On statistical classification with incomplete covariates via filtering

Majid Mojirsheibani[1] and My-Nhi Nguyen[2]
Department of Mathematics, California State University Northridge, CA, 91330, USA[1]
Department of Preventive Medicine, University of Southern California, CA, 90032, USA[2]

Abstract

This article deals with the problem of classification when some of the covariates may have missing or unobservable parts. Here, it is allowed for both the training sample as well as the new unclassified observation to have missing parts in the covariates. In fact, it is shown in Remark 3 that in classification the reconstruction/imputation of the missing part of a new unclassified observation (which is to be classified) can be counter-productive in terms of the error rates. Furthermore, unlike many of the results in the literature, where covariate fragments are usually assumed to be missing completely at random, we do not impose any such assumptions here. Given the observed parts of the covariates, we construct a kernel-type classifier which is quite straightforward to implement in practice. The proposed classifier is constructed based on d-dim covariate vectors that are obtained from the original covariates (by moving from the space L^2 to ℓ_2), where d ($<\infty$) itself is a parameter that has to be estimated. To estimate various parameters, we employ an easy-to-implement data-splitting approach.

Keywords: Classification, kernel, incomplete covariates, asymptotics.

1 Introduction

The problem of statistical classification and pattern recognition based on covariate functions has received considerable attention in recent years; this is particularly true when the data are fully observable. In a standard two-group classification problem, this amounts to considering the random pair $(\boldsymbol{\chi}, Y)$, where the covariate curve $\boldsymbol{\chi}$ can take values in some metric space, $(\mathcal{M}, d)$, and $Y \in \{0,1\}$, called the class membership or class variable, has to be predicted based on $\boldsymbol{\chi}$. Here, one would like to find a classifier (a function) $g : \mathcal{M} \to \{0,1\}$ for which the misclassification error, $L(g) := \mathbb{P}\{g(\boldsymbol{\chi}) \neq Y\}$, is as small as possible. The optimal classifier, i.e., the classifier with the lowest misclassification error, is given by $g_{\mathrm{B}}(\chi) = 1$ if $\mathbb{P}\{Y = 1|\boldsymbol{\chi} = \chi\} > 1/2$, and $g_{\mathrm{B}}(\chi) = 0$ otherwise; see, for example, Cérou and Guyader [7], Abraham et al. [1], as well as the monograph by Devroye, et al. ([16]; Ch. 2). Although we have presented our setup for the popular binary case where $Y \in \{0,1\}$, our results in this paper can be generalized in a straightforward manner to the multi-group classification problem where $Y \in \{1, 2, \ldots, C\}$, for some positive integer $C \geq 2$.

[1] Email: majid.mojirsheibani@csun.edu
[2] Corresponding author. Email: ngocmynh@usc.edu

In practice the optimal classifier g_{B} is virtually always unknown and one only has access to a set of n independent and identically distributed (iid) data values $\mathbb{D}_n = \{(\boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{\chi}_n, Y_n)\}$ from the underlying distribution of $(\boldsymbol{\chi}, Y)$. The task is then to use the data $\mathbb{D}_n$ to construct a classification rule g_n that can predict the class membership, Y, of a new curve $\boldsymbol{\chi}$ with low error rates. A variety of techniques have been proposed for the classification of functional data in the literature. One may divide these techniques into roughly two types: (a) those approaches that use the whole covariate curve $\boldsymbol{\chi}$ to predict Y and (b) those that use the *filtered* curves to carry out classification; here, a filtered curve is a representation of a curve in the form of a vector. Relevant results corresponding to the approach used under (a) include the nonparametric functional approach of Ferraty and Vieu [18], the nearest neighbor method used by Cérou and Guyader [7], the kernel classifier of Abraham et al. [1], the depth-based classifier of López-Pintado and Romo [27], the robust functional classification of Cuevas et al. [9], the wavelet approach of Chang et al. [8], the robust functional classification of Alonso et al. [2], and the work of Meister [28] on the optimality properties of kernel regression and classification with functional covariates taking values in a general complete separable metric space.

On the other hand, relevant work under (b) includes the discrimination method of Hall et al. [23], the functional classification method of Biau et al. [4], the results of Leng and Müller [26] on the classification of gene expression data as well as that of Song et al. [35], the wavelet approach of Berlinet, et al. [3], the componentwise classification approach of Delaigle, et al. [14], the classification method in Delaigle and Hall [13], the *depth-depth* plot approach of Mosler and Mozharovskyi [32], the functional classification method of Dai and Müller [10], and the regularized linear classifiers of Kraus and Stefanucci [25]. Some other relevant results (but in the context of functional regression) include the work of Cai and Hall [6] on prediction in functional linear regression, the results of Hall and Horowitz [22] on the estimation of a slope function in functional linear regression, and those of Yao and Müller [36] on functional quadratic regression.

In this paper we employ methods that primarily fall under (b) above. More specifically, assuming that the functional covariates take values in a separable Hilbert space (and using the fact that such spaces are isomorphic to the space ℓ_2), the functional covariates will be replaced by d-dim vectors where $d \equiv d(n)$ is to be determined by the data; here, $d(n) \to \infty$, as $n \to \infty$. For the missing data framework, we follow the general setup proposed by Bugni [5]; this is described in Sections 2. In this paper, it is allowed for the covariate curves to have missing fragments in the training sample and in the new unclassified observation. This is quite different from most results in the literature where missing covariates only appear in the training sample. In fact, as argued in Remark 3, a rather peculiar consequence of this difference is that while the imputation of the missing fragments of a covariate curve $\boldsymbol{\chi}_i$ in the data, based on Y_i and the observed part of $\boldsymbol{\chi}_i$, can be helpful in classification, the imputation of the missing part of a new unclassified observation can actually be counter-productive in terms of the error rates. Furthermore, unlike most results in the literature, where covariate fragments are usually tacitly assumed to be missing completely at random, here we do not impose any such assumptions. In section 3 we propose a kernel classifier, under multiple missing patterns, and study its asymptotic properties; our main results are summarized in Theorem 2. Some numerical examples are also given; these appear in Section 4. All proofs are deferred to Section 5.

2 Covariates with missing parts and the setup

In standard classification with functional covariates, one typically assumes that each covariate function $\boldsymbol{\chi}(t)$ is a smooth curve on some compact domain $\mathcal{I} \subset \mathbb{R}$. Furthermore, the great majority of existing results assume that $\boldsymbol{\chi}(t)$ and $\boldsymbol{\chi}_i(t)$, $i = 1, \ldots, n$, do not have any missing or unobservable fragments over the domain $\mathcal{I}$. Here we allow $\boldsymbol{\chi}$ to be possibly missing (unobservable) on some subset(s) of its domain, i.e., the situation where one may only be able to observe certain parts of the full curve $\boldsymbol{\chi}$. More specifically, let $(\Omega, \mathcal{A}, \mathbb{P})$ be the underlying probability space and let $\mathcal{M}$ be the space of square-integrable functions $L^2(\mathcal{I})$, where $\mathcal{I}$ is an interval on the real line. Therefore, $\boldsymbol{\chi}$ is a random function on $(\Omega, \mathcal{A}, \mathbb{P})$ with values (i.e., with sample paths) in $L^2(\mathcal{I})$. But, instead of observing the full curve $\boldsymbol{\chi} : \Omega \to L^2(\mathcal{I})$, one might only be able to observe certain segments of the curve denoted by $\boldsymbol{\chi}|_s$, i.e., the restriction of the curve $\boldsymbol{\chi}(t)$ to $t \in s \subset \mathcal{I}$.

The problem of functional classification with possibly incomplete covariates has received some attention in the literature in recent years. These include the work of Delaigle and Hall [12] who consider a quadratic discriminant classifier for censored functional data based on the observed fragments of covariates with overlapping domains that are not *too short*. Zhou et al. [37] propose a *wrapping* function to predict/estimate the residual life distribution based on partially observed signal data. Kraus [24] proposes methods to estimate parameters and to carry out principal component analysis. Delaigle and Hall [11] proposed a method based on Markov chains to reconstruct the missing parts of the curve, which are used in linear prediction for functional data. Gromenko et al. [19] consider the problem of functional regression with incomplete curves. Mojirsheibani and Shaw [29] studied the problem of classification with incomplete covariate curves taking values in $L^1([a,b])$ (instead of $L^2([a,b])$), whereas Kraus and Stefanucci [25] consider regularized linear classifiers. Most of the above results assume that the covariates are *missing completely at random.* In this paper we do not impose such assumptions. Furthermore, unlike the above papers where missing covariates typically appear in the data but not in the new unclassified observation, here the missing values can appear anywhere. Our remarks below (as well as Remark 3) make this distinction very clear.

Unfortunately, the problem of classification can be substantially different and more complicated when incomplete covariates can also appear in the new unclassified observation. To appreciate this, one can consider the simple case based on the Euclidean covariate vector $\mathbf{X} = (\mathbf{Z}, \mathbf{V}) \in \mathbb{R}^{d+p}$, $d, p \geq 1$, where $\mathbf{V} \in \mathbb{R}^p$ may be missing but not $\mathbf{Z}$. Let $Y \in \{0, 1\}$ be the class variable to be predicted, and define the Bernoulli random variable $\delta = 0$ if $\mathbf{V}$ is missing (and $\delta = 1$ otherwise). Then, as shown by Mojirsheibani and Montazeri [31], and further studied by Mojirsheibani [30], and Demirdjian and Mojirsheibani [15], the theoretically optimal classifier in this case can be expressed as (see, for example, Mojirsheibani ([30], eq. (12)):

$$\textit{Assign a new observation to class 1} \quad \textit{if} \quad \delta\,\frac{\mathbb{E}(\delta Y|\mathbf{X})}{\mathbb{E}(\delta|\mathbf{X})} + (1-\delta)\,\frac{\mathbb{E}[(1-\delta)Y|\mathbf{Z}]}{\mathbb{E}[(1-\delta)|\mathbf{Z}]} > \frac{1}{2},$$

(otherwise, assign it to class 0), with the convention 0/0=0. This classifier is very different from the usual optimal classifier that assigns a new observation to class 1 if $\mathbb{E}[Y|\mathbf{X}] > \frac{1}{2}$. *Furthermore, it turns out that any attempt to reconstruct the missing part of a new unclassified observation can be counter-productive in the sense that it can increase the theoretical misclassification error; see Remark 3 for more on this.*

To set up our framework for possible missing patterns in the curve $\boldsymbol{\chi}$, we follow Bugni [5]. In Bugni's [5] setup, it is assumed that for a fine enough partition of $\mathcal{I}$ into $J < \infty$ subintervals $\mathcal{I}_1, \ldots, \mathcal{I}_J$, each sample function of $\boldsymbol{\chi}$ is either completely observed or completely unobserved within each of these J subintervals. Some examples of such functional variables can be found in [5]. In the rest of this paper we assume that there are $M < 2^J$ possible missing patterns in the data where M is usually much smaller than 2^J. Therefore, under the k-th pattern, one observes the fragment $\boldsymbol{\chi}|_{s_k}$, $k = 1, \ldots, M$. Next, let δ be the $\{1, \ldots, M\}$-valued random variable defined as

$$\delta = k \quad \text{if pattern } k \text{ (i.e., the fragment } \boldsymbol{\chi}|_{s_k}\text{) is observed}, \quad k = 1, \ldots, M.$$

In passing, we briefly recall that since $\boldsymbol{\chi} \in L^2(\mathcal{I})$, i.e., a separable Hilbert space, it can be expressed by the expansion $\boldsymbol{\chi}(t) = \sum_{j=1}^{\infty} X_j \psi_j(t)$, where $\{\psi_1, \psi_2, ...\}$ is a complete orthonormal basis for $L^2(\mathcal{I})$ and $X_j = \langle \boldsymbol{\chi}, \psi_j \rangle := \int_{\mathcal{I}} \boldsymbol{\chi}(t) \psi_j(t) dt$. Here the infinite sum converges in L^2. Similarly, given the data $(\boldsymbol{\chi}_i, Y_i)$, $i = 1, ..., n$, we can write $\boldsymbol{\chi}_i(t) = \sum_{j=1}^{\infty} X_{ij} \psi_j(t)$, with $X_{ij} = \int_{\mathcal{I}} \boldsymbol{\chi}_i(t) \psi_j(t) dt$. Since any infinite-dimensional separable Hilbert space is isomorphic to the space $\ell_2 = \{\mathbf{x} = (x_1, x_2, \ldots) \big| \sum_{i=1}^{\infty} |x_i|^2 < \infty\}$, the *scores* X_{ij}, $j \geq 1$, are used as surrogates for the datum $\boldsymbol{\chi}_i$ in the literature in the sense that knowing $\mathbf{X}_i := (X_{i1}, X_{i2}, \ldots)$ is the same as knowing $\boldsymbol{\chi}_i$; see, for example, Hall et al [23] or Biau et al [4]. Of course, in most practical situations, one usually observes discretized versions of the curves (and not the true curves themselves). In such cases, all integrals may be approximated as weighted averages over the grid of points at which the curves are observed. In fact, this is the approach we have adopted in our numerical studies of Section 4.

