The Annals of Applied Statistics
2020, Vol. 14, No. 4, 1724–1742
https://doi.org/10.1214/20-AOAS1404
© Institute of Mathematical Statistics, 2020

CLASSIFICATION FROM ONLY POSITIVE AND UNLABELED FUNCTIONAL DATA

BY YOSHIKAZU TERADA[1,2], ISSEI OGASAWARA[3,*] AND KEN NAKATA[3,†]

[1] *Graduate School of Engineering Science, Osaka University, terada@sigmath.es.osaka-u.ac.jp*
[2] *RIKEN Center for Advanced Intelligence Project (AIP)*
[3] *Graduate School of Medicine, Osaka University, *ogasawaraissei@hss.osaka-u.ac.jp; †ken-nakata@umin.ac.jp*

In various fields, data recorded continuously during a time interval and curve data, such as spectral data, become common. These kinds of data can be interpreted as functional data. In this paper we have studied binary classification from only positive and unlabeled functional data (PU classification for functional data). Our first contribution is to present a simple classification algorithm for this problem. The key feature of the algorithm is that it is not required an estimation of the unknown class prior (or the constant probability that a positive object is labeled). It is worth noting that the idea of our method can be applied to kernel linear discriminant analysis for general data. Our second contribution is to prove that, under mild regularity conditions similar to those in a supervised context, the proposed algorithm can achieve perfect asymptotic classification in the context of PU classification. In fact, we show that the proposed algorithm works well not only in numerical experiments but also for real data examples. Moreover, as an important practical application, we have used the proposed algorithm to identify handball players at risk for anterior cruciate ligament (ACL) injury based on ground reaction force data.

1. Introduction. Functional data analysis (FDA) is commonly used in various fields such as chemometrics, sports medicine and biology. For example, it is often the case in the real data fields that data is recorded intermittently during a time interval, but the recorded time points could not be the same among objects. In this case we cannot directly apply the usual time series models. On the other hand, in FDA we consider a hidden stochastic process on the time interval, and we can deal with a variety of situations, including the above case, in a unified manner. For a general introduction to FDA, see Ramsay and Silverman (2002, 2005), and Wang, Chiou and Müller (2016). For the theoretical aspects of FDA, we refer the reader to Ferraty and Vieu (2006), Horváth and Kokoszka (2012) and Hsing and Eubank (2015).

Classifying functional data is one of the most important tasks in FDA. A number of classification methods for functional data have been proposed, and Section 4.2 of Wang, Chiou and Müller (2016) provides a helpful overview of these methods. From the Karhunen–Loève expansion of a random function, we can see the intrinsic high dimensionality of functional data. Focusing on this feature of functional data, we can obtain an interesting property of functional discriminant analysis. Delaigle and Hall (2012), Delaigle and Hall (2013) provide simple linear and quadratic functional discriminant methods (FLDA and FQDA) and show that, in the supervised classification problem for functional data, asymptotic perfect classification can often be achieved by using these methods.

In the standard binary classification problem for functional data, a training data $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ is obtained, where X_i is a curve on a compact interval $\mathcal{I}$, which is a feature of the ith object and Y_i is the group label of the ith object. The main purpose is to construct a classifier from the training data that can correctly predict the class label of a new curve whose class label is unknown. However, in some practical situations we cannot obtain the complete label information for the training data. For example, in the sports medicine field it

Received February 2019; revised September 2020.
Key words and phrases. Functional data analysis, binary classification, PU learning.

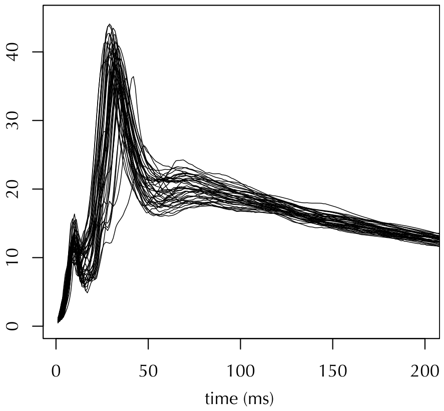


FIG. 1. *GRF data of subjects* 8 *and* 15 *who ruptured their right ACL.*

is important to identify players who are at risk for career-threatening injuries based on the various functional data reflecting individual motor dynamics. Here, we consider the problem to identify handball players who are at risk for anterior cruciate ligament (ACL) injury based on ground reaction force (GRF) data. The detailed description of this problem is described in Section 6. Figure 1 and Figure 2 show the GRF curves of injured subjects and those of noninjured subjects, respectively.

In the classical classification approach for this problem, to create the training data, each curve of injured players is assigned to the at-risk class (say positive class), and each curve of other players is assigned to the nonrisk class (say, negative class). Then, the usual binary classification method is applied for this training data. However, not all at-risk players need to have a serious injury during the experimental period. Whereas the injured players can be considered at-risk players, other players cannot be considered nonrisk players. Thus, we only have positive and unlabeled data in this situation, and the usual classification methods seem not to be appropriate.

The classification problem from only positive and unlabeled data is called *PU learning* or *PU classification*. This problem is the important issue in the various application fields, and PU classification recently has drawn much attention in the machine learning community (Blanchard, Lee and Scott (2010), du Plessis, Niu and Sugiyama (2014), du Plessis, Niu and Sugiyama (2015), du Plessis and Sugiyama (2014), Elkan and Noto (2008), Menon et al. (2015), Scott and Blanchard (2009)). A good survey of other practical situations of PU

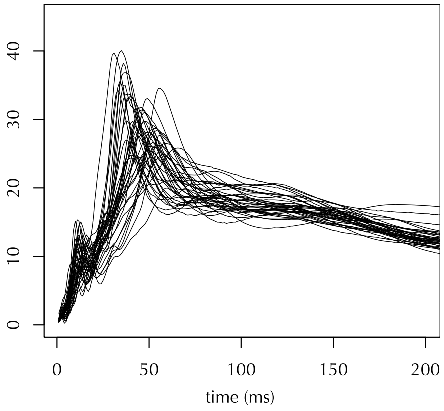
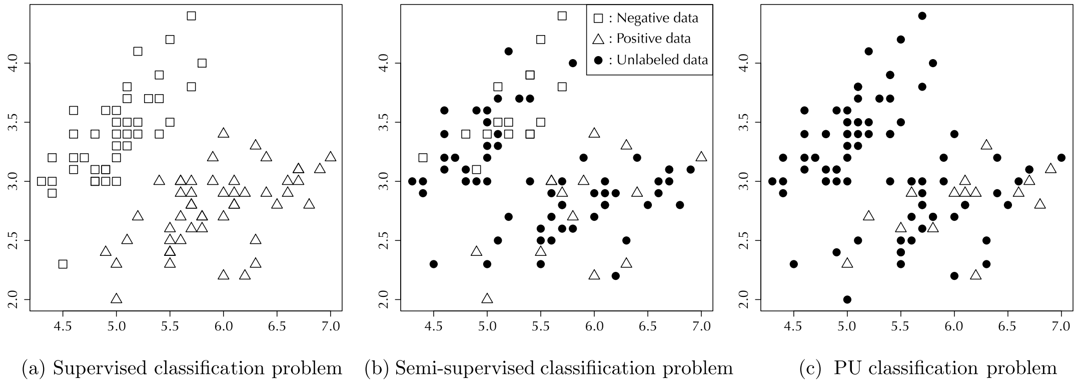


FIG. 2. *Randomly chosen* 40 *curves from GRF data of other subjects.*



(a) Supervised classification problem (b) Semi-supervised classifiication problem (c) PU classification problem

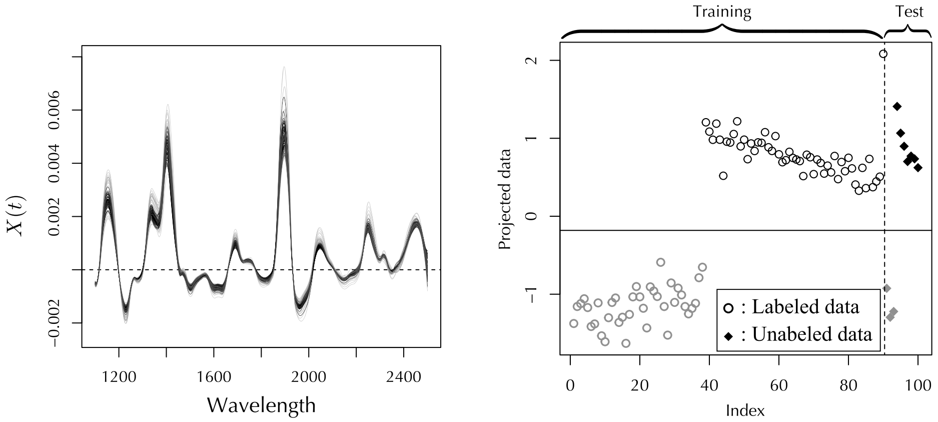
FIG. 3. *Images of training data (two-dimensional data) in three classification problems (Positive samples: White triangle; Negative samples: White square; Unlabeled samples: Filled circle).*

classification can be found in du Plessis, Niu and Sugiyama (2015). In Appendix A of the Supplementary Material (Terada, Ogasawara and Nakata (2020)), we briefly introduce the PU classification in a general setting (not only for functional data).

