds to a major model misspecification in the current example setting since their method is a functional linear method. Thus it is not fair to compare our functional nonlinear classification with theirs. The functional quadratic discriminant analysis, if available, may perform better. However to our limited knowledge, there is no existing software available for functional quadratic discriminant analysis although the idea of the corresponding extension was discussed in their paper.

As defined above, the Bayes classification rule only depends on ξ_1 and ξ_2. Its corresponding M is 2, which is correctly selected by the joint tuning of the RKHS-based functional nonlinear RSVM for 48 repetitions out of the total 100 repetitions. In the other 52 repetitions, 25, 19, 7, and 1 repetitions select M to be 3, 4, 5, and 6, respectively.

To further demonstrate how the proposed methods work, we plot the sparse and irregular predictor data in Figure 2 for one random repetition of Example 6.1. The corresponding estimated mean curve and first two eigen functions are given in the left and right panels, respectively, of Figure 3 in comparison to the corresponding true functions. Here we only plot the first two eigen functions because the true classifier only depends on the leading two.

7 Real data

In this section, we apply our new functional RSVMs to two real data sets: the spinal bone mineral density data and spectral data.

7.1 Spinal bone mineral density data

First we consider the spinal bone mineral density data studied in James and Hastie (2001). The data set consists of sparse and irregular measurements of spinal bone mineral density for 280 individuals. There are 2–4 measurements available for each individual. In addition, the ethnicity of each individual is also available and we will use it as the categorical response. Each individual belongs to one and only one group of Asian, Black, Hispanic or White in terms of ethnicity. Among these 280 individuals, 153 are females and 127 are males. We only consider the classification problem of these 153 females. The sparse and irregular measurements of these 153 females are illustrated in Figure 4.

As presented above, the first step is to apply the PACE to the sparse and irregular spinal bone mineral density measurements to estimate the mean curve and eigen functions. The estimated mean curve is given in the left panel of Figure 5. The first three estimated eigen functions are plotted in the right panel of Figure 5.

We consider both the proposed functional linear and functional nonlinear RSVMs. As there is no independent tuning data set available, we use a 5-fold cross validation to select λ and M as discussed in Section 5. Once an optimal pair of λ and M is identified, we use it in the functional linear and functional nonlinear RSVMs with data from all 153 females. We compare the number of correct classification with that of James and Hastie (2001).

The numbers of correct classification are 63 and 71 for the functional linear and the RKHS-based functional nonlinear RSVMs, respectively. The corresponding number of correct classification is 66 for the functional linear discriminant analysis (James and Hastie, 2001). These numbers show that the functional RSVMs give similar performances as the functional linear discriminant analysis. The RKHS-based functional nonlinear RSVM performs slightly

better than the functional linear discriminant analysis while the functional linear RSVM does a little bit worse for this data set.

To compare the performance of the RKHS-based functional nonlinear RSVM further with that of the functional linear discriminant analysis, we report the confusion matrix of classification for the four ethnicities in Table 1 as in the same format of Table 1 of James and Hastie (2001). A direct comparison of these two confusion matrices shows that the RKHS-based functional nonlinear RSVM gives more correctly classified samples for the Black and White ethnicity groups while less correctly classified samples for the other two. Especially for the White group, the percentage of correct classification increases from 18.8% to 35.42%.

As a remark, the joint tuning selects M to be 3 and 2 for function linear RSVM and RKHS-based functional nonlinear RSVM, respectively.

7.2 Spectral data

Next we consider the spectral data reported in Borggaard and Thodberg (1992). The predictor is the absorbance trajectory recorded on a Tecator Infratec Food and Feed Analyzer which works in the wavelength range of 850–1050 nm. The response is the fat content of meat, which is a scalar response. The original task is to predict the fat content of meat based on absorbance spectrum. The data set is available online at http://lib.stat.cmu.edu/datasets/tecator. Interested readers may find more revelent background information there.