To simplify our presentation, we first look at the hypothetical (oversimplified) case where there is only one missing pattern. More specifically, write $\mathcal{I} = [a,b] = [a,c] \cup (c,b]$, for some $a < c < b$, where $\boldsymbol{\chi}(t)$ may be missing on $(c,b]$ only. Therefore, we have the expansions

$$\begin{aligned}\boldsymbol{\chi}(t) &= \sum_{j=1}^{\infty} \langle \boldsymbol{\chi}, \psi_j \rangle_{[a,b]} \psi_j(t) = \sum_{j=1}^{\infty} \left[\int_a^c \boldsymbol{\chi}(t)\psi_j(t)dt + \int_c^b \boldsymbol{\chi}(t)\psi_j(t)dt \right] \psi_j(t) \\ &= \sum_{j=1}^{\infty} \left(\langle \boldsymbol{\chi}, \psi_j \rangle_{[a,c]} + \langle \boldsymbol{\chi}, \psi_j \rangle_{[c,b]} \right) \psi_j(t).\end{aligned}$$

Now the surrogate vector of score functions can be written as

$$\begin{aligned}\mathbf{X} &= (X_1, X_2, \dots) := \big(\langle \boldsymbol{\chi}, \psi_1 \rangle_{[a,\,b]}\,,\ \langle \boldsymbol{\chi}, \psi_2 \rangle_{[a,\,b]}\,, \dots \big) \\ &= \big(\langle \boldsymbol{\chi}, \psi_1 \rangle_{[a,\,c]}\,,\ \langle \boldsymbol{\chi}, \psi_2 \rangle_{[a,\,c]}\,, \dots \big) + \big(\langle \boldsymbol{\chi}, \psi_1 \rangle_{[c,\,b]}\,,\ \langle \boldsymbol{\chi}, \psi_2 \rangle_{[c,\,b]}\,, \dots \big) \\ &=: (Z_1, Z_2, \dots) + (V_1, V_2, \dots) \\ &=: \mathbf{Z} + \mathbf{V},\end{aligned}$$

where $\mathbf{V}$ may be missing, but not $\mathbf{Z}$. Here, we note that if $\mathbf{V}$ is not missing then $\mathbf{X}$ ($= \mathbf{Z} + \mathbf{V}$) is fully observable, otherwise our classification has to be carried out be based on $\mathbf{Z}$ only. For the more general setting with M missing patterns, if we let $X_j^{(k)} = \langle \boldsymbol{\chi}, \psi_j \rangle_{s_k}$, then, with $s_1 := \mathcal{I}$, we have the M vectors of scores

$$\mathbf{X}^{(k)} = (X_1^{(k)}, X_2^{(k)}, \dots) = \big(\langle \boldsymbol{\chi}, \psi_1 \rangle_{s_k}, \langle \boldsymbol{\chi}, \psi_2 \rangle_{s_k}, \dots\dots \big), \qquad k = 1, \dots, M. \tag{1}$$

Clearly, when $\delta = k$, we only observe $\mathbf{X}^{(k)}$ in which case a classifier is any function of the form $g_k : \ell_2 \to \{0,1\}$. Hence, any classifier can be written in the general form

$$\Gamma(\mathbf{X}^{(\delta)}) := \sum_{k=1}^{M} I\{\delta = k\} \cdot g_k(\mathbf{X}^{(k)}), \qquad \text{where } \ \mathbf{X}^{(\delta)} := \textstyle\sum_{k=1}^{M} \mathbf{X}^{(k)} I\{\delta = k\}. \tag{2}$$

Suppose that g_k is the theoretically optimal classifier for the k^{th} pattern, i.e., $g_k(\mathbf{X}^{(k)}) = 1$ if $\mathbb{E}[Y|\mathbf{X}^{(k)}] > \frac{1}{2}$, (and $g_k(\mathbf{X}^{(k)})$=0, otherwise). Then, with this choice of g_k, one may be inclined to consider $\Gamma(\mathbf{X}^{(\delta)})$ in (2) to be the optimal classifier for the current setup with incomplete covariates; however, this turns out to be incorrect, in general. As for the optimal classifier, let

$$\phi_k(\mathbf{X}^{(k)}) = \mathbb{E}\left[(2Y-1) I\{\delta = k\} \,\middle|\, \mathbf{X}^{(k)} \right], \ k = 1, \dots, M, \tag{3}$$

and define the following classifier

$$\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) = \sum_{k=1}^{M} I\{\delta = k\} \cdot I\{\phi_k(\mathbf{X}^k) > 0\}. \tag{4}$$

This amounts to choosing g_k in (2) to be $g_k(\mathbf{X}^{(k)}) = I\{\phi_k(\mathbf{X}^k) > 0\}$. Then, part (i) of the following result shows that the classifier in (4) is optimal.

Theorem 1 *Let Γ^B be the classifier given in (4).*

(i) *The classifier Γ^B has the lowest misclassification error, i.e., for any other classifier Γ, one has $\mathbb{P}\{\Gamma(\mathbf{X}^{(\delta)}) \neq Y\} - \mathbb{P}\{\Gamma^B(\mathbf{X}^{(\delta)}) \neq Y\} \geq 0$.*

(ii) *Let Γ be any classifier of the form $\Gamma(\mathbf{X}^{(\delta)}) = \sum_{k=1}^{M} I\{\delta = k\} \cdot I\{\varphi_k(\mathbf{X}^{(k)}) > 0\}$ for arbitrary functions $\varphi_k : \ell_2 \to [-1, 1]$, $k = 1, \ldots, M$. Then, with $\phi_k(\mathbf{X}^{(k)})$ is as in (3), one has*

$$\mathbb{P}\big\{\Gamma(\mathbf{X}^{(\delta)}) \neq Y\big\} - \mathbb{P}\big\{\Gamma^B(\mathbf{X}^{(\delta)}) \neq Y\big\} \leq \textstyle\sum_{k=1}^{M} \mathbb{E}\Big|\phi_k(\mathbf{X}^{(k)}) - \varphi_k(\mathbf{X}^{(k)})\Big|.$$

The proof of this theorem is given at the end of Section 5.

In passing, we note that part (ii) of Theorem 1 provides a useful tool to bound the difference between the misclassification error of $\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)})$ and that of any other classifier $\Gamma(\mathbf{X}^{(\delta)})$ in terms of the difference between $\phi_k(\mathbf{X}^{(k)})$ that appears in (3) and the function $\varphi_k(\mathbf{X}^{(k)})$. Here, one can think of $\varphi_k(\mathbf{X}^{(k)})$ as an approximation to the unknown function $\phi_k(\mathbf{X}^{(k)}) = \mathbb{E}\left[(2Y-1)I\{\delta = k\} \,\middle|\, \mathbf{X}^{(k)}\right]$.

3 The proposed classifier

Here we begin by considering finite-dimensional versions (with increasing dimensions) of the classifier Γ^{B} defined in (4), where $\mathbf{X}^{(k)}$ $(\in \ell_2)$ will be replaced by the d-dimensional vector $\mathbf{X}^{(d,k)} = (X_1^{(k)}, \ldots, X_d^{(k)}) = \big(\langle \chi, \psi_1 \rangle_{s_k}, \ldots, \langle \chi, \psi_d \rangle_{s_k}\big)$, $k = 1, \ldots, M$. A data-driven choice of the parameter d is discussed later in this section. We first start by defining the function $\phi_{d,k} : \mathbb{R}^d \to [-1, 1]$ according to

$$\phi_{d,k}(\mathbf{X}^{(d,k)}) := \mathbb{E}\left[(2Y-1)I\{\delta = k\} \,\middle|\, \mathbf{X}^{(d,k)}\right] = \mathbb{E}\left[(2Y-1)I\{\delta = k\} \,\middle|\, X_1^{(k)}, \ldots, X_d^{(k)}\right], \tag{5}$$

$k = 1, \ldots, M$. Now, consider the following version of the classifier in (4)

$$\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)}) = \sum_{k=1}^{M} I\{\delta = k\} \cdot I\left\{\phi_{d,k}(\mathbf{X}^{(d,k)}) > 0\right\}. \tag{6}$$

Here, $\mathbf{X}^{(d,\delta)} = \sum_{k=1}^{M} \mathbf{X}^{(d,k)} \cdot I\{\delta = k\}$. The fact that all distributions are unknown implies that the classifier in (6) is not available in practice and has to be constructed based on the available data. Here we propose a kernel-type methodology. To construct our classifier, we also employ the following data-splitting approach which is in the spirit of the method proposed by Biau et al [4] to deal with functional nearest neighbor classification (without any missing data). Let $\mathbf{X}^{(\delta)}$ be as in (2) and start by randomly splitting the data $\mathbb{D}_n = \{(\mathbf{X}_1^{(\delta_1)}, Y_1, \delta_1), \ldots, (\mathbf{X}_n^{(\delta_n)}, Y_n, \delta_n)\}$ into a *training* sample $\mathbb{D}_m$ of size m and a *testing sequence* $\mathbb{D}_\ell$ of size $\ell = n - m$. Here, m and ℓ typically depend on n (they grow with n). Next, for each fixed integer $d \geq 1$, put

$$\widehat{\phi}_{m,d,\,h_k}(\mathbf{X}^{(d,k)}) = \sum_{i:\ (\mathbf{X}^{(\delta_i)}, Y_i, \delta_i) \in \mathbb{D}_m} (2Y_i - 1) I\{\delta_i = k\} \cdot \mathcal{K}_k \left(\frac{\mathbf{X}^{(d,k)} - \mathbf{X}_i^{(d,k)}}{h_k} \right), \tag{7}$$

where $\mathcal{K}_k : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel used with the smoothing parameter h_k, and define the following sample-based counterpart of (6), which is based on $\mathbb{D}_m$ only,