In this paper we consider the binary classification problem from only positive and unlabeled functional data. We refer to this classification problem as PU classification for functional data or functional PU classification. Here, we note that, in the usual semisupervised classification problem in FDA (e.g., Kawano (2013)), both positive and negative curves are available as labeled data and that the existing techniques cannot be applied for this problem. Figure 3 shows the differences in the training data among three classification problems. Now, we provide an illustrative example to see why the PU classification is required (why the supervised methods are not enough for the context of PU classification).

EXAMPLE 1. With the real data example we show the problem of the supervised classification method in the context of PU classification. We use the near-infrared spectral data of wheat samples with actual moisture content in Kalivas (1997). We describe the details about this data in Section 5. The wheat samples with moisture content less than 15% are considered negative samples, and other samples are considered positive samples. These labels are considered as the true labels in this example. To see the generalized performance of FLDA, we randomly select 10 samples as test samples, and the remaining samples (training data) are used to construct the classifiers. If we have fully labeled training data, FLDA often provides near perfect performance. In fact, Figure 4(b) shows the result of FLDA for the fully labeled training data. We see that FLDA can clearly classify positive and negative samples without any error.

Next, let us see how FLDA performs in the context of the PU classification. The labeled samples (40% of positive samples) are selected randomly from positive ones in the training data. We may consider unlabeled samples as negative ones to apply the supervised classification method for this partly labeled training data. In this way we forcibly apply FLDA for this training data consisting of only positive and unlabeled data. Figure 5(a) shows the classification result of FLDA with the projected data onto the discriminant subspace. From this result we can see that FLDA tried to distinguish positive labeled samples and unlabeled samples. However, since unlabeled data contains the positive samples in this setting, there are many misclassification samples not only for the training data but also for the test data. On the other hand, Figure 5(b) shows the result of the functional PU classification method (the proposed method) with the projected data. We can see that, by treating unlabeled samples appropriately, the PU learning detects an appropriate discriminant subspace in which we can



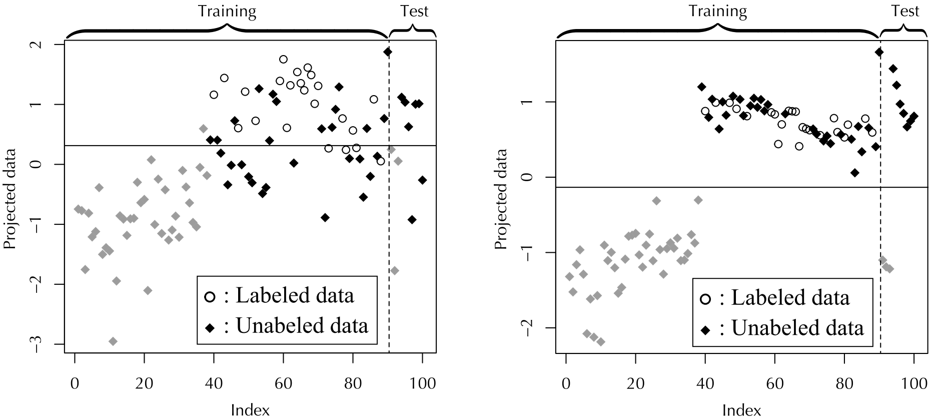
(a) The NIR spectra of the training data colored by the true labels.

(b) Projected data of FLDA for fully labeled training data.

FIG. 4. *The NIR spectra training data of wheat samples used in Example 1 and the result of FLDA in the context of the supervised classification in which all training samples are labeled. In both* (*a*) *and* (*b*), *the positive and negative samples are colored in black and grey*, *respectively.*

distinguish the underlying groups clearly. Here, we note that the true percentage (40%) of labeled samples in the positive training data is an unknown parameter and is not used in PU learning.

Throughout the paper we consider the setting introduced by Elkan and Noto (2008): only positive curves are labeled, and labeled positive curves are chosen completely randomly from all positive curves. Let π be the unknown class prior and λ be the constant probability that a positive curve is labeled. It is known that the PU classification problem can be solved by cost-sensitive learning between positive and unlabeled data if we know (or can consistently estimate) the values of λ or π (Appendix A of the Supplementary Material; du Plessis, Niu and Sugiyama (2014)). Thus, λ (or π) plays a key role in general PU classification problems. Although several methods for estimating λ or π have been proposed (Blanchard, Lee and Scott (2010), du Plessis and Sugiyama (2014), Elkan and Noto (2008), Jain et al. (2016),



(a) The result of FLDA in which unlabeled samples are treated as negative objects.

(b) The result of the proposed method (functional PU classification) with k-means.

FIG. 5. *Projected data of NIR spectra of wheat samples. The black line is the classification boundary in each method. The points are colored by the true labels* (*Class* 1: *Black*; *Class* −1: *Grey*).

Menon et al. (2015)), it is still difficult to construct a consistent estimator for λ or π in general. Most PU learning methods focus on multivariate data analysis (i.e., data space $\mathcal{X} \subseteq \mathbb{R}^d$). Note that the estimation framework of Blanchard, Lee and Scott (2010) can be applied to a general PU classification problem in theory. However, Blanchard, Lee and Scott (2010) mention that the empirical risk minimization algorithm described in Section 4 of their paper is computationally infeasible, and thus they employe a plug-in kernel density estimation classifier in their experiments. The kernel density estimation on $\mathbb{R}^d$ cannot directly be applied to functional data. Therefore, it is difficult that these methods are applied to functional data in general.

In this paper, based on the idea of linear discriminant analysis, we develop a new simple classification algorithm for the functional PU classification problem which utilizes the intrinsic high dimensionality of functional data. The proposed algorithm does *not require* an estimation of the unknown class prior π or the probability λ that a positive object is labeled. Moreover, we show that, even in the PU classification problem for functional data, asymptotic perfect classification can often be achieved using the proposed method. The proposed algorithm worked well not only in numerical experiments but also for real data examples. It is worth noting that the idea of our method can be applied to the PU learning problem for multivariate data. In fact, we describe the details about the kernel discriminant PU classification method for multivariate data in Appendix G of the Supplementary Material (Terada, Ogasawara and Nakata (2020)).

2. Preliminaries. Let (Y, R, X) and (Y_i, R_i, X_i) $(i = 1, \dots, n)$ be independent and identically distributed data triples, where Y_i is a class label taking the values -1 or 1, R_i is the response indicator for label Y_i, that is, $R_i = 1$ if Y_i is observed; otherwise, $R_i = 0$, and X_i is a random function defined on a compact interval $\mathcal{I}$. Here, as in many studies of functional data analysis, it is assumed that a sample of random functions is observed in the continuum and without measurement noise. In practice, however, functional data are observed discretely on a grid or at randomly distributed points, and substantial smoothing is used to convert these discrete data points into functions. The main theoretical property in this paper holds even without assuming that a sample of random functions is observed in the continuum and without measurement noise. This assertion is based on Theorem 1 in Hall, Müller and Wang (2006) and the results of Zhang and Wang (2016).

As done by Elkan and Noto (2008), we assume the following missing mechanism: $\mathbb{P}(R = 1 \mid X, Y = 0) = 0$ and $\lambda := \mathbb{P}(R = 1 \mid X, Y = 1) = \mathbb{P}(R = 1 \mid Y = 1) > 0$. The population is a mixture of subpopulations Π_{-1} and Π_1 corresponding to $Y_i = -1$ and $Y_i = 1$, respectively. Let π be the probability that a data curve comes from subpopulation Π_1, and assume $\pi > 0$. We write X_{ki} $(1 \le i \le n_k)$ for the ith function among functions for which the corresponding label equals k, and then $n = n_{-1} + n_1$. We denote by $n_{\text{obs}} = \sum_{i=1}^n \mathbb{1}(R_i = 1)$ the number of labeled positive objects. Suppose that X_{ki} is a second-order measurable process, and denote by $c_k(u, v) := \mathrm{Cov}_k(X(u), X(v))$ the covariance function of Π_k.

Henceforth, we assume the following general assumption: $\mathbb{E}[X \mid Y = 1] = 0$, $\mathbb{E}[X \mid Y = -1] = \mu \neq 0$ and that the covariance functions c_{-1} and c_1 are continuous, strictly positive definite and uniformly bounded. Let $\tilde{\mu} := \mathbb{E}[X] = (1 - \pi)\mu$. We can consistently estimate the mean function of the positive class from the labeled functions. Hence, the assumption $\mathbb{E}[X \mid Y = 1] = 0$ does not essentially affect the results presented in the following section. Note that, even in the setting of du Plessis, Niu and Sugiyama (2014), our method works, and the theoretical properties described in the following sections are still hold.