The scalar response, the fat content, ranges between 0.9 and 49.1. To fit into the classification framework, we consider the prediction of whether the fat content is larger than 20, which leads to a binary classification problem. The total sample size is 215. As one referee pointed out that the previous example reports an in-sample performance, here we randomly split the data into training and testing sets of size 155 and 60, respectively. A 5-fold cross validation is applied to the training set to select λ and M using the joint tuning method of Section 5. The classification accuracy over the test data we set aside is reported for different methods. The functional linear discriminant analysis gives an accuracy of 52/60. The functional linear RSVM and RKHS-based functional nonlinear RSVM lead to classification accuracy of 59/60 and 58/60, respectively. The joint tuning selects the optimal M to be 4 for both methods. Improvement over the functional linear discriminant analysis is observed for this data. To save space, we skip plotting the original data and estimated mean and eigen functions.

8 Conclusion

In this work, we propose functional linear and nonlinear RSVMs. Motivated by longitudinal data, we focused on sparse and irregular functional data even though it works for densely observed functional data as well. The new methods are based on estimating functional principal component scores first using the PACE (Yao, Müller and Wang, 2005a). Once the FPC scores are estimated, functional linear and nonlinear RSVMs are simplified to the corresponding counterparts with a multivariate predictor. In this paper we choose the multicategory RSVM for demonstration. The same idea can be applied to other multicategory classification methods such as the multicategory ψ-learning (Liu and Shen, 2006) and the multicategory SVM (Liu and Yuan, 2010).

In some applications, one may also be interested in estimating the conditional probability of each curve belonging to each class in addition to predicting the class membership. For the case with a multivariate predictor, new methods (Wang et al., 2008; Wu et al., 2010) have

been recently devised to estimate the conditional class probabilities using large-margin classifiers with the aid of assigning different weights to different classes. These new methods may be extended to the case of sparse and irregular functional predictors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Professor Richard A. Levine, the associate editor, and two referees for their constructive comments and suggestions that led to significant improvement of the article. The authors also thank Professor Gareth James for sharing the spinal bone mineral density data. The work is partially supported by NSF Grants DMS-0747575 (Liu), DMS-0905561 (Wu), and DMS-1055210 (Wu), and NIH Grant NIH/NCI R01 CA-149569 (Liu and Wu).

References

An LTH, Tao PD. Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms. Journal of Global Optimization. 1997; 11:253–285.

Bartlett P, Jordan M, McAuliffe J. Convexity, classification, and risk bounds. Journal of the American Statistical Association. 2006; 101:138–156.

Borggaard C, Thodberg HH. Optimal minimal neural interpretation of spectra. Analytical Chemistry. 1992; 64:545–551.

Bredensteiner E, Bennett K. Multicategory classification by support vector machines. Computational Optimizations and Applications. 1999; 12:53–79.

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. The Proceeding of National Academy of Sciences. 2000; 97:262–267.

Canu, S.; Mary, X.; Rakotomamonjy, A. Advances in Learning Theory: Methods, Models and Applications NATO Science Series III: Computer and Systems Sciences. IOS Press; 2002. Functional learning through kernel; p. 89-110.

Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research. 2001; 2:265–292.

Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines. Cambridge University Press; 2000.

Diggle, P.; Heagerty, P.; Liang, K.; Zeger, S. Analysis of Longitudinal Data. 2nd ed.. New York: Oxford University Press; 2002.

Hall P, Poskitt DS, Presnell B. A functional data-analytic approach to signal discrimination. Technometrics. 2001; 43:1–9.

Hastie, T.; Tibshirani, R.; Friedman, JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed.. New York: Springer-Verlag; 2009.

James GM, Hastie TJ. Functional linear discriminant analysis for irregularly sampled curves. Journal of Royal Statistical Society Series B. 2001; 63:533–550.

Kimeldorf G, Wahba G. Some results on Tchebycheffian spline functions. Journal of Mathematical Analysis and Applications. 1971; 33:82–95.

Lee, H. PhD thesis. Department of Statistics, Texas, A&M University; 2004. Functional data analysis: classification and regression.

Lee Y, Lin Y, Wahba G. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association. 2004; 99:67–81.

Leng X, Müller H-G. Classification using functional data analysis for temporal gene expression data. Bioinformatics. 2006; 22:68–76. [PubMed: 16257986]

Li B, Yu Q. Classification of functional data: A segmentation approach. Computational Statistics and Data Analysis. 2008; 52:4790–4800.

Lin Y. A note on margin-based loss functions in classification. Statistics and Probability Letters. 2004; 68:73–82.

Liu Y. Fisher consistency of multicategory support vector machines. Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics. 2007:289–296.