$$\Gamma_m^d(\mathbf{X}^{(d,\delta)}) = \sum_{k=1}^{M} I\{\delta = k\} I\left\{\widehat{\phi}_{m,d,h_k}(\mathbf{X}^{(d,k)}) > 0\right\}, \tag{8}$$

where d and h_k, $k = 1, \ldots, M$ are the free parameters to be estimated. Here, $\mathbf{X}^{(d,k)}$ and $\mathbf{X}_i^{(d,k)}$ represent the first d components of $\mathbf{X}^{(k)}$ and $\mathbf{X}_i^{(k)}$, respectively. Let $\mathcal{H} \equiv \mathcal{H}_n$ be a grid of positive values from which $h_1, \ldots h_M$ are to be selected, and define $\widehat{d}$ and $\widehat{h}_k$ to be the empirically chosen values of d and h_k, $k = 1, \ldots, M$ that minimize the empirical error committed by the classifier (8) on the testing sequence $\mathbb{D}_\ell$, i.e.,

$$(\widehat{d}, \widehat{h}_1, \ldots, \widehat{h}_M) = \underset{1 \leq d \leq d_n,\, h_k \in \mathcal{H}_n,\, k=1,\ldots,M}{\operatorname{argmin}} \ \ell^{-1} \sum_{i:\ (\mathbf{X}^{(\delta_i)}, Y_i, \delta_i) \in \mathbb{D}_\ell} I\{\Omega_i(m, d, h_1, \ldots, h_M)\}, \tag{9}$$

where the set Ω_i is given by

$$\Omega_i(m, d, h_1, \ldots, h_M) = \left\{ \sum_{k=1}^{M} I\{\delta_i = k\} \cdot I\left\{\widehat{\phi}_{m,d,h_k}(\mathbf{X}_i^{(d,k)}) > 0\right\} \neq Y_i \right\}, \tag{10}$$

and where d_n in (9) diverges with n, but not too rapidly; see Remark 1. In passing we also note that in our estimation steps above, no part of the data is discarded. Our final classifier is the plug-in version of (8) given by

$$\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)}) := \sum_{k=1}^{M} I\{\delta = k\} I\left\{\widehat{\phi}_{n,\widehat{d},\widehat{h}_k}(\mathbf{X}^{(\widehat{d},k)}) > 0\right\}, \tag{11}$$

where the subscript n used in the definition of $\widehat{\Gamma}_n$ in (11) indicates that it is constructed based on the entire data of size n. How good is the classifier $\widehat{\Gamma}_n$ in (11)? The next theorem shows that under rather standard assumptions, $\widehat{\Gamma}_n$ is strongly optimal. To state this formally, we first state the following assumption on the kernels used in (7).

Assumption (K).
The kernel $\mathcal{K}_k$ used in (7) is *regular:* A nonnegative kernel $\mathcal{K}$ is said to be regular if there are positive constants $b > 0$ and $r > 0$ for which $\mathcal{K}(\mathbf{x}) \geq bI\{\mathbf{x} \in S_{0,r}\}$ and $\int \sup_{\mathbf{y}\in\mathbf{x}+S_{0,r}} \mathcal{K}(\mathbf{y})d\mathbf{x} < \infty$, where $S_{0,r}$ is the ball of radius r centered at the origin. (For more on regular kernels see, for example, Györfi et al [20].)

Theorem 2 *[The Main Result.] Suppose that Assumption (K) holds. Also assume that, as $n \to \infty$, we have $\ell \equiv \ell(n) \to \infty$, $m \equiv m(n) \to \infty$, $\ell^{-1} \log |\mathcal{H}_n| \to 0$, and $\ell^{-1} \log d_n \to 0$, where $|\mathcal{H}_n|$ is the cardinality of the set $\mathcal{H}_n$. Suppose that for each $k = 1, \ldots, M$, there is an $h_k \equiv h_k(n) \in \mathcal{H}_n$ such that $\max_{1\leq k\leq M} h_k \to 0$ and $\min_{1\leq k\leq M} m h_k^{d_n} \to \infty$, as $n \to \infty$. Then the classifier $\widehat{\Gamma}_n$ in (11) is asymptotically strongly optimal, i.e.,*

$$\mathbb{P}\left\{\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)}) \neq Y \,\middle|\, \mathbb{D}_n\right\} \longrightarrow^{\text{a.s.}} \mathbb{P}\left\{\Gamma^B(\mathbf{X}^{(\delta)}) \neq Y\right\},$$

as $n \to \infty$, where Γ^B is the theoretically optimal classifier appearing in Theorem 1.

Remark 1 *[Magnitude of the smoothing parameters.]* The conditions imposed on $h_k \equiv h_k(n)$ and d_n in the statement of Theorem 2 are satisfied if d_n does not grow too rapidly and, additionally, h_k converges to zero slowly, as $n \to \infty$. In fact, if we take $d_n = (\log n^{c_0})^{1-\gamma}$ for any $c_0 > 0$ and any $0 < \gamma < 1$, and if, for example, $h_k = (\log n^{c_k})^{-1}$ for any $c_k > 0$, then it is straightforward to see that $m h_k^{d_n} \to \infty$, as $n \to \infty$. Intuitively, the slow rate of convergence (logarithmic) of h_k to zero is not necessarily unrealistic here and, in a sense, can be tied to the increasing dimension d_n. In fact, in what Ferraty and Vieu ([17], p. 211) refer to as the *curse of infinite dimensionality*, the authors argue that in the problem of kernel regression estimation for the general regression function $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ with a functional covariate $\mathbf{X}$, the smoothing parameter $h \equiv h(n)$ can be of order $(\log n)^u$ for some $u < 0$.

Remark 2 *[Number of missing patterns in practice.]* As explained in Section 2, our framework for missing patterns is the same as Bugni's [5], where it is assumed that for a fine enough partition of the domain of $\boldsymbol{\chi}$ into $J < \infty$ subintervals, each $\boldsymbol{\chi}$ is either fully observed or completely unobserved within each subinterval. This framework has also been used by Kraus ([24], p. 781) to estimate various parameters for functional data. In practice, unless n is quite large, it is tacitly assumed that the actual number of missing patterns, M, is much smaller than the 2^J possible missing patterns; this would ensure that there will be enough data to estimate various parameters. Therefore, the main focus of this article (and many other papers on missing covariates) is to deal with those missing patterns that give rise to most of the missing values in the data. Such difficulties and hurdles are not confined to functional covariates and can also plague the problem of classification for the simpler case of $\mathbf{X} \in \mathbb{R}^d$, where there could be as high as $2^d - 1$ possible missing patterns.

See, for example, Mojirsheibani and Montazeri [31]. It should also be mentioned that, unlike the results of Delaigle and Hall [12] and Kraus [24], we do not require any moment assumptions in this paper.

Remark 3 *[Drawbacks of imputing/reconstructing missing parts of the covariate of an unclassified observation in classification.]* One particular feature of this paper is that it also allows for missing covariates to appear in the new unclassified observation (and not just the training sample). In its simplest form, if $\mathbf{V}$ is missing from the covariate vector $\mathbf{X} = (\mathbf{Z}, \mathbf{V})$, then it may be suggested that one should first approximate/reconstruct the missing $\mathbf{V}$ by some $\mathbf{V}^*$ and then apply the theoretically optimal classifier to $\mathbf{X}^* = (\mathbf{Z}, \mathbf{V}^*)$ for predicting the class variable $Y \in \{0, 1\}$. As in the previous sections, the covariate $\mathbf{X} = (\mathbf{Z}, \mathbf{V})$ could be either functional or Euclidean. Here, $\mathbf{V}^*$, which is typically a function of $\mathbf{X}$ (or a function of $\mathbf{X}$ and the data $\mathbb{D}_n$), is called the *imputed* value of $\mathbf{V}$. Unfortunately, in classification, imputation does not work well for new unclassified observations and may, in fact, be counter-productive. To appreciate this, consider the popular method of regression imputation, where the missing $\mathbf{V}$ will be replaced by the estimates of $\mathbb{E}(\mathbf{V}|\mathbf{Z})$. To simplify our example, we further assume that the regression function $r(\mathbf{z}) := \mathbb{E}(\mathbf{V}|\mathbf{Z} = \mathbf{z})$ is completely known (thus there is no need to estimate it). Therefore, replacing the missing $\mathbf{V}$ with its imputed value $\mathbf{V}^* := r(\mathbf{Z})$, the optimal classifier is

$$g_{\mathrm{B}}(\mathbf{X}^*) = \begin{cases} 1 & \text{if } \mathbb{P}\{Y = 1|\mathbf{Z}, \mathbf{V}^*\} > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \quad \text{where } \ \mathbf{X}^* = (\mathbf{Z}, \mathbf{V}^*) = (\mathbf{Z}, r(\mathbf{Z})). \tag{12}$$

It turns out that g_{B}, given by (12), is not even as good as the classifier that ignores $\mathbf{V}$ completely and classifies Y based on $\mathbf{Z}$ alone. More specifically, if we let $\tilde{g}_{\mathrm{B}}$ be the optimal classifier based on $\mathbf{Z}$ only, i.e.,

$$\tilde{g}_{\mathrm{B}}(\mathbf{Z}) = \begin{cases} 1 & \text{if } \mathbb{P}\{Y = 1|\mathbf{Z}\} > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

then (by Theorem 3.3 of Devroye et al. [16]) one finds $\mathbb{P}\{\tilde{g}_{\mathrm{B}}(\mathbf{Z}) \neq Y\} \leq \mathbb{P}\{g_{\mathrm{B}}(\mathbf{X}^*) \neq Y\}$. That is, the theoretically optimal classifier g_{B} in (12), which uses both $\mathbf{Z}$ and $\mathbf{V}^* = r(\mathbf{Z})$ to predict Y, can actually perform worse than the classifier $\tilde{g}_{\mathrm{B}}$ that ignores $\mathbf{V}$ (and uses $\mathbf{Z}$ only). In fact, this conclusion holds true for any $\mathbf{V}^*$ which is a function of $\mathbf{Z}$ (and not just the regression imputation $\mathbf{V}^* = E(\mathbf{V}|\mathbf{X})$). Of course, if missing values appear in the data, then proper regression imputation is available since for each data point $(\mathbf{X}_i, Y_i)$, where part of $\mathbf{X}_i$ may be missing, the variable Y_i is always available. Clearly imputation can be beneficial in such cases.

Remark 4 *[The number of parameters.]* The proposed classifier in (11) involves the estimation of $M + 1$ parameters: $d, h_1, \ldots, h_{\mathrm{M}}$. Here, it is assumed that M is not too large compared to the sample size n. When n is small or M is large, one can simply consider one common bandwidth h, and it is not hard to see that the conclusion of Theorem 2 continues to hold. Of course, in finite samples, the resulting classifier can have slightly higher error rates when a common bandwidth h is used. On the other hand, when n is very much larger than M, one has the luxury of considering $d_1, \ldots, d_{\mathrm{M}}$ instead of a common d. Once again, as in Theorem 2, one can show that the resulting classifier is strongly consistent. From an applied point of view, and as in many results in statistics, the question of how large is large (in terms of n) could be difficult to quantify and can vary from one situation to another. Our limited experience shows that the gain from the inclusion of additional parameters may not be worth the extra computational burden needed to estimate all parameters.