2.1. *Karhunen–Loève expansion and intrinsic high dimensionality of functional data.* Let (θ_{kj}, ϕ_{kj}) denote the eigenvalue and eigenfunction pairs of the integral operator with

TABLE 1
Essential differences between usual high-dimensional data and the intrinsic high dimensional nature of functional data

	High dim.	Functional
Dimensionality	Goes to infinity	Infinity
Correlation	Not zero	Zero
Variances	Nondegenerate	Degenerate
Variables	Observed	Latent

kernel c_k, where $\theta_{k1} \geq \theta_{k2} \geq \cdots > 0$. From Mercer's theorem (cf. Lemma 3.1 of Bosq (2000) or Theorem 4.6.5 of Hsing and Eubank (2015)), c_k has the following representation: $c_k(u, v) = \sum_{j=1}^{\infty} \theta_{kj}\phi_{kj}(u)\phi_{kj}(v)$. Since c_k is uniformly bounded, we have $\sum_{j=1}^{\infty} \theta_{kj} < \infty$. Moreover, from the Karhunen–Loève theorem (cf. Theorem 1.5 of Bosq (2000) or Theorem 7.3.5 of Hsing and Eubank (2015)), we have the following expansion (KL expansion) for a sample curve X'_k from subpopulation Π_k:

$$X'_k(u) = \eta_k(u) + \sum_{s=1}^{\infty} \sqrt{\theta}_{ks} Z_{ks}\phi_{ks}(u), \tag{2.1}$$

where $\eta_k := \mathbb{E}_k[X'_k]$ and Z_{ks} are real-valued random variables satisfying $\mathbb{E}_k[Z_{ks}] = 0$ and $\mathbb{E}_k[Z_{ks}Z_{kt}] = \delta_{s,t}$ for $s, t \in \mathbb{N}$. Since c_k is strictly positive definite so that $\{\phi_{kj}\}_{j\in\mathbb{N}}$ is a complete orthonormal system (CONS), we can write $\eta_k(u) = \sum_{j=1}^{\infty} \eta_{kj}\phi_{kj}(u)$ for the generalized Fourier decomposition of η_k with respect to the CONS $\{\phi_{kj}\}_{j\in\mathbb{N}}$. Combining these results, we can write

$$X'_k(u) = \sum_{s=1}^{\infty} W_{ks}\phi_{kj}(u), \tag{2.2}$$

where $W_{ks} = \eta_{ks} + \sqrt{\theta_{ks}} Z_{ks}$. Here, we have $\mathbb{E}_k[W_{ks}] = \eta_{ks}$, $\mathrm{Var}_k(W_{ks}) = \theta_{ks}$, and $\mathrm{Cov}_k(W_{ks}, W_{kt}) = 0$ $(s \neq t)$ for $s, t \in \mathbb{N}$. We can thus see that functional data X'_k is essentially constructed using high-dimensional data $\boldsymbol{W}_k^{(\infty)} = (W_{kj})_{j\in\mathbb{N}}$. However, in contrast to high-dimensional data, $\mathrm{Var}_k(W_{kj})$ degenerates with increasing j, and we cannot observe the value of W_{kj}. Moreover, we have $\sum_{j=1}^{\infty} \eta_{kj}^2 < \infty$ and $\eta_{kp}^2 \to 0$ as $p \to \infty$. We summarize in Table 1 the essential differences between the usual high-dimensional data and the intrinsic high dimensional nature $\boldsymbol{W}_k^{(\infty)}$ of functional data. Since the variance can be the amount of information the variable contains, functional data can be interpreted as an intermediate between finite-dimensional data and high-dimensional data.

3. PU classification for functional data.

Since most existing methods focus on PU classification problem for multivariate data, it is difficult to estimate the missing probability λ (or the class prior π) from functional data. As a simple approach for PU classification problem for functional data, we may consider the method to identify curves which are near to positive label curves (or the mean curve of the positive class) in the sense of L_2-distance. The advantage of this approach is that neither an estimation of λ nor an estimation of π is required. However, there are many situations in which the appropriate functional classification methods, such as Delaigle and Hall (2012), Delaigle and Hall (2013), perform well, but the methods based on the L_2-distance do not. Roughly speaking, the methods based on the L_2-distance work only when there is a visually clear difference between the mean curves of two classes. For example, there is no visually clear difference between two classes in the near-infrared reflectance (NIR) spectra data of wheat samples, described in Example 1 and Section 5.2, and the methods using the L_2-distance do not work well. In Appendix B of the Supplementary Material (Terada, Ogasawara and Nakata (2020)), we describe the theoretical reason why the L_2 distance is not appropriate for the PU classification problem of functional data. Thus, at first, we propose a new distance function which is more appropriate for the functional PU classification problem. Next, we describe the PU classification algorithm for functional data based on the proposed distance.

3.1. *New distance function for PU classification.* In the supervised classification problem, Delaigle and Hall (2012) show that the projection of functional data is useful to extract the intrinsic high dimensionality of functional data. This high dimensionality leads to an excellent classification performance, as shown in Figure 4(b). If we can construct the projection function of FLDA in the context of PU classification, then we can obtain a discriminant subspace in which hidden two groups can be clearly distinguished. The main part of the proposed method is to construct such an appropriate projection function from only positive and unlabeled samples. Since the negative samples are not labeled in the context of PU classification, it seems to be impossible. Surprisingly, we will show that, by focusing on the scale indeterminacy of the projection function, it is possible to construct an appropriate projection function or the discriminant subspace of FLDA in the context of PU classification.

Let X and X' be two independent sample curves from Π_1 and Π_{-1}, respectively. We temporarily assume that the covariance functions of two classes are the same, that is, $c_{-1} = c_1$. Here, we note that this assumption is not necessary to achieve good performance but is needed for only describing the main idea simply. In Section 4 we will prove that good performance can be achieved by the proposed method without this assumption. First, we consider an optimal projection of functional data into $\mathbb{R}$ for the binary classification problem. For a function ψ defined on $\mathcal{I}$, we consider the difference between the projected objects on $\mathbb{R}$ by ψ,

$$d^2(X, X' \mid \psi) := \langle X - X', \psi\rangle^2,$$

where $\langle f, g\rangle := \int f(u)g(u)\,du$ is the inner product of two functions f and g. Here, we have a question which ψ is best for distinguishing X and X'. By a simple calculation the expectation of the squared distance can be decomposed into the following two terms:

$$\mathbb{E}\big[d^2(X, X' \mid \psi)\big] = \langle \mu, \psi\rangle^2 + \mathbb{E}\big[\langle X - \tilde{X}', \psi\rangle^2\big],$$

where $\tilde{X}' = X' - \mathbb{E}[X']$. The first term and the second term can be considered the between-class dissimilarity (the mean difference) and the within-class dissimilarity (the within-class variance), respectively. Thus, we consider the following criterion of the separability:

$$Q(\psi) := \frac{\langle \mu, \psi\rangle^2}{\mathbb{E}[\langle X - \tilde{X}', \psi\rangle^2]}.$$

This criterion is commonly used as the objective function of the linear discriminant analysis. A large value of the separability $Q(\psi)$ means that two groups are well separated in the subspace spanned by ψ.

Now, we will find an optimal function ψ which maximizes the separability $Q(\psi)$. We recall that the eigenfunctions $\{\phi_j\}_{j\in\mathbb{N}}$ are a complete orthonormal system. Using $\{\phi_j\}_{j\in\mathbb{N}}$, a function can be expressed in the form of $\sum_{j=1}^{\infty} a_j\phi_j$. We denote by $\psi = \sum_{j=1}^{\infty} a_j\phi_j$ the generalized Fourier series expansion of ψ and by $\psi_r := \sum_{j=1}^{r} a_j\phi_j$ the truncated expansion for each $r \in \mathbb{N}$. Then, we have

$$\langle \mu, \psi\rangle^2 = \left(\sum_{j=1}^{\infty} \mu_j a_j\right)^2$$

and

$$\mathbb{E}[\langle X - \tilde{X}', \psi\rangle^2] = \mathrm{Var}[\langle X, \psi\rangle] + \mathrm{Var}[\langle X', \psi\rangle] = 2\sum_{j=1}^{\infty} \theta_j a_j^2.$$

By choosing the eigenfunctions as a CONS, the separability $Q(\psi)$ can be rewritten as

$$Q(\psi) = Q(\{a_j\}) = \frac{(\sum_{j=1}^{\infty} \mu_j a_j)^2}{2\sum_{j=1}^{\infty} \theta_j a_j^2}.$$

Here, we note that the representation of $Q(\psi)$ with the generalized Fourier coefficients is more complicated if we choose the other CONS to expand ψ. Since we cannot deal with infinitely many parameters in practice, we consider only the truncated function $\psi_r = \sum_{j=1}^{r} a_j \phi_j$.