Liu Y, Shen X. Multicategory ψ-learning. Journal of the American Statistical Association. 2006; 101:500–509.

Liu Y, Yuan M. Reinforced multicategory support vector machines. Journal of Computational and Graphical Statistics to appear. 2010

Müller H-G, Stadtmüller U. Generalized functional linear models. Annals of Statistics. 2005; 33:774–805.

Müller HG, Yao F. Functional additive models. Journal of the American Statistical Association. 2008; 103:1534–1544.

Preda C. Regression models for functional data by reproducing kernel hilbert spaces methods. Journal of Statistical Planning and Inference. 2007; 137:829–840.

Rossi F, Villa N. Support vector machine for functional data classification. Neurocomputing. 2006; 69:730–742.

Shen X, Tseng G, Zhang X, Wong W. On ψ-learning. Journal of the American Statistical Association. 2003; 98:724–734.

Wahba, G. Spline models for observational data. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics; 1990.

Wang J, Shen X, Liu Y. Probability estimation for large margin classifiers. Biometrika. 2008; 95:149–167.

Weston, J.; Watkins, C. Support vector machines for multi-class pattern recognition. In: Verleysen, M., editor. Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN-99). Belgium: Bruges; 1999. p. 219-224.

Wu Y, Liu Y. Robust truncated-hinge-loss support vector machines. Journal of the American Statistical Association. 2007; 102:974–983.

Wu Y, Zhang HH, Liu Y. Robust model-free multiclass probability estimation. Journal of the American Statistical Association. 2010; 105:424–436. [PubMed: 21113386]

Yao F, Müller H-G. Functional qudratic regression. Biometrika. 2010; 97:49–64.

Yao F, Müller H-G, Wang J-L. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association. 2005a; 100:577–590.

Yao F, Müller H-G, Wang J-L. Functional linear regression analysis for longitudinal data. The Annals of Statistics. 2005b; 33:2873–2903.

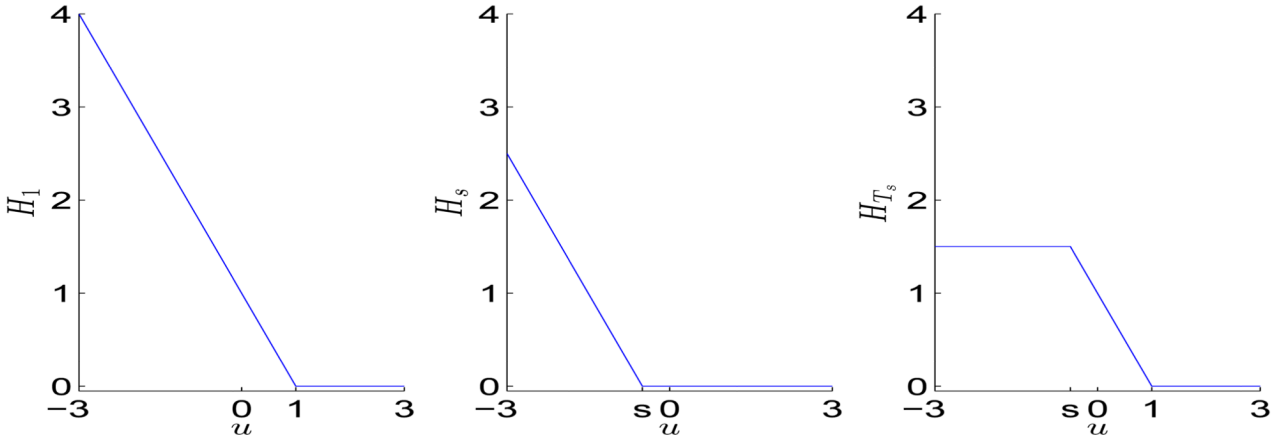

Figure 1.
Plots of $H_1(u)$, $H_s(u)$, and $H_{T_s}(u)$ (from left to right).