4 Numerical examples

4.1 Simulated Data

Here, we provide some numerical examples to assess the performance of the methods proposed in the previous section. In this analysis, we develop classifiers to predict the unknown class $Y = 0$ or $Y = 1$ of a functional covariate $\boldsymbol{\chi}(t)$, taking values in $L^2([0, 1])$, that may have missing fragments. Adopting the missing pattern setup of Section 2, without loss of generality let $s_1 := \mathcal{I} = [0, 1]$. Also, let $s_2 = [0, 0.3] \cup [0.5, 1] \subset \mathcal{I}$, $s_3 = [0, 0.1] \cup [0.2, 0.45] \cup [0.6, 0.85] \cup [0.9, 1] \subset \mathcal{I}$, $s_4 = [0.25, 0.5] \cup [0.65, 1] \subset \mathcal{I}$, and $s_5 = [0, 0.2] \cup [0.3, 0.55] \cup [0.75, 0.9] \subset \mathcal{I}$. We consider two cases of missing patterns: $M = 3$ and $M = 5$. In the case of $M = 3$, the patterns used are s_1, s_2, and s_3. Next, samples of functional observations $\left(\boldsymbol{\chi}_i^{(\delta_i)}, Y_i, \delta_i\right)$, $i = 1, \ldots, n$, are generated based on rules which are similar to the approach of Rachdi and Vieu [33] as follows:

$$\boldsymbol{\chi}_i(t) = (t - 0.5)^2 \xi_i + \zeta_i, \quad i = 1, 2, \ldots, n$$

where $t \in s = s_1, s_2, s_3, s_4$, or s_5 depending on whether $\delta_i = 1, 2, 3, 4$ or 5. Since in practice one typically observes discretized versions of the curves (instead of the curves themselves), all covariates were generated in a dicretized form based on a grid of 500 equispaced values of t in $[0, 1]$. Regarding the independent random variables ξ_i and ζ_i, if $Y_i = 1$ then $\xi_i \overset{\text{iid}}{\sim} N(5, 2^2)$ and $\zeta_i \overset{\text{iid}}{\sim} N(1, 0.5^2)$, otherwise if $Y_i = 0$ then $\xi_i \overset{\text{iid}}{\sim} \text{Unif}\,(0, 5)$ and $\zeta_i \overset{\text{iid}}{\sim} \text{Unif}\,(0, 1)$. The class probabilities are taken to be $P(Y = 1) = 0.5 = P(Y = 0)$. With respect to the missing probability mechanism, we consider the popular logistic-type model

$$P_s\{\delta = 1 \big| Y = y, \boldsymbol{\chi} = \chi\} = \lambda(s, y, \chi)/[1 + \lambda(s, y, \chi)], \tag{13}$$

where
$$\lambda(s, y, \chi) = \exp\Big\{a(1-y) + b\int_s \chi(u)\, du + c\int_{[0,1]\setminus s} u \cdot \chi(u)\, du\Big\}, \tag{14}$$

and the set s can be selected to be any one of the patterns s_k, $k = 2, \ldots, M$, with probability $1/(M - 1)$. The coefficients a, b, and c in the above expression can be adjusted to control the missing data rate. They can also be adjusted to control the level of dependency of the missing probability in (13) on Y and on the observed and unobserved segments of the curve. As for the choice of the basis functions, we used the Fourier basis $\{\psi_1(t) = 1, \psi_{2k}(t) = \sqrt{2}\cos(\pi kt), \psi_{2k+1}(t) = \sqrt{2}\sin(2\pi kt), k \geq 1\}$ which forms a complete orthonormal basis for $L^2([0,1])$; see, for example, Zygmund [38] or Sansone [34]. Figure 1 shows a few realizations of the simulated curves $\boldsymbol{\chi}|_{s_k}$.

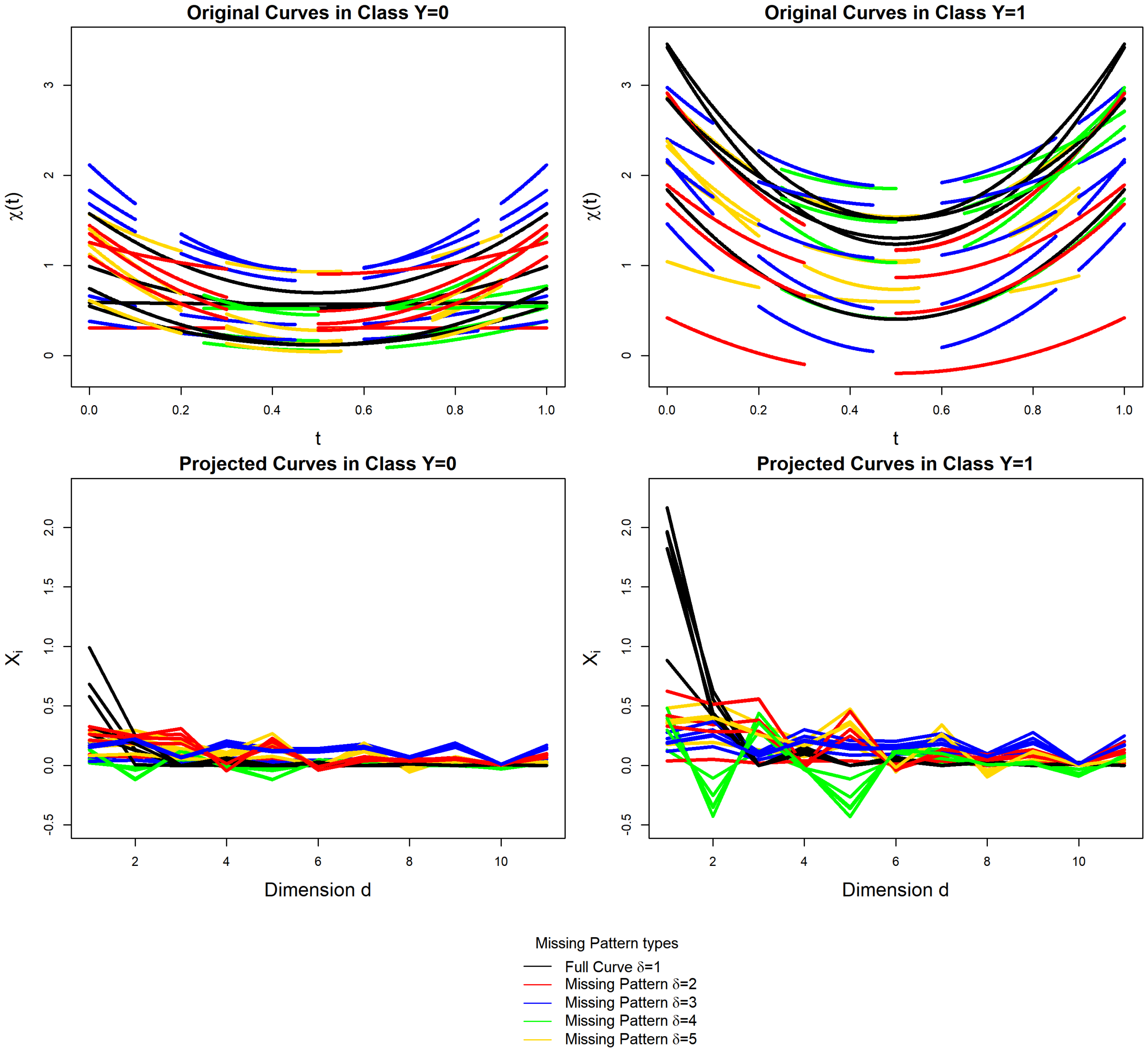


Figure 1: A sample of simulated curves with their projected vectors for a few values of d. Here, 30% of the data contain some unobserved fragments

Next, we constructed the proposed classifier $\widehat{\Gamma}_n$, given by (11), based on two different sample sizes, $n = 100$ and 200, as well as several choices for the constants a, b, and c for each of the missing patterns. The parameters h_k and d were selected from a grid of equally spaced values of h in $[0, 1]$ and $1 \leq d \leq d_n$, based on the procedure in (8) and (9), with Gaussian kernels, and with a random split of the data into $\mathbb{D}_m$ and $\mathbb{D}_\ell$ of sizes $m = 0.65n$ and $\ell = n - m$. Here, we took $d_n \approx 2.5 \ln(n)$; see Remark 1 for details and the justification for the choice of d_n. This process was repeated for 20 such random sample splits and the values of h_k and d that minimized the average error were selected; these are denoted by $\widehat{h}_k$ and $\widehat{d}$ which appear in (9). In addition to the proposed classifier $\widehat{\Gamma}_n$, we also constructed the classifier based on the complete case analysis, which will be denoted by $\widehat{\Gamma}_{CC}$, (this classifier uses the complete cases only), as well as the classifier corresponding to the case with no missing data (i.e., when all covariates are fully observable), to be denoted by $\widetilde{\Gamma}_n$, which was proposed by Biau et al. [4]. Furthermore, in our analysis here, we have considered different missingness mechanisms such as the "Not Missing At Random" (NMAR), the "Missing At Random" (MAR), and the "Missing Completely At Random" (MCAR) scenarios. These classifiers are then used to classifying 1000 additional observations from the same underlying distribution of the data. The entire above process was repeated a total of 100 times and the average misclassification errors (over 100 Monte Carlo runs) were computed. Our findings are summarized in Table 1 and Table 2 with the percentage of missing data of 30% and 80% accordingly. The constants a, b, c (of equation (14)) corresponding to pattern s_2 are reported in columns a_2, b_2, c_2 of the tables, those corresponding to s_3 are reported in columns a_3, b_3, c_3, and so on. A total of 50 cases can be identified corresponding to our classifiers at different settings (different sample sizes, different missing rates, different values of a, b, c, etc.); these cases are labeled as C1, C2, C3, $\ldots$, C50 in the two tables. The numbers appearing in parentheses in the two tables are the standard errors of the reported misclassification errors. Figure 2 provides boxplots of the error rates of various classifiers. As shown in Table 1, Table 2 and Figure 2, for both sample sizes, the classifier $\widehat{\Gamma}_n$ outperforms $\widehat{\Gamma}_{CC}$ regardless of the missingness mechanism or the number of missing patterns involved. This is particularly true when the percentage of missing data is at 80%.

In passing, we note that the proposed classifier $\widehat{\Gamma}_n$ can also perform better than $\widetilde{\Gamma}_n$ whenever the dependence of the missing probability mechanism on class Y (as defined via (13)) dominates its dependence on the observed and/or unobserved segments of the curves (i.e., the constant a is orders of magnitude larger than b and c in (14)). Since the missing probability mechanism is essentially the conditional law of δ, in such cases the correlation between Y and δ can be higher than that between Y and the missing covariate. As a result, one can expect better performance compared to the case of fully observed covariates. In other words, in such cases, the random variable δ which is always observable can sometimes work better at predicting Y than the missing part of the covariate curve. See, for example, the cases C7, C8, C22, C23 in Table 1 and the cases C31, C32, C43, C44 in Table 2.