From the proof of Theorem 1 in Delaigle and Hall (2012), we can maximize $Q(\psi_r)$ with respect to ψ_r by taking $a_j = b \times \theta_j^{-1} \mu_j$ $(j = 1, \dots, r)$ for any nonzero constant $b \in \mathbb{R} \setminus \{0\}$. The important point to note here is that an optimal function has *an indeterminacy of scale*. That is, for an optimal function $\psi_r^{(1)} = \sum_{j=1}^{r} \theta_j^{-1} \mu_j \phi_j$, functions $\psi_r^{(b)} = b \times \sum_{j=1}^{r} \theta_j^{-1} \mu_j \phi_j$ with $b \neq 0$ are also optimal. Thus, substituting $(1 - \pi)$ for a constant b, we notice that the following function is also optimal:

$$\psi_r = (1 - \pi) \sum_{j=1}^{r} \theta_j^{-1} \mu_j \phi_j = \sum_{j=1}^{r} \theta_j^{-1} (1 - \pi) \mu_j \phi_j = \sum_{j=1}^{r} \theta_j^{-1} \tilde{\mu}_j \phi_j.$$

If we can estimate the eigenvalues and eigenfunctions $\{(\theta_j, \phi_j)\}_{j=1}^{r}$ of the covariance and the function $(1 - \pi)\mu$, then we can construct an optimal ψ_r empirically. Here, we recall that, in the context of PU classification, (θ_j, ϕ_j) can be estimated by using functional principal component analysis (FPCA) for the labeled data, and $\tilde{\mu} = (1 - \pi)\mu$ can be estimated by averaging all curves. Therefore, we can empirically construct an optimal ψ_r, even in the context of PU classification.

To obtain deeper insight into why the projection by ψ is useful, we focus on the optimal value of $Q(\psi)$ which is given by

$$\max Q(\psi_r) = \frac{1}{2} \sum_{j=1}^{r} \theta_j^{-1} \mu_j^2.$$

We recall that $\sum_{j=1}^{\infty} \theta_j < \infty$, so $\theta_j \searrow 0$ as $j \to \infty$ and also recall that the mean difference between the two groups is denoted by μ. Even when the mean difference $\|\mu\|^2 = \sum_{j=1}^{\infty} \mu_j^2 < \infty$ takes a small value, the maximum value $\max Q(\psi_r) = \frac{1}{2} \sum_{j=1}^{r} \theta_j^{-1} \mu_j^2$ may diverge to infinity or may take a large value as $r \to \infty$. Focusing on the term $\sum_{j=1}^{r} \theta_j^{-1} \mu_j^2$, we can see that the projection by ψ amplifies the difference between two groups. (The squared mean difference μ_j^2 is amplified by multiplying θ_j^{-1}.) Since a large value of $Q(\psi_r)$ means that two groups are well separated in the subspace spanned by ψ_r, this implies that the projected data by ψ_r may be clearly separated, even if there is only a small difference in the sense of the L_2 distance. Thus, we can extract the intrinsic high dimensionality of functional data effectively by using the projection. We can expect a good classification performance in the single-dimensional subspace with the optimal ψ_r.

Based on these ideas, we propose a new distance between two functions for PU classification.

DEFINITION 1. For $r \in \mathbb{N}$, the (truncated) PU distance between two functional objects X_i and X_j is defined as

$$d_r(X_i, X_j) := \left|\langle X_i - X_j, \psi_r\rangle\right|,$$

where $\psi_r := \sum_{s=1}^{r} \theta_{1s}^{-1}\tilde{\mu}_{1s}\phi_{1s}$, $\tilde{\mu}_{1s} := \int \tilde{\mu}(u)\phi_{1s}(u)\,du$, and $\tilde{\mu} = (1-\pi)\mu$ is the mean function of the population $\pi\Pi_{+1} + (1-\pi)\Pi_{-1}$.

Obviously, this distance function satisfies the axioms of distance. The following proposition shows the fundamental property of this distance function:

PROPOSITION 3.1. *For sample curves X_i and X_j $(i \neq j)$, we have*

$$d_r(X_i, X_j) = \begin{cases} \sqrt{2}\alpha_{kr}|R| & \text{if } Y_i = Y_j = k, \\ |\kappa \times \alpha_{1r}^2 + \alpha_r R'| & \text{if } Y_i \neq Y_j, \end{cases}$$

where R and R' are random variables satisfying $\mathbb{E}[R] = \mathbb{E}[R'] = 0$ *and* $\mathrm{Var}(R) = \mathrm{Var}(R') = 1$, *respectively*, $\kappa := 1/(1-\pi)$, $\alpha_r^2 := \alpha_{-1r}^2 + \alpha_{1r}^2$, *and* $\alpha_{kr}^2 := \sum_{s=1}^{\infty} \theta_{ks}\langle\phi_{ks}, \psi_r\rangle^2$ $(k = -1, 1)$.

PROOF. See Appendix C of the Supplementary Material (Terada, Ogasawara and Nakata (2020)). □

Here, we note that, in the proposition, we do not assume that two covariance functions c_{-1} and c_{+1} are the same. Since the term α_{kr} monotonically increases with r, we expect that $\alpha_{1r}^2 \gg \alpha_{kr}$ and thus $d_r(U, V) \gg d_r(V, V')$ for random functions $U \sim \Pi_k$ and $V, V' \sim \Pi_\ell$ $(k, l = -1, 1; k \neq l)$. That is, with increasing r, the PU distance d_r between two curves from different classes may take a much larger value than the PU distance between two curves in the same class. Thus, we can expect that two classes are distinguished clearly using the clustering method with the PU distance d_r.

3.2. *PU classification algorithm based on PU distance.* From the above discussion, it is expected that the projected data $\langle X_i, \psi_r\rangle$ $(i = 1, \dots, n)$ are well separated in accordance with class labels. In practice, θ_{1j}, ϕ_{1j} and $\tilde{\mu}_{1j}$ are unknown and must be estimated from the data to construct the estimator of ψ_r. Let $\{(\hat{\theta}_{1j}\hat{\phi}_{1j})\}_{j=1}^{r}$ be the estimator of $\{(\theta_{1j}, \phi_{1j})\}_{j=1}^{r}$ obtained by FPCA of the labeled data. We can estimate $\tilde{\mu}$ by

$$\hat{\tilde{\mu}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_1),$$

where $\hat{\mu}_1 = \sum_{R_i=1} X_i/n_{\text{obs}}$. In addition, $\tilde{\mu}_{1j}$ also can be estimated by using

$$\hat{\tilde{\mu}}_{1j} = \int \hat{\tilde{\mu}}(u)\hat{\phi}_{1j}\,du.$$

Hence, we can construct an estimator of ψ_r as follows:

$$\hat{\psi}_r = \sum_{j=1}^{r} \hat{\theta}_{1j}^{-1}\hat{\tilde{\mu}}_{1j}\hat{\phi}_{1j}.$$

By the construction of the estimator $\hat{\psi}_r$, the projected data $\{\langle X_i, \hat{\psi}_r\rangle\}_{i=1}^{n}$ should be similar to the projected data of FLDA with fully labeled training data. In fact, Figure 6 shows the comparison between the (Procrustes-adjusted) projected data by FLDA with fully labeled training data and the projected data by the proposed method with only positive and unlabeled data in Example 1. This figure shows that the discriminant subspace constructed by

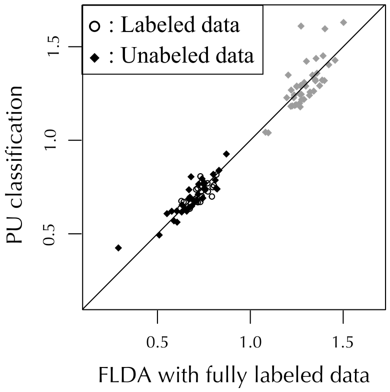


FIG. 6. *The comparison between the* (*Procrustes-adjusted*) *projected data by FLDA with fully-labeled data and the projected data by functional PU classification* (*the proposed method*). *The points are colored by the true labels* (*Class* 1: *Black*; *Class* −1: *Grey*).

the proposed method is reasonably similar to the discriminant subspace of FLDA with fully labeled training data. Here, we again emphasize that the proposed method uses incompletely labeled training data (only positive and unlabeled data), whereas FLDA uses completely labeled training data in this experiment. Thus, we see that an appropriate discriminant subspace can be constructed from only positive and unlabeled data.

Let us denote by $\hat{d}_r(\cdot, \star)$ the estimated PU distance by $\hat{\psi}_r$. The PU classification based on the estimated PU distance can be described as Algorithm 1. It is worth noting that the idea of our method can be applied to (kernel) linear discriminant analysis (LDA) for multivariate PU learning problems. For the details about the PU classification algorithm based on kernel LDA, see Appendix G of the Supplementary Material (Terada, Ogasawara and Nakata (2020)).