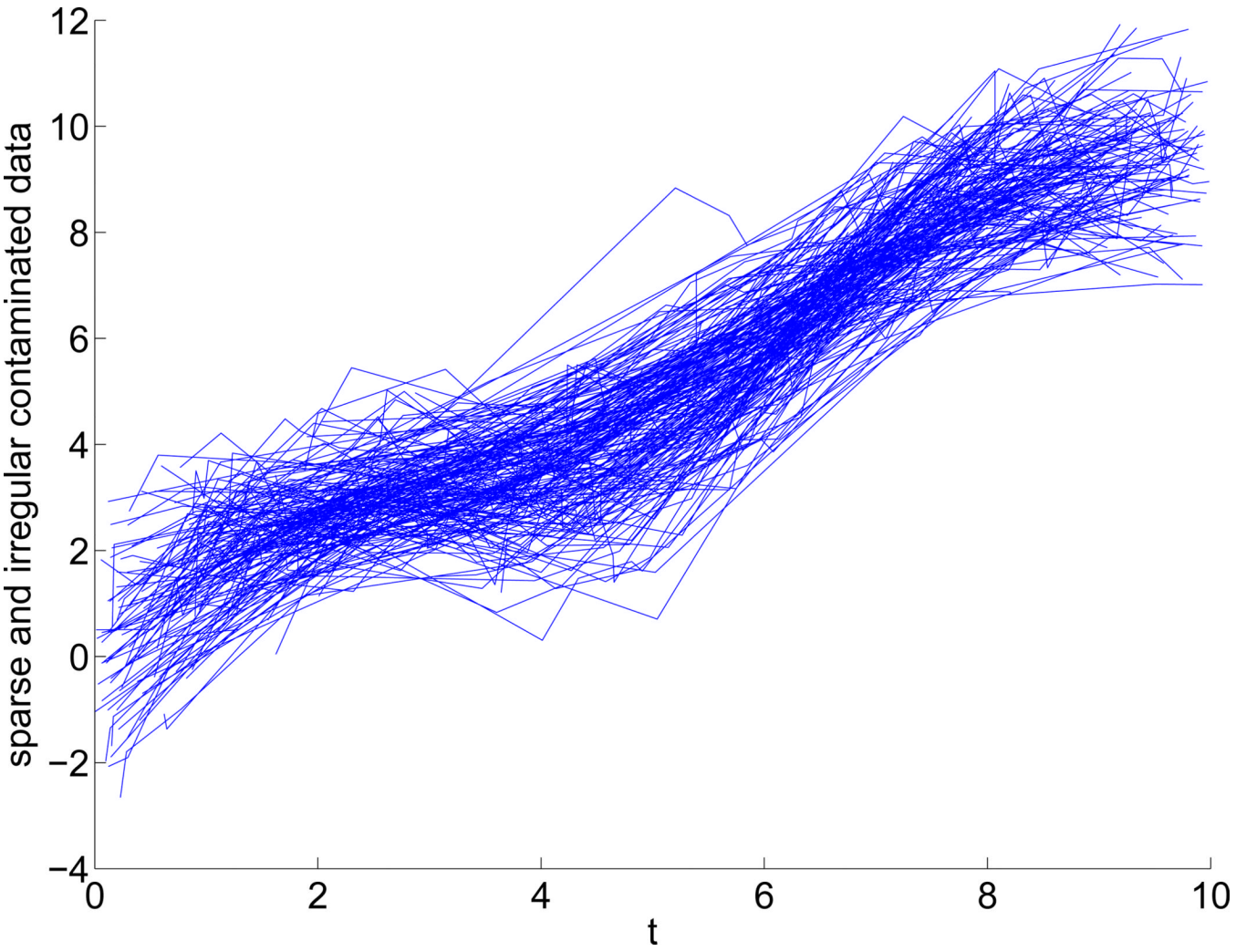

Figure 2.
Sparse and irregular functional data for one random repetition of Example 6.1.