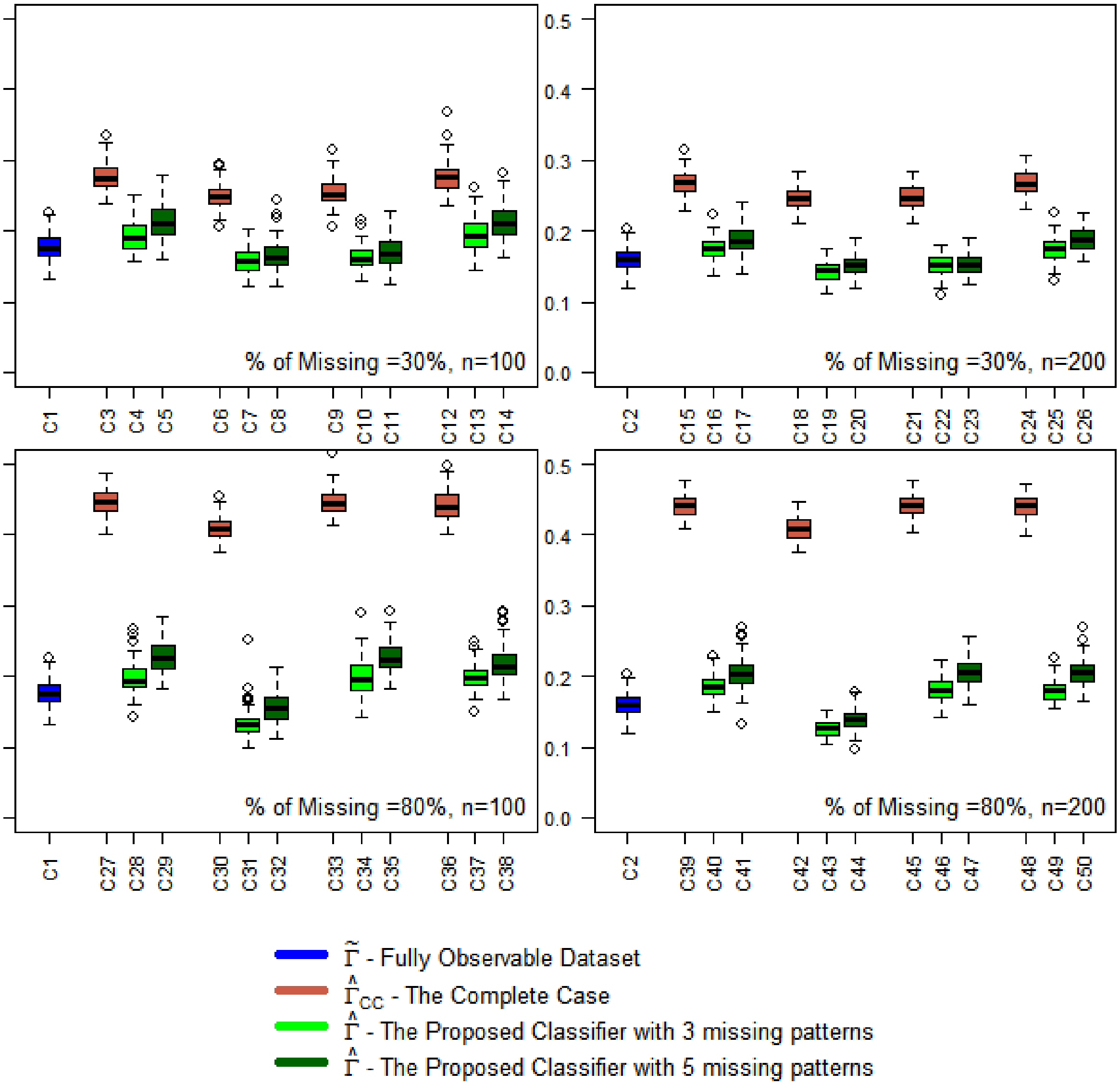


Figure 2: Boxplots of the error rates of all classifiers (C1, C2, . . . , C50) that appear in Table 1 and Table 2

Table 1: Misclassification errors of $\widetilde{\Gamma}_n$ (fully observable data), $\widehat{\Gamma}_{CC}$ (complete case analysis), $\widehat{\Gamma}_n$ (the proposed classifier with $M = 3$ and 5 missing patterns). The percentage of missing data is 30%. The numbers in parentheses are the standard errors over 100 Monte Carlo runs.

% of Missing	**n**	**Missing Mechanism**	**a2**	**b2**	**c2**	**a3**	**b3**	**c3**	**a4**	**b4**	**c4**	**a5**	**b5**	**c5**	**Error of** $\widetilde{\Gamma}_n$	**Error of** $\widehat{\Gamma}_{CC}$	**Error of** $\widehat{\Gamma}_n$ **with** M**=3 missing patterns**	**Error of** $\widehat{\Gamma}_n$ **with** M**=5 missing patterns**
30%	100	NMAR	0	0.95	0.13	0	0.9	1	0	1.05	0.13	0	1.2	1	C1 0.1771 (0.0178)	C3 0.2774 (0.0179)	C4 0.1920 (0.0215)	C5 0.2124 (0.0241)
		NMAR	2	0.01	0.8	2	0.01	0.4	2	0.01	0.3	2	0.01	0.3		C6 0.2502 (0.0185)	C7 0.1581 (0.0176)	C8 0.1650 (0.0199)
		MAR	1.9	0.075	0	1.9	0.075	0	2	0.085	0	2	0.085	0		C9 0.2549 (0.0187)	C10 0.1626 (0.0165)	C11 0.1700 (0.0231)
		MCAR		NA			NA			NA			NA			C12 0.2773 (0.0227)	C13 0.1939 (0.0224)	C14 0.2121 (0.0238)
	200	NMAR	0	0.95	0.13	0	0.9	1	0	1.05	0.13	0	1.2	1	C2 0.1607 (0.0154)	C15 0.2672 (0.0159)	C16 0.1743 (0.0157)	C17 0.1883 (0.0202)
		NMAR	2	0.01	0.8	2	0.01	0.4	2	0.01	0.3	2	0.01	0.3		C18 0.2451 (0.0137)	C19 0.1430 (0.0128)	C20 0.1517 (0.0135)
		MAR	1.9	0.075	0	1.9	0.075	0	2	0.085	0	2	0.085	0		C21 0.2470 (0.0161)	C22 0.1510 (0.0147)	C23 0.1529 (0.0146)
		MCAR		NA			NA			NA			NA			C24 0.2686 (0.0180)	C25 0.1744 (0.0171)	C26 0.1871 (0.0159)

Table 2: Misclassification errors of $\widetilde{\Gamma}_n$ (fully observable data), $\widehat{\Gamma}_{CC}$ (complete case analysis), $\widehat{\Gamma}_n$ (the proposed classifier with $M = 3$ and 5 missing patterns). The percentage of missing data is 80%. The numbers in parentheses are the standard errors over 100 Monte Carlo runs.

% of Missing	**n**	**Missing Mechanism**	**a2**	**b2**	**c2**	**a3**	**b3**	**c3**	**a4**	**b4**	**c4**	**a5**	**b5**	**c5**	**Error of $\widetilde{\Gamma}_n$**	**Error of $\widehat{\Gamma}_{CC}$**	**Error of $\widehat{\Gamma}_n$ with M=3 missing patterns**	**Error of $\widehat{\Gamma}_n$ with M=5 missing patterns**
80%	100	NMAR	0	-1.9	1.5	0	-1.45	-3	0	-2.1	1.5	0	-2	-3	C1 0.1771 (0.0178)	C27 0.4463 (0.0184)	C28 0.1970 (0.0205)	C29 0.2284 (0.0240)
		NMAR	-5	-0.4	0.25	-4	-0.5	-0.15	-5	-0.4	0.25	-4	-0.5	-0.15		C30 0.4096 (0.0158)	C31 0.1335 (0.0183)	C32 0.1554 (0.0207)
		MAR	1	-3	0	-1.45	-0.95	0	0.6	-3	0	-1.9	-0.95	0		C33 0.4464 (0.0165)	C34 0.1991 (0.0244)	C35 0.2273 (0.0202)
		MCAR		NA			NA			NA			NA			C36 0.4419 (0.0187)	C37 0.1978 (0.0181)	C38 0.2178 (0.0256)
	200	NMAR	0	-1.9	1.5	0	-1.45	-3	0	-2.1	1.5	0	-2	-3	C2 0.1607 (0.0154)	C39 0.4410 (0.0157)	C40 0.1872 (0.0151)	C41 0.2048 (0.0227)
		NMAR	-5	-0.4	0.25	-4	-0.5	-0.15	-5	-0.4	0.25	-4	-0.5	-0.15		C42 0.4087 (0.0164)	C43 0.1264 (0.0118)	C44 0.1394 (0.0150)
		MAR	1	-3	0	-1.45	-0.95	0	0.6	-3	0	-1.9	-0.95	0		C45 0.4412 (0.0153)	C46 0.1824 (0.0167)	C47 0.2061 (0.0194)
		MCAR		NA			NA			NA			NA			C48 0.4398 (0.0179)	C49 0.1800 (0.0141)	C50 0.2058 (0.0181)

4.2 Application: Share Price Increase Data

In this section, we use a real data to illustrate the proposed classifier $\widehat{\Gamma}_n$ in (11). A company's stock price reflects investor perception of its ability to earn and grow its profits in the future. Many studies suggests that near future events can be predicted using historical stock prices (Fama [39], Khan et. al. [40] and Bonde and Khaled [41]). Here, we study the dataset comprising of daily prices of 965 companies listed on the NASDAQ 100 companies, which is available from `http://www.timeseriesclassification.com/dataset.php`. The aim is to predict whether the share price of a company will rise significantly after quarterly announcement of the Earning Per Share based on its 60-day price movement before the reporting date. Here, each observation was a series of 60-day percentage changes of the close price from the day before. The class variable y_i was coded as $0 =$ price did not increase by more than 5 percent after the company report released and $1 =$ price increased by more than 5 percent after the company report released.

Since our main goal is to compare the performance of our proposed classifier to the ones based on the full data and the complete case, we extracted fragments from the full curves to form 3 missing patterns $\delta_i = 1, 2, 3$ where $s_1 := \mathcal{I} = [1, 60], s_2 = [16, 60] \subset \mathcal{I}$, and $s_3 = [31, 60] \subset \mathcal{I}$. For each observation, the set s can be selected to be any one of the patterns s_1, s_2 or s_3 with probabilities a_i, b_i, c_i respectively. We considered the MCAR scenario where $a_i = b_i = c_i = 1/3, \forall i = 1, \ldots, 965$ and the NMAR scenario where $b_i = I\{8 \leqslant ||\mathbf{X}_i|| \leqslant 14\}/3$, $c_i = I\{(Y_i = 0 \ \& \ ||\mathbf{X}_i|| < 8)$ or $(Y_i = 1 \ \& \ ||\mathbf{X}_i|| > 18)\}/1.5$, and $a_i = 1 - b_i - c_i$, $i = 1, \ldots, 965$ (here, $||\mathbf{X}_i||$ is the norm of the original 60-day percentage change vector of the i^{th} observation).

We compared the performance of our proposed classifier, $\widehat{\Gamma}_n$, with that of the classifiers based on the full data, $\widetilde{\Gamma}_n$, and the complete case analysis, $\widehat{\Gamma}_{CC}$. To do this, the sample of $n = 965$ companies was split into a training sequence and a testing sequence of ratio 70:30. The smoothing parameters h_k and d were selected using the same data splitting approach described in Section 3. Table 3 provides the average error rates of each classifier committed on the testing sequence over 100 such sample splits with standard errors given in parenthesis as well as a visual display of classifier performance. In this example, we see that the proposed classifier consistently performs well compared to the classifier based on full data regardless of the missing mechanism and the missing percentage. This phenomenon has been explained in the example in Section 4.1. In the MCAR case with 66.67% of the observations being fragemented curves, a complete case classifier eliminates much of the available information and performs poorly compared to the classifier based on the filtered curves.