In the proposed algorithm we need to choose r first, and the performance is affected by choice of r. To choose an appropriate r, empirically, we suggest using the leave-m labeled data-out cross-validation described below. We randomly create p partitions of the labeled data. For $s = 1, \dots, p$, each partition splits the labeled data into two subsamples, $\{X^*_{1s}, \dots, X^*_{ms}\}$ and $\{X^*_{(m+1)s}, \dots, X^*_{n_{\text{obs}}s}\}$, where $\{X^*_{1s}, \dots, X^*_{n_{\text{obs}}s}\}$ denote a random permu-

Algorithm 1 PU classification based on PU distance

```
Set $r$ to number of principal components constructing $\hat{\psi}_r$.
Compute $\hat{\mu}_1 = \sum_{R_i=1} X_i/n_{\text{obs}}$ and $\hat{\hat{\mu}} = \sum_{i=1}^{n}(X_i - \hat{\mu}_1)/n$.
Apply FPCA to the labeled data and obtain $(\hat{\theta}_{1j}, \hat{\phi}_{1j})$ $(j = 1, \dots, r)$.
Compute $\hat{\hat{\mu}}_{1j} = \int \hat{\hat{\mu}}(t)\phi_{1j}(t)\,dt$ $(j = 1, \dots, r)$.
Compute $Z_i = \langle X_i, \hat{\psi}_r \rangle$ $(i = 1, \dots, n)$ where
    $$\hat{\psi}_r = \sum_{j=1}^{r} \hat{\theta}_{1j}^{-1} \hat{\hat{\mu}}_{1j} \hat{\phi}_{1j}.$$
Use conventional clustering algorithm (e.g., Ward's clustering) for $Z_1, \dots, Z_n$ to divide all objects into two clusters.
Assign positive label $\hat{Y}_i = 1$ if unlabeled object $Z_i$ is in the cluster containing more labeled objects than the other cluster; assign negative label $\hat{Y}_i = -1$ otherwise.
```

tation of the labeled data. Then, we estimate the classification error using

$$\widehat{\mathrm{err}}(r) = \frac{1}{pn_{\mathrm{obs}}} \sum_{s=1}^{p} \sum_{i=1}^{n_{\mathrm{obs}}} \mathbb{1}(\hat{Y}_{-s}(X_i^* \mid r) \neq 1),$$

where $\hat{Y}_{-s}(X_i^* \mid r)$ denotes the label of X_i^* estimated using $\{X_{(m+1)s}^*, \dots, X_{n_{\mathrm{obs}}s}^*\}$. Then, we choose r by minimizing $\widehat{\mathrm{err}}(r)$.

4. Theoretical properties.

In this section we establish the theoretical properties of the proposed classification algorithm in which we construct the estimator of ψ_r, empirically. The following theorem provides the theoretical guarantee for the performance of the proposed algorithm with conventional hierarchical clustering such as Ward's method. Note that we *do not assume* that two covariance functions, c_{-1} and c_{+1}, are the same in the following theorem:

THEOREM 4.1. *We assume the general assumption introduced in Section 2. In addition, we suppose that*:

(a) $\sup_{t\in\mathcal{I}} \mathbb{E}_k[|X(t)|^4] < \infty$ *for* $k = -1$ *and* $k = 1$,
(b) *there are no ties among the eigenvalues* θ_{1j} *and*
(c)

$$\left(\sum_{j=1}^{r} \theta_{1j}^{-1} \tilde{\mu}_{1j}^2\right)^2 \Big/ \left(\sum_{j=1}^{\infty} \theta_{kj} \langle \phi_{kj}, \psi_r \rangle^2\right) \to \infty$$

as $r \to \infty$ *for* $k = -1$ *and* $k = 1$.

Let $\{\eta_n\}_{n\in\mathbb{N}}$ *be a decreasing sequence such that* $\eta_n \to 0$ *and* $n\eta_n^5 \to \infty$ *as* $n \to \infty$. *Let* $\hat{R}_n + 1 := \inf\{j \in \mathbb{N} \mid \hat{\theta}_{1j} - \hat{\theta}_{1,j+1} < \eta_n\}$ *and* r_n *be an increasing sequence such that* $r_n \to \infty$ *and* $r_n \le \hat{R}_n$. *Then, for any random functions* $U \sim \Pi_k$ *and* $V, V' \sim \Pi_\ell (k, l = -1, 1; k \neq l)$,

$$\mathbb{P}(\hat{d}_{r_n}(U, V) > \hat{d}_{r_n}(V, V')) \to 1 \quad \text{as } n \to \infty.$$

Moreover, if we assume the following condition instead of condition (*a*):

(a′) $\exists M > 0; \mathbb{P}(\|X\| < M) = 1$;

then,

$$\mathbb{P}\Big(\min_{Y_i \neq Y_j} \hat{d}_{r_n}(X_i, X_j) > \max_{1\le k\le n} \min_{\substack{k\neq l,\\ Y_k = Y_l}} \hat{d}_{r_n}(X_k, X_l)\Big) \to 1 \quad \text{as } n \to \infty, \tag{4.1}$$

and thus the probability that there are no misclassified curves converges to one.

PROOF. See Appendix D of the Supplementary Material (Terada, Ogasawara and Nakata (2020)). □

This theorem ensures that perfect asymptotic classification is often possible using the proposed method even in the PU classification context. Condition (a') is simply for mathematical convenience. If we take η_n as $n\eta_n^5 = \Omega(n^{1/2})$, then condition ($a'$) can be replaced by condition (a) even for (4.1). Assumption (c) requires that the covariance functions of the two classes not be much different. For detailed discussions about the conditions of Theorem 4.1, see Appendix E of the Supplementary Material (Terada, Ogasawara and Nakata (2020)).

We recall that the proposed method constructs the discriminant subspace of FLDA with fully labeled data, approximately. Hence, at least when two covariances are the same, the performance of the proposed method cannot be better than the performance of FLDA with fully labeled data. This is one of the limitations of the proposed method. Moreover, we can use only the positive labeled data to estimate the eigenvalues and eigenfunctions. When the sample size of the positive labeled data is too small (e.g., n_{obs} less than 10), the proposed method fails to construct an appropriate subspace. In other words, whenever FLDA with fully labeled training data can provide good performance and the sample size is moderately large, the proposed method also can provide a comparable performance from only positive and unlabeled data.

5. Numerical experiments.

5.1. *Simulated examples.* The performance of the proposed algorithm was evaluated through numerical experiments. We used the settings of the numerical examples in Delaigle and Hall (2012). Let $\mathcal{I} = [0, 1]$. For $i = 1, \dots, n_k$ $(k = -1, 1)$, we took $X_{ki} = \sum_{j=1}^{40} (\eta_{kj} + \sqrt{\theta^*_{kj}} Z_{kj})\phi_j$ where $\phi_j(t) = \sqrt{2}\sin(\pi j t)$. As Z_{kj}s, we generated independent standard normal random variables $Z_{kj} \sim_{\text{i.i.d.}} N(0, 1)$ or independent exponential variables $Z_{kj} - 1 \sim_{\text{i.i.d.}} \mathrm{Exp}(1)$. In all settings we generated n curves from two subpopulations Π_{-1} and Π_1, and we set $n_{-1} = n_1 = n/2$ and $n_{\mathrm{obs}} = n_1/2$. We completely randomly selected n_{obs} curves as the labeled objects from curves generated from subpopulation Π_1. We used the following four settings:

Setting 1. The Z_{kj}s were independently generated from the centered exponential distribution Exp(1), that is, $Z_{kj} - 1 \sim_{\text{i.i.d.}} \mathrm{Exp}(1)$. For $j = 1, \dots, 40$, we set $\theta^*_{kj} = j^{-1}$ $(k = -1, 1)$, for $j > 6$ we set $\eta_{kj} = 0 (k = -1, 1)$ and we set

$$(\eta_{-1,1}, \eta_{-1,2}, \eta_{-1,3}, \eta_{-1,4}, \eta_{-1,5}, \eta_{-1,6}) = (0, -0.75, 0.75, -0.15, 1.4, 0.1)$$
$$\text{and} \quad (\eta_{1,1}, \eta_{1,2}, \eta_{1,3}, \eta_{1,4}, \eta_{1,5}, \eta_{1,6}) = (0, -0.5, 1, -0.5, 1, -0.5).$$

Setting 2. We used the same settings for Setting 1, but, for the subpopulation Π_1, we replaced θ^*_{1j} with $\theta^*_{1j} = 1.5 \times \theta^*_{-1j}$.

Setting 3. We used the same settings for Setting 1, except for $X_{1i} = \sum_{j=1}^{40} (\eta_{1j}\phi_j + \sqrt{\theta^*_{1j}} Z_{1j}\phi_{1j})$, where $\phi_{1j} = \sqrt{2}\cos(\pi j t)$.

Setting 4. $Z_{kj} \sim_{\text{i.i.d.}} N(0, 1)$. For $j = 1, \dots, 40$, we set

$$\theta^*_{kj} = \exp\{-[2.1 - (j-1)/20]^2\} \quad (k = -1, 1),$$

$\eta_{1j} = 0$ and $\eta_{-1j} = 0.75(-1)^{j+1}\mathbb{1}(j \le 3)$.