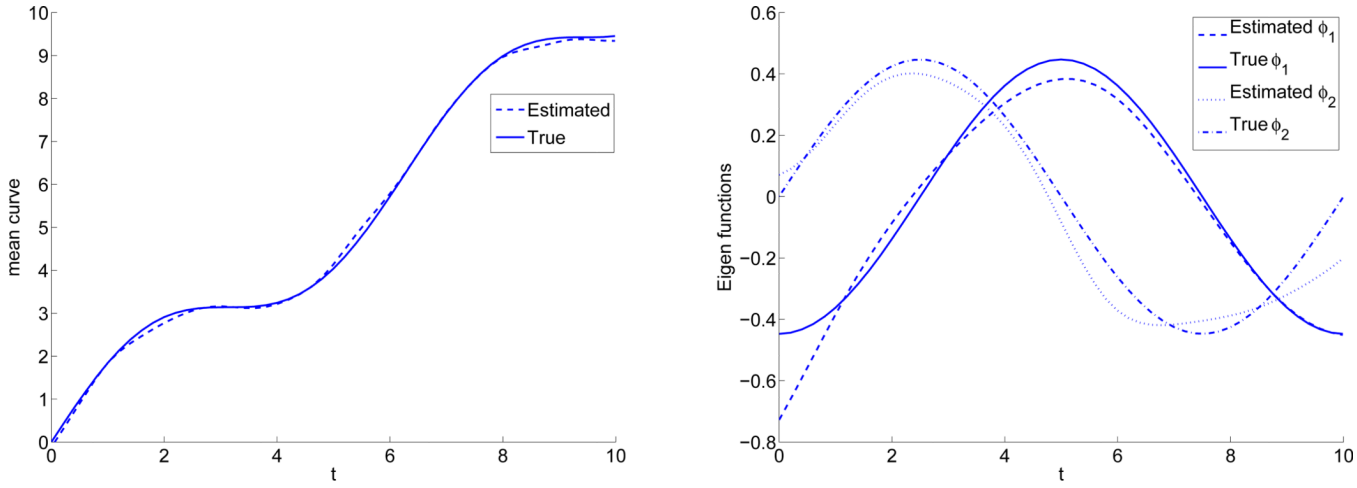

Figure 3.
The estimated mean curve and first two eigenfunctions given by the PACE for one random repetition of Example 6.1.