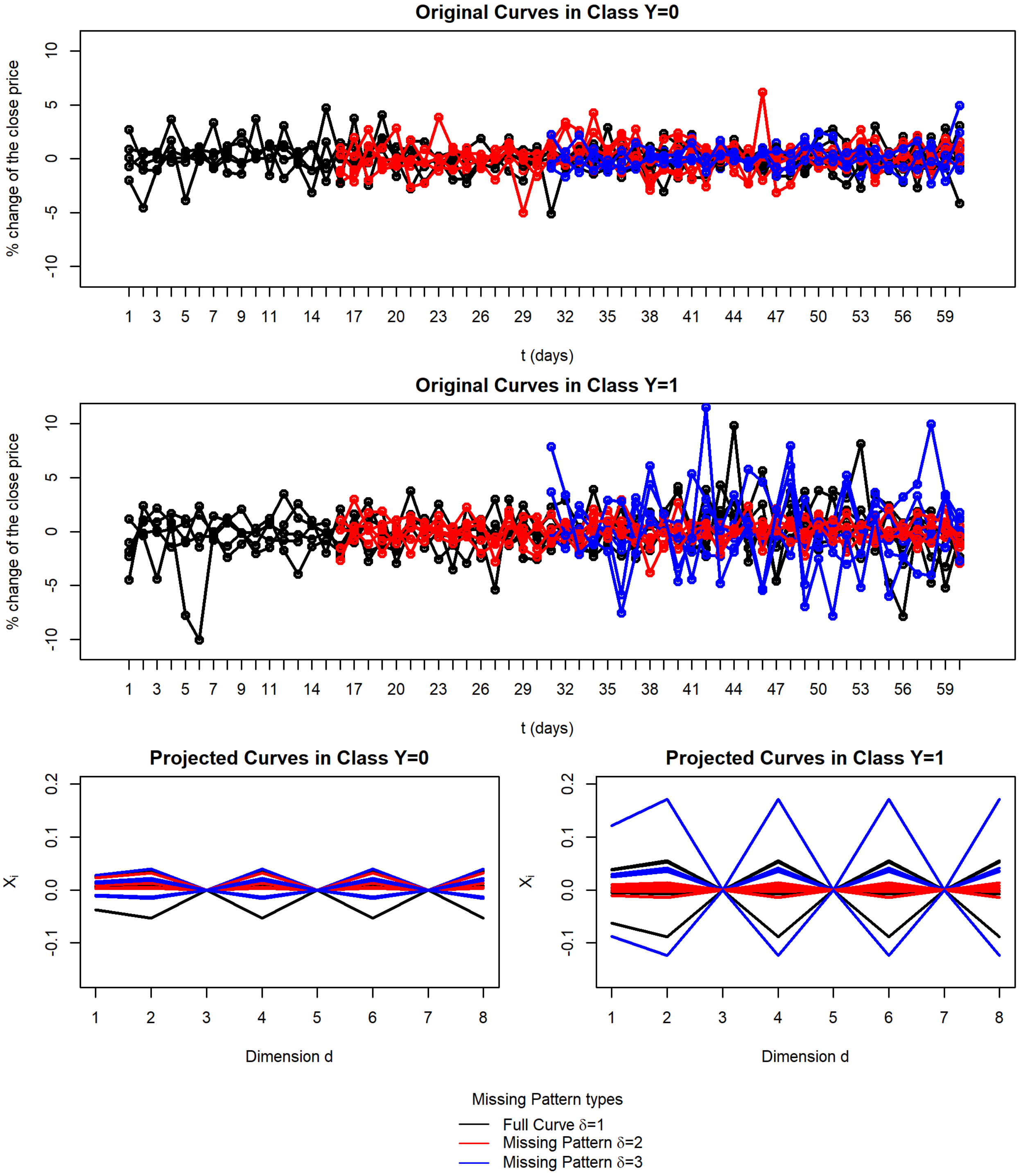
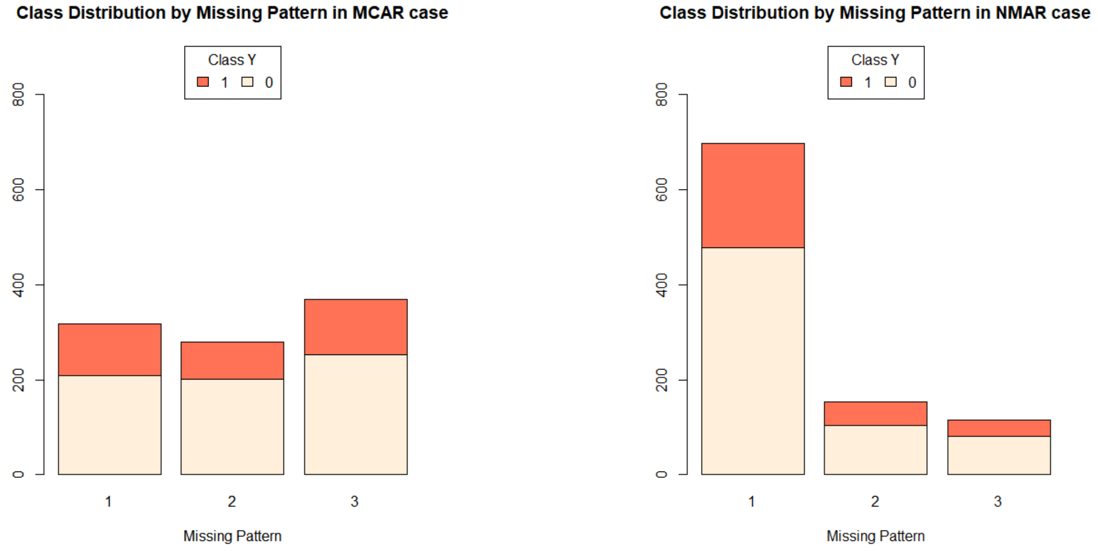


Figure 3: A sample of curves χ_i showing percentage changes in share price measured over 60 days. Various colors were used to plot the curves according to their missing pattern. The corresponding d-dimensional vector of the projected curves, $d = 1, \ldots, 8$, were also displayed.



(a) (b)

Figure 4: Distributions and proportion of class membership for each missing pattern created in the data. (a) MCAR (b) NMAR

Missing Mechanism	**% Missing Data**	$\widetilde{\Gamma}_n$	$\widehat{\Gamma}_n$	$\widehat{\Gamma}_{CC}$
MCAR	66.67%	0.3054 (0.0139)	0.3152 (0.0161)	0.4385 (0.0189)
NMAR	27.77%		0.3018 (0.0149)	0.3657 (0.0186)

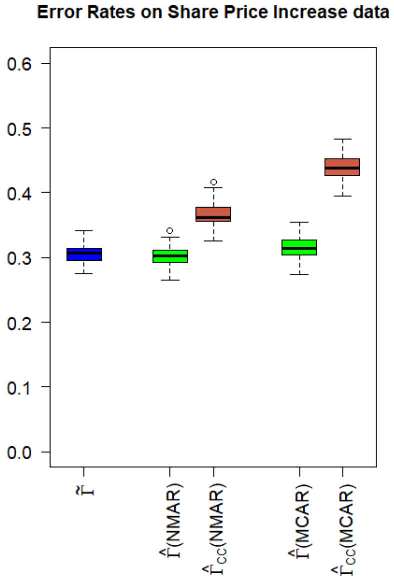


Table 3: Error rates for $\widehat{\Gamma}_n$ (the classifier based on filtered curves), $\widetilde{\Gamma}_n$ (the classifier based on full data) and $\widehat{\Gamma}_{CC}$ (the complete case analysis).

5 Proofs

In order to prove Theorem 2, we first state a number of lemmas. In what follows, we use the following notation:

$$\mathcal{R}_m(d,h_1,\ldots,h_M) \;=\; \mathbb{P}\left\{\Gamma_m^d(\mathbf{X}^{(d,\delta)}) \neq Y \big| \mathbb{D}_m\right\} \tag{15}$$

$$\widehat{\mathcal{R}}_{m,\ell}(d,h_1,\ldots,h_M) \;=\; \ell^{-1} \sum_{i:\ (\mathbf{X}^{(\delta_i)},Y_i,\delta_i)\in\mathbb{D}_\ell} I\big\{\Omega_i(m,d,h_1,\ldots,h_M)\big\}, \tag{16}$$

where $\Gamma_m^d(\mathbf{X}^{(d,\delta)})$ and $\Omega_i(m,d,h_1,\ldots,h_M)$ are as in (8) and (10), respectively.

Lemma 1 *Let $\widehat{\mathcal{R}}_{m,\ell}$ and $\mathcal{R}_m$ be as in (16) and (15). If $\ell^{-1}\log|\mathcal{H}_n| \to 0$ and $\ell^{-1}\log d_n \to 0$, where $|\mathcal{H}_n|$ is the cardinality of the set $\mathcal{H}_n$, then, as $n\to\infty$,*

$$\sup_{1\le d\le d_n\ ,\, h_1,\ldots,h_M\in\mathcal{H}_n} \Big|\widehat{\mathcal{R}}_{m,\ell}(d,h_1,\ldots,h_M) - \mathcal{R}_m(d,h_1,\ldots,h_M)\Big| \;\;\longrightarrow^{a.s.}\;\; 0\,.$$

PROOF OF LEMMA 1

First observe that for any given constant $\beta>0$,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{1\le d\le d_n\ ,\, h_1,\ldots,h_M\in\mathcal{H}_n} \Big|\widehat{\mathcal{R}}_{m,\ell}(d,h_1,\ldots,h_M) - \mathcal{R}_m(d,h_1,\ldots,h_M)\Big| > \beta\right\}\\
&\le \sum_{1\le d\le d_n}\ \sum_{h_1,\ldots,h_M\in\mathcal{H}_n} \mathbb{P}\left\{\Big|\widehat{\mathcal{R}}_{m,\ell}(d,h_1,\ldots,h_M) - \mathcal{R}_m(d,h_1,\ldots,h_M)\Big| > \beta\right\}\\
&\le d_n|\mathcal{H}_n|^M \sup_{1\le d\le d_n}\ \sup_{h_1,\ldots,h_M\in\mathcal{H}_n} \mathbb{E}\left[\mathbb{P}\left\{\Big|\widehat{\mathcal{R}}_{m,\ell}(d,h_1,\ldots,h_M) - \mathcal{R}_m(d,h_1,\ldots,h_M)\Big| > \beta\Big|\mathbb{D}_m\right\}\right]
\end{aligned}$$

where $|\mathcal{H}_n|$ is the cardinality of the set $\mathcal{H}_n$. But, with $\Omega_i(m,d,h_1,\ldots,h_M)$ as in (10),

$$\begin{aligned}
&\mathbb{P}\left\{\Big|\widehat{\mathcal{R}}_{m,\ell}(d,h_1,\ldots,h_M) - \mathcal{R}_m(d,h_1,\ldots,h_M)\Big| > \beta\,\Big|\,\mathbb{D}_m\right\}\\
&\quad= \mathbb{P}\left\{\left|\ell^{-1}\sum_{i:\ (\mathbf{X}^{(\delta_i)},Y_i,\delta_i)\in\mathbb{D}_\ell} I\big\{\Omega_i(m,d,h_1,\ldots,h_M)\big\} - \mathbb{P}\big\{\Omega_1(m,d,h_1,\ldots,h_M)\big\}\right| > \beta\,\middle|\,\mathbb{D}_m\right\}\\
&\quad\le 2\,e^{-2\ell\beta^2}, \qquad \text{(by Hoeffding's inequality)},
\end{aligned}$$

which does not depend on $\mathbb{D}_m$ or any of the parameters $d,h_1,\ldots,h_M$. Therefore

$$\mathbb{P}\left\{\sup_{1\le d\le d_n\ ,\, h_1,\ldots,h_M\in\mathcal{H}_n} \Big|\widehat{\mathcal{R}}_{m,\ell}(d,h_1,\ldots,h_M) - \mathcal{R}_m(d,h_1,\ldots,h_M)\Big| > \beta\right\} \;\le\; 2\,d_n|\mathcal{H}_n|^M e^{-2\ell\beta^2}.$$

Furthermore, the conditions of Lemma 1 imply that $\sum_{n=1}^{\infty} d_n|\mathcal{H}_n|^M e^{-\ell\beta^2/2} < \infty$. The result now follows from an application of the Borel-Cantelli lemma.