For each setting, $B = 100$ independent samples were generated. For each sample we applied the proposed algorithm with empirically chosen $\hat{r} = \min_r \widehat{\mathrm{err}}(r)$ and the plug-in algorithm with the true class-prior π described in Appendix F of the Supplementary Material (Terada, Ogasawara and Nakata (2020)). We calculated the misclassification rate, $P_{\mathrm{Miss}} = 100 \times \sum_{i=1}^{n} \mathbb{1}(\hat{Y}_i \neq Y_i)\%$, for the estimated labels of each method. Note that the class prior is unknown in practice. Table 2 shows the means and standard deviations of the values of P_{Miss}.

Settings 1 to 3 are for the cases in which the curves are quite dispersed, but the mean curves are not much different. The covariance functions of the two groups are clearly different in Settings 2 and 3. In Figure 7 the first column illustrates an example of labeled data, and the second column in Figure 7 shows an example of unlabeled data, colored in accordance with the true label. Distinguishing the curves into two classes appears difficult. However, in these cases, $\sum_{j=1}^{r} \theta_{1j}^{-1}\mu_j^2$ is large even for small r, where θ_{1j} is the jth eigenvalue of the

TABLE 2

Percentage of misclassifications for simulated and real data examples: mean of P_{Miss} with standard deviation of P_{Miss}, calculated from $B = 100$ simulations

			Error (Std)	
Data set	n	n_{obs}/n_1	Proposed	Plug-in
Setting 1	100	25/50	3.88% (3.46)	3.77% (2.22)
Setting 2	100	25/50	7.73% (4.39)	7.08% (3.09)
Setting 3	100	25/50	7.63% (3.84)	5.80% (3.22)
Setting 4.1	100	25/50	44.97% (6.97)	39.2% (7.64)
Setting 4.2	200	50/100	3.63% (4.52)	5.85% (2.23)
Wheat	100	25/59	1.54% (1.30)	3.72% (5.11)
Tecator	215	30/138	5.04% (4.53)	10.66% (4.92)

covariance operator of the positive class and μ_j is the projection, on the jth eigenfunction, of the difference between the mean curves of the two classes. The first three rows of Table 2 show that the proposed method provided nearly perfect classification results in these three cases. Note that, although the covariance functions are not the same in Settings 2 and 3, the proposed method worked nicely. This result supports the theoretical property of the proposed method in Theorem 4.1, empirically. It is worth noting that the performance of the proposed algorithm is comparable with the performance of the plug-in algorithm, although the proposed algorithm does not require the estimation of the class prior.

Setting 4 is for the case in which $\sum_{j=1}^{r} \theta_{1j}^{-1} \mu_j^2$ is not large for small r. In this case we have $\sum_{j=1}^{r} \theta_{1j}^{-1} \mu_j^2 = 0$ for $r < 38$. Since the proposed method is mainly based on FPCA for the labeled data, we need to estimate reasonably well the smallest 38th–40th eigenvalues and the corresponding eigenfunctions to obtain good performance with the proposed method. That is, in this setting the number of labeled curves must be somewhat larger than 40. Focusing on the fourth row in Table 2 and the projected data in the second column in Figure 7, we see that the proposed method did not work well when $n_{\text{obs}} = 25$ ($n = 100$). On the other hand,

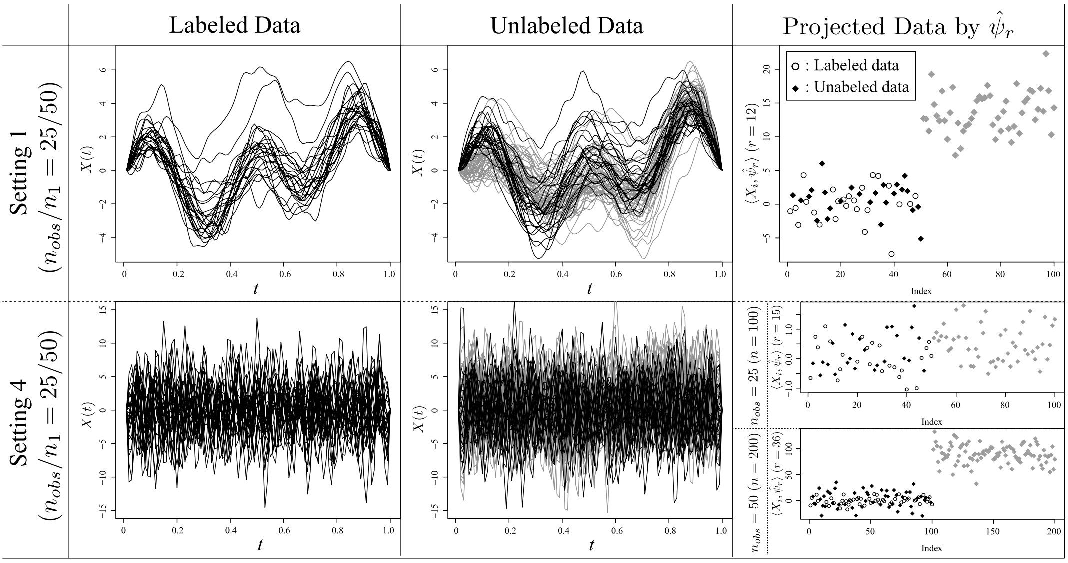


FIG. 7. *Results of numerical experiments: Labeled data, unlabeled data (colored in accordance with true label—Class +1: Black; Class −1: Grey) and data projected using $\hat{\psi}_r$ for each numerical experiment.*

focusing on the fifth row in Table 2 and the projected data in the second column of Figure 7, we see that the proposed method worked very well when $n_{\text{obs}} = 50$ ($n = 200$).

5.2. *Real-data examples.* We applied the proposed algorithm and the plug-in algorithm with the true class-prior to two chemometrics data sets. For each one we generated PU data by randomly selecting n_{obs} curves from subsample Π_1, and we applied the proposed algorithm to the data with empirically chosen $\hat{r}$. Then, we evaluated the misclassification rate, $P_{\text{Miss}} = 100 \times \sum_{i=1}^{n} \mathbb{1}(\hat{Y}_i \neq Y_i)\%$. For each data set, we generated PU data $B = 100$ times and then calculated the B values of P_{Miss}. The means and standard deviations of these values are shown in Table 2.

In the first example we use the near-infrared spectral data set described in Kalivas (1997). It contained data on the near-infrared reflectance (NIR) spectra of 100 wheat samples with known moisture content, measured in *two* nm intervals from 1100 to 2500 nm. Here, we refer to this data set as Tecator data. This data set is available at `fds`-package (Shang and Hyndman (2013)) of R. We used the moisture content to separate the data into two subpopulations, Π_{-1} (moisture content less than 15) and Π_1 (moisture content greater than 15). Here, $n_{-1} = 41$ and $n_1 = 59$. We set $n_{\text{obs}} = 25$. In accordance with the custom of chemometrics data analysis, we used the derivative curves of the spectra which were estimated using the method described in Ferraty and Vieu (2006). The first row of Figure 8 shows an example of the labeled and unlabeled data and projected data $\langle X_i, \hat{\psi}_r \rangle$ of these curves. For the PU data in Figure 8, $\hat{r} = 10$ was estimated. The average for 100 values of the misclassification rate was very small, indicating that nearly-perfect classification was achieved from only positive and unlabeled data.

In the second example, we used the NIR spectral data set described in Ferraty and Vieu (2006). It contained data on the NIR spectra of 215 pieces of finely chopped meat, measured in *two* nm intervals from 850 to 1050 nm. We also had the percentages of fat, water and protein for each sample. This data set is available at `fda.usc`-package (Bande et al. (2014)) of R. We assigned a label to each spectrometric curve X_i as follows: $Y_i = -1$, if the fat content was less than 20; $Y_i = 1$, if the fat content was greater than 20. Here, $n_{-1} = 77$ and $n_1 = 138$. We also used the derivative curves of the spectra. Figure 9 shows the covariances of