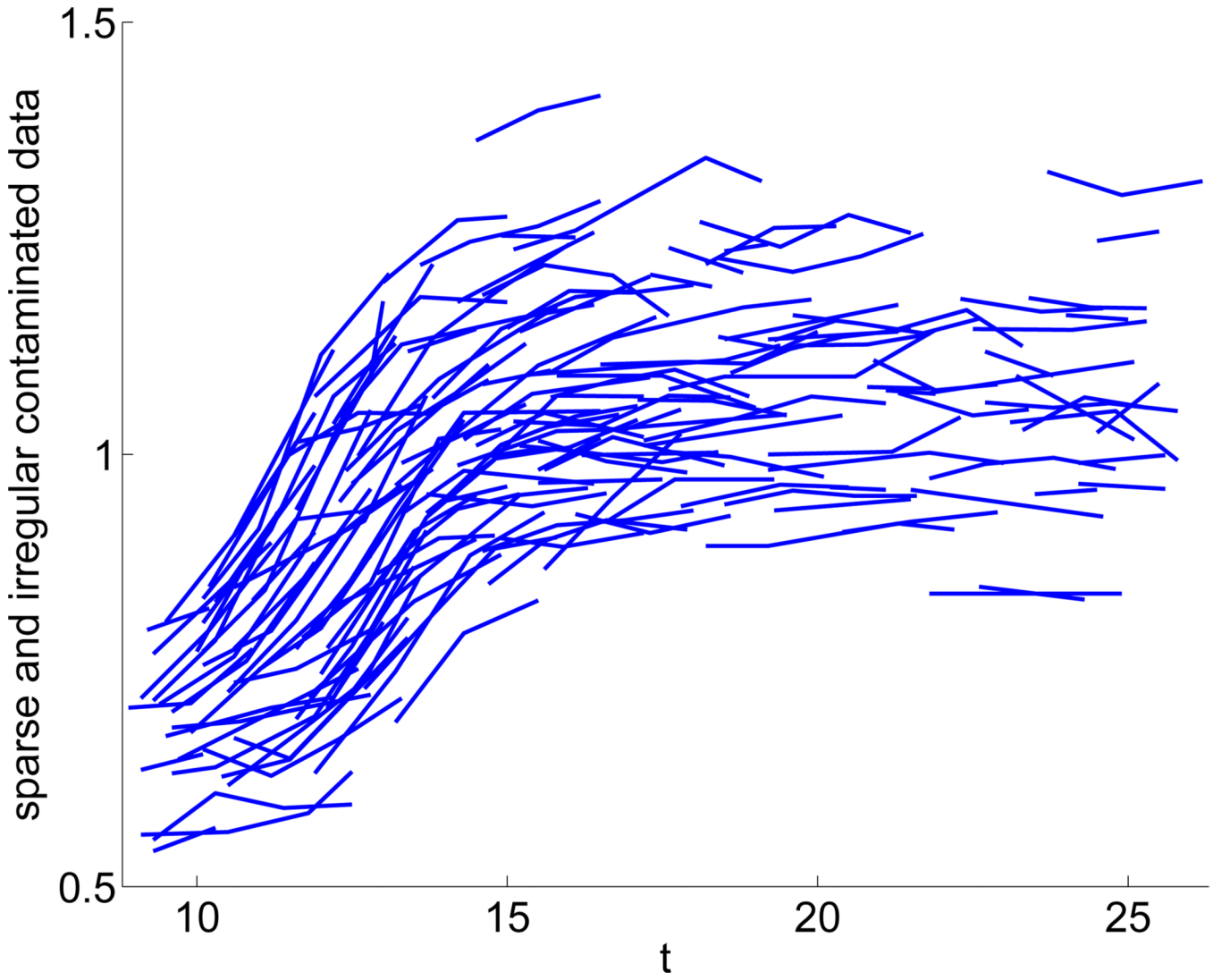

Figure 4.
Sparse and irregular functional data for the spinal bone mineral density data.

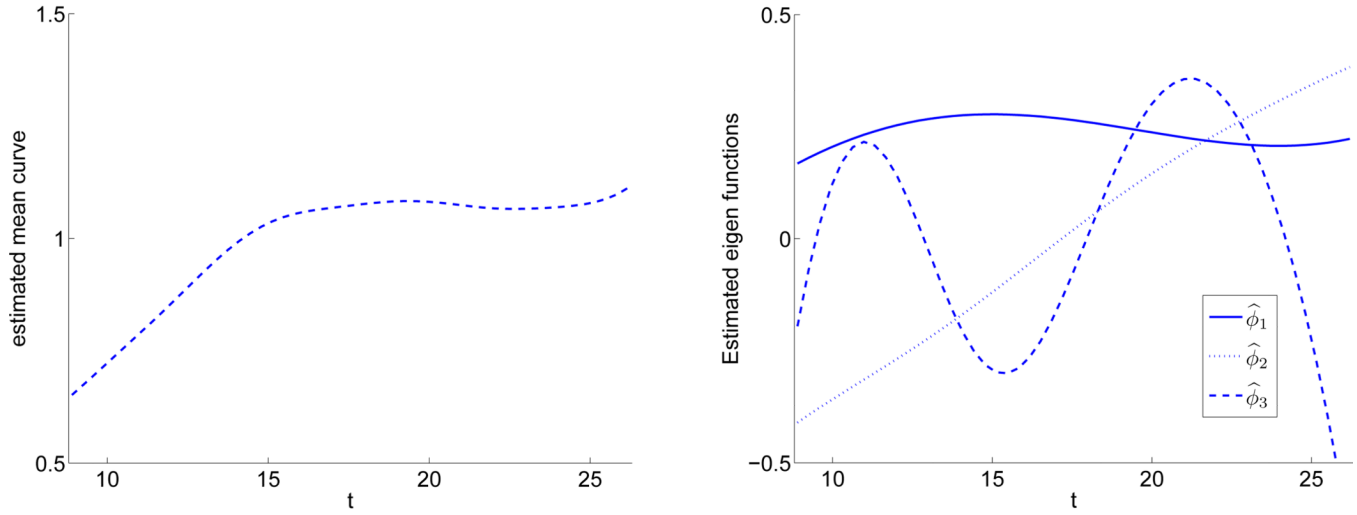

Figure 5.
The estimated mean curve and first three eigenfunctions given by the PACE for the spinal bone mineral density data.

Table 1

Confusion matrix of classifications for the four ethnicities

Prediction	Classifications for the following true ethnicities:				Total
	Asian	Black	Hispanic	White	
Asian	17(48.57%)	4(09.30%)	5(18.52%)	11(22.92%)	37
Black	12(34.29%)	35(81.40%)	11(40.74%)	17(35.42%)	75
Hispanic	2(05.71%)	1(02.33%)	2(07.41%)	3(06.25%)	8
White	4(11.43%)	3(06.98%)	9(33.33%)	17(35.42%)	33
Total	35(100.00%)	43(100.00%)	27(100.00%)	48(100.00%)	153