□

Lemma 2 *Let $\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)})$ be the classifier in (11). Also, let $\widehat{\mathcal{R}}_{m,\ell}$ and $\mathcal{R}_m$ be as in (16) and (15). Then*

$$\mathbb{P}\left\{\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)}) \neq Y \Big| \mathbb{D}_n\right\} \; - \inf_{1\leq d\leq d_n\,,\,h_1,\ldots,h_M\in\mathcal{H}_n} \; \mathcal{R}_m(d,h_1,\ldots,h_M)$$
$$\leq \; 2 \sup_{1\leq d\leq d_n\,,\,h_1,\ldots,h_M\in\mathcal{H}_n} \left|\widehat{\mathcal{R}}_{m,\ell}(d,h_1,\ldots,h_M) - \mathcal{R}_m(d,h_1,\ldots,h_M)\right|.$$

PROOF of LEMMA 2

The proof of this lemma, which is similar to that of Lemma 8.2 of Devroye et al [16], is straightforward and will not be given here. □

Lemma 3 *Let $\Gamma^{B,d}(\mathbf{X}^{(d,\delta)})$ be the classifier defined via (6) and (5). Let $d \geq 1$ be any fixed integer and consider any classifier of the form $\Gamma^d(\mathbf{X}^{(d,\delta)}) := \sum_{k=1}^M I\{\delta = k\} \cdot g_{d,k}(\mathbf{X}^{(d,k)})$, where $g_{d,k}(\mathbf{X}^{(d,k)}) = I\{G_{d,k}(\mathbf{X}^{(d,k)}) > 0\}$ for some function $G_{d,k} : \mathbb{R}^d \to [-1,1]$, and $\mathbf{X}^{(d,k)}$ represents the first d components of $\mathbf{X}^{(k)}$ in (1). Then*

$$\mathbb{P}\left\{\Gamma^d(\mathbf{X}^{(d,\delta)}) \neq Y\right\} - \mathbb{P}\left\{\Gamma^{B,d}(\mathbf{X}^{(d,\delta)}) \neq Y\right\} \;\; \leq \;\; \sum_{k=1}^M \mathbb{E}\left|\phi_{d,k}(\mathbf{X}^{(d,k)}) - G_{d,k}(\mathbf{X}^{(d,k)})\right|,$$

where $\phi_{d,k}(\mathbf{X}^{(d,k)})$ is as in (5).

PROOF OF LEMMA 3

It is not difficult to show that

$$\mathbb{P}\left\{\Gamma^d(\mathbf{X}^{(d,\delta)}) \neq Y\right\} - \mathbb{P}\left\{\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)}) \neq Y\right\}$$
$$\leq \; \sum_{k=1}^M \mathbb{E}\left(I\left\{g_{d,k}(\mathbf{X}^{(d,k)}) \neq I\left\{\phi_{d,k}(\mathbf{X}^{(d,k)}) > 0\right\}\right\} \times \left|\phi_{d,k}(\mathbf{X}^{(d,k)})\right|\right).$$

Now, observe that on the set $\left\{g_{d,k}(\mathbf{X}^{(d,k)}) \neq I\left\{\phi_{d,k}(\mathbf{X}^{(d,k)}) > 0\right\}\right\}$, one has

$$\mathbb{E}\left(I\left\{g_{d,k}(\mathbf{X}^{(d,k)}) \neq I\left\{\phi_{d,k}(\mathbf{X}^{(d,k)}) > 0\right\}\right\} \times \left|\phi_{d,k}(\mathbf{X}^{(d,k)})\right|\right) \;\; \leq \;\; \mathbb{E}\left|\phi_{d,k}(\mathbf{X}^{(d,k)}) - G_{d,k}(\mathbf{X}^{(d,k)})\right|,$$

which completes the proof of the lemma.

□

The following result is an immediate corollary to Lemma 3.

Corollary 1 *Let $\Gamma^{B,d}(\mathbf{X}^{(d,\delta)})$ be the classifier defined via (6) and (5). Also, for $k = 1,\ldots,M$, let $G_{m,d,k}(\mathbf{X}^{(d,k)})$ be any sample-based version of the function $G_{d,k}(\mathbf{X}^{(d,\delta)})$ that appears in Lemma 3, based on the training sample $\mathbb{D}_m$, and consider the classifier*

$$\widetilde{\Gamma}_m(\mathbf{X}^{(d,\delta)}) = \sum_{k=1}^M I\{\delta = k\} \cdot g_{m,d,k}(\mathbf{X}^{(d,k)}),$$

where $g_{m,d,k}(\mathbf{X}^{(d,k)}) = I\left\{G_{m,d,k}(\mathbf{X}^{(d,k)}) > 0\right\}$. *Then*

$$\mathbb{P}\left\{\widetilde{\Gamma}_m(\mathbf{X}^{(d,\delta)}) \neq Y \Big| \mathbb{D}_m\right\} - \mathbb{P}\left\{\Gamma^{B,d}(\mathbf{X}^{(d,\delta)}) \neq Y\right\} \;\; \leq \;\; \sum_{k=1}^{M} \mathbb{E}\left[\left|\phi_{d,k}(\mathbf{X}^{(d,k)}) - G_{m,d,k}(\mathbf{X}^{(d,k)})\right| \Big| \mathbb{D}_m\right].$$

PROOF of COROLLARY 1

The proof of Corollary 1 is the same as that of Lemma 3 and is obtained by conditioning on the training data $\mathbb{D}_m$.

□

The next lemma is a well-known result on the performance of the L^1-norm of kernel regression estimators.

Lemma 4 *[Györfi et al ([20], Lemma 23.9).]*
Let $(U, \mathbf{V}) \in [-B, B] \times \mathbb{R}^d$, *where* $B < \infty$, *and let* $\phi(\mathbf{v}) = \mathbb{E}[U|\mathbf{V} = \mathbf{v}]$ *be the regression function. Let* $\mathbb{D}_n = \{(U_1, \mathbf{V}_1), \ldots, (U_n, \mathbf{V}_n)\}$ *be the data (iid), where* $(U_i, \mathbf{V}_i) \overset{iid}{=} (U, \mathbf{V})$, *and define* $\widehat{\phi}_n(\mathbf{v}) = \sum_{i=1}^n U_i \mathcal{K}((\mathbf{v} - \mathbf{V}_i)/h_n) \big/ \{n\, \mathbb{E}[\mathcal{K}(\mathbf{v} - \mathbf{V})/h_n)]\}$, *where* $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}_+$ *is regular. If* $h_n \to 0$ *and* $nh_n^d \to \infty$, *as* $n \to \infty$, *then for any distribution of* $(U, \mathbf{V})$, *any* $\epsilon > 0$, *and* n *large enough,*

$$\mathbb{P}\left\{\mathbb{E}\left[\left|\widehat{\phi}_n(\mathbf{V}) - \phi(\mathbf{V})\right| \Big| \mathbb{D}_n\right] > \epsilon\right\} \leq e^{-n\epsilon^2/(8B\rho)^2},$$

where $\rho \equiv \rho(\mathcal{K})$ *is a positive constant depending on the kernel* $\mathcal{K}$ *only.*

PROOF OF THEOREM 2

Let $\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)})$ and $\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)})$ be as in (6) and (4), respectively, and observe that, in view of part (ii) of Theorem 1, one has

$$\begin{aligned}&\mathbb{P}\{\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)}) \neq Y\} - \mathbb{P}\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) \neq Y\} \\ &\quad \leq \sum_{k=1}^{M} \mathbb{E}\left|\mathbb{E}\Big[(2Y-1)I\{\delta = k\}\Big|\mathbf{X}^{(k)}\Big] - \mathbb{E}\Big[(2Y-1)I\{\delta = k\}\Big|\mathbf{X}^{(d,k)}\Big]\right|, \qquad (17)\end{aligned}$$

which follows upon taking the function $\varphi_k(\mathbf{X}^{(k)})$ that appears in part (ii) of Theorem 1 to be the same as the right side of (5) (in which case the classifier $\Gamma(\mathbf{X}^{(\delta)})$ of Theorem 1(ii) will coincide with the classifier $\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)})$ in (6)). Here, as before, $\mathbf{X}^{(k)} = (X_1^{(k)}, X_2^{(k)}, \ldots)$ and $\mathbf{X}^{(d,k)} = (X_1^{(k)}, \ldots, X_d^{(k)})$. Let $S_d^{(k)} = \mathbb{E}\big[(2Y-1)I\{\delta = k\}\big|\mathbf{X}^{(d,k)}\big]$ and $S_\infty^{(k)} = \mathbb{E}\big[(2Y-1)I\{\delta = k\}\big|\mathbf{X}^{(k)}\big]$, and observe that for any $k = 1, \ldots, M$, and any integers $d_1 < d_2$, one has

$$\mathbb{E}\Big[S_{d_2}^{(k)}\Big|X_1^{(k)}, \ldots, X_{d_1}^{(k)}\Big] \overset{\text{a.s.}}{=} S_{d_1}^{(k)}.$$

Furthermore, $\sup_{d\geq 1}\left|S_d^{(k)}\right| \leq 1$, and therefore $\{S_d^{(k)}, d = 1, 2, \dots\}$ is a martingale with respect to the increasing sequence of σ-fields, $\sigma(X_1^{(k)}, \dots, X_d^{(k)})$. Invoking the martingale convergence theorem (see, for example, Sec. 1.3 of Hall and Heyde [21]), and arguing as in Biau et al. [4], we find $S_d^{(k)} \to^{a.s.} S_\infty^{(k)}$, as $d \to \infty$. This fact together with the bound in (17) and an application of the dominated convergence theorem yield $\mathbb{P}\{\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)}) \neq Y\} - \mathbb{P}\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) \neq Y\} \to 0$, as $d \to \infty$. Consequently, for every $\epsilon > 0$, and n sufficiently large, there is a $d_\epsilon \in [1, d_n]$ such that

$$\mathbb{P}\{\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)}) \neq Y\} - \mathbb{P}\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) \neq Y\} \;\leq\; \epsilon \quad \text{for all } d \geq d_\epsilon$$

(recall $d_n \to \infty$ as $n \to \infty$). Therefore, for any $\widetilde{h}_k \equiv \widetilde{h}_k(n) \in \mathcal{H}_n$, $k = 1, \dots, n$, satisfying the conditions of Theorem 2, any $\epsilon > 0$, and n large enough, one has