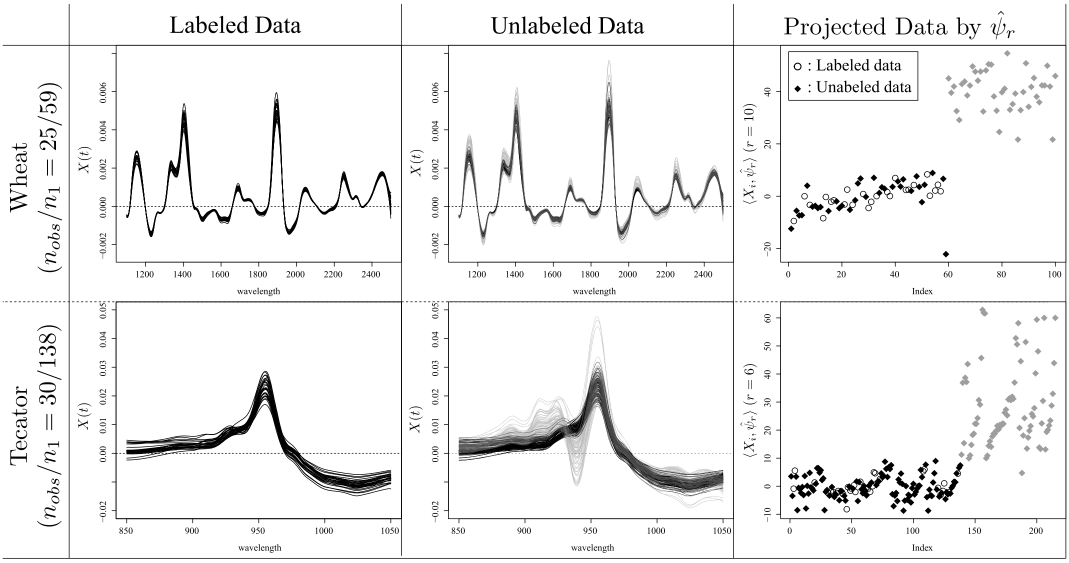
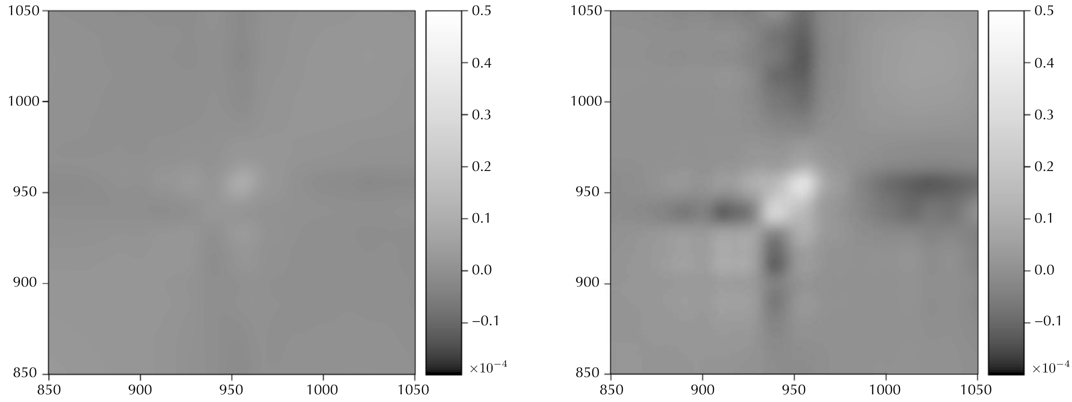


FIG. 8. *Results of real data examples*: *Labeled data*, *unlabeled data* (*colored in accordance with true label—-Class* +1: *Black*; *Class* −1: *Grey*) *and data projected using* $\hat{\psi}_r$ *for each real data example.*



(a) Positive data ($Y = +1$)

(b) Negative data ($Y = -1$)

FIG. 9. *The sample covariance functions of the positive and negative groups in Tecator data.*

two groups, and we can see that there is a clear difference between in the covariance structures of two groups. The number of observations, n_{obs}, was 30. The second row in Figure 8 shows an example of the labeled and unlabeled data and the data projected using $\hat{\psi}_r$. For these data, $\hat{r} = 6$ was chosen. The average for 100 values of the misclassification rate was again very small, confirming that the proposed method performs very well, even when two covariances are different, as was shown theoretically.

In these real data examples the proposed method provides better performance than the plug-in method. We recall that the clustering method is used in the classification step, whereas the plug-in method is a centroid-based classifier. Thus, the proposed method sometimes outperforms the plug-in method when the variances of two groups in the subspace of $\hat{\psi}_r$ are different.

6. Application: Identify players who are at risk for ACL injury. As a more practical application, we have used our method to identify handball players at risk for anterior cruciate ligament (ACL) injury based on ground reaction force data.

Subject. Twenty-two healthy female volunteers (age $= 20.4 \pm 1.3$ yr, mass $= 58.7 \pm 5.6$ kg, height $= 163 \pm 6.9$ cm), with no history of orthopedic lower limb injury prior to *six* months of the experiment, were recruited to this study. All the subjects were elite-level handball players belonging to the division-one category of western Japan. The local ethics board approved the procedure of this experiment, and the written informed consent was obtained from each subject before data collection.

Landing task. The experimental task was the single-legged drop landing task. Subjects were asked to make a forward jump from a wooden platform of 0.2 m in height and land on the force plate (Type9281B, Kistler, Switzerland) with the right leg. After landing, subjects were required to keep single-legged standing as quiet as possible for at least *eight* seconds. For all landings, subjects were instructed to put their arm crossed in front of their chest and not to release throughout the trial to eliminate the effect of arm swinging. The trials that the subject fell from the force plate or could not keep single-legged standing were regarded as unsuccessful and were discarded from further analysis.

Measurement of ground reaction force. The ground reaction force (GRF) signal from the force plate was amplified using the signal conditioner (Type 9865E1Y28, Kistler, Switzerland) and digitized with the data acquisition device (NI-USB6218 BNC, National Instruments, U.S.) with the sampling frequency of 1000 Hz.

Follow up of anterior cruciate ligament (ACL) injury. Within 24 months of follow-up period, seven cases of noncontact type ACL injuries were reported. Five cases among them occurred during handball practice or game with no physical contact from opponent players. For each case, the rupture of ACL is diagnosed by an orthopedic surgeon via arthroscopy. The mean number of days from experiment to injury was 360.1 ± 283 days. The shortest number of days were 43 days, and the longest was 470 days.

Identifying at-risk players using the proposed algorithm. Here, we apply our proposed algorithm to identify at-risk subjects based on the GRF data of the right leg landing task. As input data, we used the vertical component of GRF data during the first 200 ms after the initial foot impact. GRF data were normalized with the body mass for each subject to eliminate the effect of body mass. ACL injured subjects were, No. 2, 4, 7, 8, 15, 19 and 20. Subjects 7, 8 and 15 ruptured their right ACL during the follow-up period, but another sport caused the ACL injury of subject 7. Thus, GRF data of subjects 8 and 15 were labeled as positive, and the data of other subjects were considered as unlabeled data in the PU classification context. Figure 1 shows the GRF data of subjects 8 and 15 and Figure 2 shows randomly chosen 200 GRF curves of unlabeled data. For this positive and unlabeled GRF data, we applied the proposed PU classification method and the supervised classification method of Delaigle and Hall (2012). In both methods, we used empirically chosen $\hat{r} = \min_r \widehat{\text{err}}(r)$. Note that in the supervised classification based on Delaigle and Hall (2012), the GRF data of subjects 8 and 15 were considered positive, and others were forcibly considered negative In addition, we randomly chose 50 GRF curves as test data from all GRF data.

Result. Figure 10 and Figure 11 illustrate the supervised classification and the results of the proposed PU classification, respectively. In both figures the red markers denote the positive data (from subjects 8 and 15). Note that the black markers denote unlabeled data in Figure 11, whereas the black markers mean the negative data in Figure 10. In Figure 10 the blue markers denote the test data in the supervised classification. In both figures the vertical axes mean the projected data based on the estimated optimal $\hat{\psi}_r$, and horizontal axes show each subject.

From the classification results of test data in Figure 10, we cannot find any at-risk players except with subjects 8 and 15. That is, the usual supervised classification just identifies whether players are injured or not. Thus, the supervised classification is not appropriate to identify at-risk players.

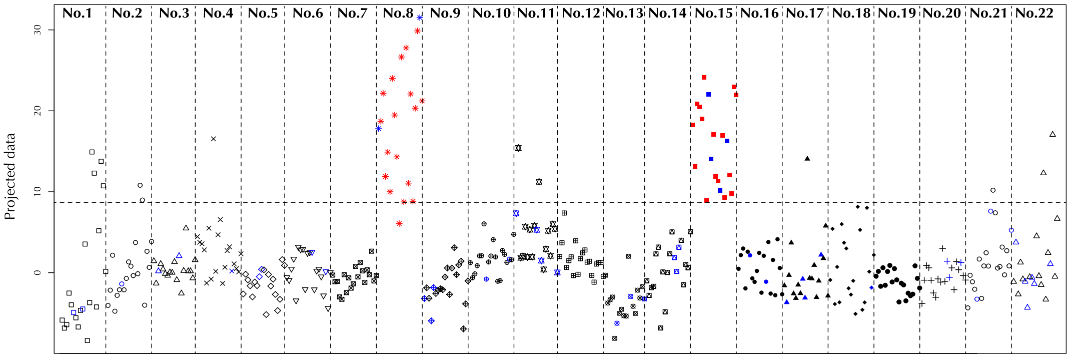


FIG. 10. *The result of the supervised classification based on Delaigle and Hall (2012) (Red markers: positive training data; Black markers: Negative training data; Blue markers: Test data). The horizontal dash line means the discrimination boundary.*

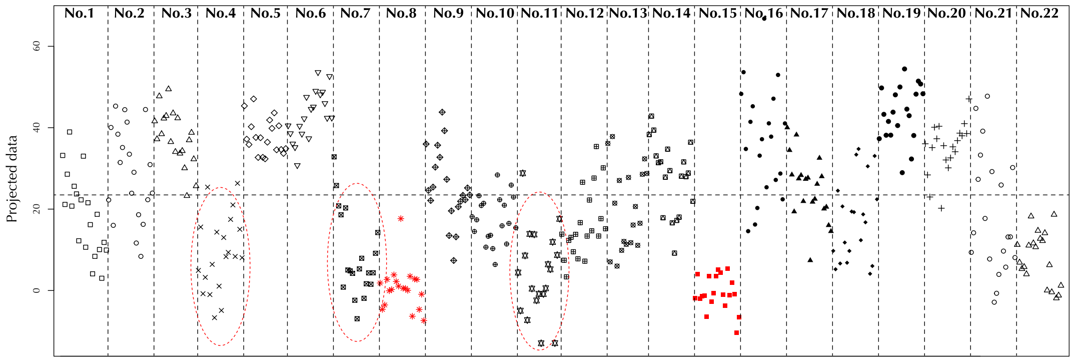


FIG. 11. *The result of the proposed PU classification algorithm* (*Red markers*: *positive data*; *Black markers*: *Unlabeled data*). *The horizontal dash line means the discrimination boundary.*

On the other hand, the result of the proposed method in Figure 11 suggests that subjects 4, 7, 11 and 22, who showed almost the same distribution as the positive data in the vertical axis, were regarded as high risk. Actually, subjects 4 and 7 suffered from the left and right ACL injuries, respectively, and, although not an ACL injury, subject 11 had a very severe ankle sprain after data collection. Subject 22 had no lower limb injury in the follow-up phase.

Noncontact type lower limb injury, such as ACL injury, is considered to occur due to dynamic postural control failure during sports play. Since the GRF signal during the landing task reflects the subject-specific postural strategy, it is suggested that our proposed algorithm successfully classified at-risk subjects through learning the signal component specific only to the injured subject which was buried in the GRF signal. Other injured subjects (No. 19 and No. 20) were distributed the opposite side of positive data, and, seemingly, they were safe. One of the possible reasons for this result is that subjects 19 and 20 ruptured their *left* ACL, whereas we used the GRF data of *right leg* landing task. Meanwhile, we could identify subject 4, who suffered from a left ACL injury, as an at-risk player in this result. Combining these facts, we can conclude that the risk factor of noncontact ACL injury, which reflective of the GRF signal, was not unique. It is wellknown in the clinical field that the mechanism of ACL injury is multifactorial; therefore, it is natural that there are differences in GRF components among the different individuals. It is suggested that our proposed algorithm feasibly distinguishes the different types of postural strategies, which both could lead to the ACL injury.

7. Conclusion.

In this paper we provide a new simple classification algorithm for the functional PU classification problem. The key feature of the proposed algorithm is that it does not require the estimation of the unknown class prior nor the constant probability that a positive object is labeled. In addition, it is worth noting that the idea of our method can be applied to kernel linear discriminant analysis for general data including multivariate data. Moreover, we demonstrated that, even in the PU classification problem for functional data, asymptotic perfect classification can often be achieved with the proposed method.

The main idea of the proposed method is to construct the discriminant subspace of FLDA with fully labeled data from only positive and unlabeled data. Thus, at least when the two covariances are the same, the performance of the proposed method cannot be better than the performance of FLDA with fully-labeled data. This is the first limitation of the proposed method. In this sense we might think that neither FLDA nor the proposed method performs well when the covariances of the two groups are not the same. However, as mentioned in Delaigle and Hall (2013), (functional) linear discriminant analysis often outperforms (functional) quadratic discriminant analysis in practice, unless the total sample size n is rather large (e.g., $n > 1000$) or the covariances of the two groups differ significantly. As with FLDA, the proposed method also can provide good performance in many practical situations.

On the other hand, we should mention that, when the sample size of the positive labeled data is too small (e.g., $n_{\mathrm{obs}} < 10$), the proposed method fails to construct an appropriate discriminant subspace. This is the second limitation of the proposed method. In fact, as shown in Setting 3 in the numerical experiments, when the sample size is not enough to detect the difference between the two groups, the performance of the proposed method is poor.

Overall, since FLDA (with fully labeled data) provides comparable performance with other classification methods including FQDA and nonparametric methods in practice (see Delaigle and Hall (2012) and Delaigle and Hall (2013)), we can expect that the proposed method with a moderate sample size of positive labeled data can provide good classification performance from only positive and unlabeled functional data. In fact, it worked well not only for numerical experiments but also for real chemometrics data and an important task in the sports medicine field. Since the proposed method can be easily implemented and has a low computational cost, we believe the proposed method is one of the possible choices for the PU classification of functional data.

Acknowledgments. The authors would like to thank the Editor and the anonymous reviewer for their helpful comments and suggestions which have significantly improved the paper. The authors also wish to express their thanks to Prof. Jeng-Min Chiou for his helpful comments. This work was supported by the Japan Society for the Promotion of Science (Y. Terada was supported by Grants 16K16024 and 20K19756; I. Ogasawara was supported by Grant 24700716).

SUPPLEMENTARY MATERIAL

Proofs and discussions (DOI: 10.1214/20-AOAS1404SUPP; .pdf). We provide technical details and additional discussions to the supplement (Terada, Ogasawara and Nakata (2020)).

REFERENCES

BLANCHARD, G., LEE, G. and SCOTT, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.* **11** 2973–3009. MR2746544

BOSQ, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. MR1783138 https://doi.org/10.1007/978-1-4612-1154-9

DELAIGLE, A. and HALL, P. (2012). Achieving near perfect classification for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 267–286. MR2899863 https://doi.org/10.1111/j.1467-9868.2011.01003.x

DELAIGLE, A. and HALL, P. (2013). Classification using censored functional data. *J. Amer. Statist. Assoc.* **108** 1269–1283. MR3174707 https://doi.org/10.1080/01621459.2013.824893

DU PLESSIS, M. C., NIU, G. and SUGIYAMA, M. (2014). Analysis of learning from positice and unlabeled data. In *Proceedings of Advances in Neural Information Processing Systems* 27 703–711.

DU PLESSIS, M. C., NIU, G. and SUGIYAMA, M. (2015). Convex formulation for learning from positive and unlabeled data. In *Proceedings of the* 32*nd International Conference on Machine Learning* 1386–1394.

DU PLESSIS, M. C. and SUGIYAMA, M. (2014). Class prior estimation from positive and unlabeled data. *IEICE Trans. Inf. Syst.* **E97-D** 1358–1362.

ELKAN, C. and NOTO, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 213–220.

FEBRERO-BANDE, M., DE LA FUENTE, M. O., GALEANO, P., NIETO, A. and GARCÍA-PORTUGUÉS, E. (2014). fda.usc: Functional data analysis and utilities for statistical computing.

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*. *Springer Series in Statistics*. Springer, New York. MR2229687

HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365 https://doi.org/10.1214/009053606000000272

HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. *Springer Series in Statistics*. Springer, New York. MR2920735 https://doi.org/10.1007/978-1-4614-3655-3

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis*, *with an Introduction to Linear Operators*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3379106 https://doi.org/10.1002/9781118762547

JAIN, S., WHITE, M., TROSSET, M. W. and RADIVOJAC, P. (2016). Nonparametric semi-supervised learning of class proportions. ArXiv.

KALIVAS, J. H. (1997). Two data sets of near infrared spectra. *Chemom*. *Intell*. *Lab*. *Syst*. **37** 255–259.

KAWANO, S. (2013). Semi-supervised logistic discrimination via labeled data and unlabeled data from different sampling distributions. *Stat*. *Anal*. *Data Min*. **6** 472–481. MR3150893 https://doi.org/10.1002/sam.11204

MENON, A. K., VAN ROOYEN, B., ONG, C. S. and WILLIAMSON, R. C. (2015). Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the* 32*nd International Conference on Machine Learning* 125–134.

RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*. *Springer Series in Statistics*. Springer, New York. MR1910407 https://doi.org/10.1007/b98886

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2168993

SCOTT, C. and BLANCHARD, G. (2009). Novelty detection: Unlabeled data definitely help. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* 464–471.

SHANG, H. L. and HYNDMAN, R. J. (2013). fds: Functional data sets.

TERADA, Y., OGASAWARA, I. and NAKATA, K. (2020). Supplement to "Classification from only positive and unlabeled functional data." https://doi.org/10.1214/20-AOAS1404SUPP

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu*. *Rev*. *Stat*. *Appl*. **3** 257–295.

ZHANG, X. and WANG, J.-L. (2016). From sparse to dense functional data and beyond. *Ann*. *Statist*. **44** 2281–2321. MR3546451 https://doi.org/10.1214/16-AOS1446