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)}) \neq Y \,\middle|\, \mathbb{D}_n\right\} - \mathbb{P}\left\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) \neq Y\right\} \\
&\quad = \; \mathbb{P}\left\{\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)}) \neq Y \,\middle|\, \mathbb{D}_n\right\} \;-\; \inf_{1\leq d\leq d_n\,,\, h_1,\dots,h_M\in\mathcal{H}_n} \;\mathcal{R}_m(d, h_1, \dots, h_M) \\
&\qquad +\; \inf_{1\leq d\leq d_n\,,\, h_1,\dots,h_M\in\mathcal{H}_n} \Big\{\mathcal{R}_m(d, h_1, \dots, h_M) - \mathbb{P}\{\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)}) \neq Y\} \\
&\qquad\qquad\qquad +\; \mathbb{P}\{\Gamma^{\mathrm{B},d}(\mathbf{X}^{(d,\delta)}) \neq Y\}\Big\} \;-\; \mathbb{P}\left\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) \neq Y\right\} \\
&\quad \leq \; \mathbb{P}\left\{\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)}) \neq Y \,\middle|\, \mathbb{D}_n\right\} \;-\; \inf_{1\leq d\leq d_n\,,\, h_1,\dots,h_M\in\mathcal{H}_n} \;\mathcal{R}_m(d, h_1, \dots, h_M) \\
&\qquad +\; \mathcal{R}_m(d_\epsilon, \widetilde{h}_1, \dots, \widetilde{h}_M) - \mathbb{P}\{\Gamma^{\mathrm{B},d_\epsilon}(\mathbf{X}^{(d_\epsilon,\delta)}) \neq Y\} \\
&\qquad +\; \epsilon
\end{aligned} \tag{18}$$

Now, in view of lemmas 1 and 2, as $n \to \infty$, we have

$$\mathbb{P}\left\{\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)}) \neq Y \,\middle|\, \mathbb{D}_n\right\} \;-\; \inf_{1\leq d\leq d_n\,,\, h_1,\dots,h_M\in\mathcal{H}_n} \;\mathcal{R}_m(d, h_1, \dots, h_M) \longrightarrow^{\text{a.s.}} 0. \tag{19}$$

Next, define

$$\widetilde{\phi}_{m,d,h_k}(\mathbf{x}) = \frac{\widehat{\phi}_{m,d,h_k}(\mathbf{x})}{m \cdot \mathbb{E}\left[\mathcal{K}_k\left(\frac{\mathbf{x}-\mathbf{X}^{(d,k)}}{h_k}\right)\right]}\,,$$

where $\widehat{\phi}_{m,d,h_k}(\mathbf{x})$ is as in (7), and observe that the classifier Γ_m^d in (8) can alternatively be written as $\Gamma_m^d(\mathbf{X}^{(d,\delta)}) = \sum_{k=1}^M I\{\delta = k\} I\big\{\widetilde{\phi}_{m,d,h_k}(\mathbf{X}^{(d,k)}) > 0\big\}$. Therefore, by Corollary 1,

$$\begin{aligned}
\mathcal{R}_m(d_\epsilon, \widetilde{h}_1, \dots, \widetilde{h}_M) - \mathbb{P}\{\Gamma^{\mathrm{B},d_\epsilon}(\mathbf{X}^{(d_\epsilon,\delta)}) \neq Y\} \;\;&\leq\;\; \sum_{i=1}^M \mathbb{E}\left[\left|\phi_{d_\epsilon,k}(\mathbf{X}^{(d_\epsilon,k)}) - \widetilde{\phi}_{m,d_\epsilon,\widetilde{h}_k}(\mathbf{X}^{(d_\epsilon,k)})\right| \middle| \mathbb{D}_m\right] \\
&\quad \longrightarrow^{\text{a.s.}} 0, \quad \text{as } n \to \infty,
\end{aligned} \tag{20}$$

(by Lemma 4 and the Borel-Cantelli lemma),

where $\phi_{d_\epsilon,k}$ is as in (5) with d replaced by d_ϵ. Therefore, in view of (18), (19), and (20), for any $\epsilon > 0$,

$$\lim_{n\to\infty} \left[\mathbb{P}\left\{\widehat{\Gamma}_n(\mathbf{X}^{(\widehat{d},\delta)}) \neq Y \,\middle|\, \mathbb{D}_n\right\} - \mathbb{P}\left\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) \neq Y\right\}\right] \;\;\leq\;\; \epsilon,$$

almost surely. This completes the proof of Theorem 2.

□

Appendix.

Proof of Theorem 1

Part (i).

The proof is similar to that of Mojirsheibani and Montazeri ([31], Theorem 3) and goes as follows. Let $\mathbf{X}^{(k)}$ be as in (1), $k = 1, \ldots, M$, and define the functions $r_k(\mathbf{x}, y) := \mathbb{P}\left\{\delta = k \,\middle|\, \mathbf{X}^{(k)} = \mathbf{x}, Y = y\right\}$, $y = 0, 1$, and $\eta_k(\mathbf{x}) := \mathbb{P}\left\{Y = 1 \,\middle|\, \mathbf{X}^{(k)} = \mathbf{x}\right\} = \mathbb{E}\left[Y \middle| \mathbf{X}^{(k)} = \mathbf{x}\right]$, and observe that the function ϕ_k in (3) can be written as

$$\begin{aligned}
\phi_k(\mathbf{X}^{(k)}) &= \mathbb{E}\left\{\mathbb{E}\left[(2Y-1)I\{\delta = k\} \,\middle|\, \mathbf{X}^{(k)}, Y\right] \middle| \mathbf{X}^{(k)}\right\}, \\
&= \mathbb{E}\left[(2Y-1)\mathbb{P}\{\delta = k \middle| \mathbf{X}^{(k)}, Y\} \,\middle|\, \mathbf{X}^{(k)}\right] \\
&= \mathbb{E}\left[(2Y-1)\Big(Y \cdot r_k(\mathbf{X}^{(k)}, 1) + (1-Y) r_k(\mathbf{X}^{(k)}, 0)\Big) \,\middle|\, \mathbf{X}^{(k)}\right] \\
&= \mathbb{E}\left[Y \cdot r_k(\mathbf{X}^{(k)}, 1) + (Y-1) \cdot r_k(\mathbf{X}^{(k)}, 0) \,\middle|\, \mathbf{X}^{(k)}\right], \text{ (because } Y^2{=}Y) \\
&= \eta_k(\mathbf{X}^{(k)}) r_k(\mathbf{X}^{(k)}, 1) + \Big(\eta_k(\mathbf{X}^{(k)}) - 1\Big) r_k(\mathbf{X}^{(k)}, 0). \qquad (21)
\end{aligned}$$

Therefore, the classifier Γ^{B} in (4) can be written as

$$\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) = \sum_{k=1}^{M} I\{\delta = k\} \cdot I\left\{\eta_k(\mathbf{X}^{(k)}) r_k(\mathbf{X}^{(k)}, 1) + \Big(\eta_k(\mathbf{X}^{(k)}) - 1\Big) r_k(\mathbf{X}^{(k)}, 0) > 0\right\},$$

and this can be used to write

$$\begin{aligned}
&\mathbb{P}\left\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) = Y\right\} \\
&\quad = \mathbb{P}\left\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(d,\delta)}) = 1, Y = 1\right\} + \mathbb{P}\left\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) = 0, Y = 0\right\} \\
&\quad = \sum_{k=1}^{M} \mathbb{P}\left\{Y = 1, \delta = k, \left[\phi_k(\mathbf{X}^{(k)}) > 0\right]\right\} + \sum_{k=1}^{M} \mathbb{P}\left\{Y = 0, \delta = k, \left[\phi_k(\mathbf{X}^{(k)}) \leq 0\right]\right\} \\
&\quad := \sum_{k=1}^{M} \pi_{k1} + \sum_{k=1}^{M} \pi_{k0}, \quad \text{(say)}.
\end{aligned}$$

But

$$\begin{aligned}
\pi_{k1} &= \mathbb{E}\Big[I\{Y = 1\} \cdot I\left\{\phi_k(\mathbf{X}^{(k)}) > 0\right\} \cdot \mathbb{P}\left\{\delta = k \middle| \mathbf{X}^{(k)}, Y\right\}\Big] \\
&= \mathbb{E}\Big[I\left\{\phi_k(\mathbf{X}^{(k)}) > 0\right\} \cdot r_k(\mathbf{X}^{(k)}, 1) \cdot \mathbb{E}\big[I\{Y = 1\} \big| \mathbf{X}^{(k)}\big]\Big]
\end{aligned}$$

$$= \quad \mathbb{E}\Big[I\left\{\phi_k(\mathbf{X}^{(k)}) > 0\right\} \cdot r_k(\mathbf{X}^{(k)}, 1) \cdot \eta_k(\mathbf{X}^{(k)})\Big].$$

Furthermore, similar arguments yield $\pi_{k0} = \mathbb{E}\big[I\left\{\phi_k(\mathbf{X}^{(k)}) \leq 0\right\} \cdot r_k(\mathbf{X}^{(k)}, 0) \cdot \big(1 - \eta_k(\mathbf{X}^{(k)})\big)\big]$. Thus, we have

$$\begin{aligned}\mathbb{P}\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) = Y\} \quad &= \quad \sum_{k=1}^{M}\Bigg(\mathbb{E}\Big[I\left\{\phi_k(\mathbf{X}^{(k)}) > 0\right\} \cdot r_k(\mathbf{X}^{(k)}, 1) \cdot \eta_k(\mathbf{X}^{(k)})\Big] \\ &\qquad + \mathbb{E}\Big[I\left\{\phi_k(\mathbf{X}^{(k)}) \leq 0\right\} \cdot r_k(\mathbf{X}^{(k)}, 0) \cdot \Big(1 - \eta_k(\mathbf{X}^{(k)})\Big)\Big]\Bigg).\end{aligned}$$

Also, for any other classifier $\Gamma(\mathbf{X}^{(\delta)})$ given by (2), it is not difficult to see that

$$\begin{aligned}\mathbb{P}\{\Gamma(\mathbf{X}^{(\delta)}) = Y\} \quad &= \quad \sum_{k=1}^{M}\Bigg(\mathbb{E}\Big[I\left\{g_k(\mathbf{X}^{(k)}) = 1\right\} \cdot r_k(\mathbf{X}^{(k)}, 1) \cdot \eta_k(\mathbf{X}^{(k)})\Big] \\ &\qquad + \mathbb{E}\Big[I\left\{g_k(\mathbf{X}^{(k)}) = 0\right\} \cdot r_k(\mathbf{X}^{(k)}, 0) \cdot \Big(1 - \eta_k(\mathbf{X}^{(k)})\Big)\Big]\Bigg).\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{P}\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) \neq Y\} - \mathbb{P}\{\Gamma(\mathbf{X}^{(\delta)}) \neq Y\} \\ &= \sum_{k=1}^{M}\mathbb{E}\Big[\Big(I\left\{\phi_k(\mathbf{X}^{(k)}) > 0\right\} - I\left\{g_k(\mathbf{X}^{(k)}) = 1\right\}\Big) \cdot r_k(\mathbf{X}^{(k)}, 1) \cdot \eta_k(\mathbf{X}^{(k)})\Big] \\ &\quad + \sum_{k=1}^{M}\mathbb{E}\Big[\Big(I\left\{\phi_k(\mathbf{X}^{(k)}) \leq 0\right\} - I\left\{g_k(\mathbf{X}^{(k)}) = 0\right\}\Big) \cdot r_k(\mathbf{X}^{(k)}, 0) \cdot \big(1 - \eta_k(\mathbf{X}^{(k)})\big)\Big] \\ &= \sum_{k=1}^{M}\mathbb{E}\Bigg[I\Big\{g_k(\mathbf{X}^{(k)}) \neq I\left\{\phi_k(\mathbf{X}^{(k)}) > 0\right\}\Big\} \\ &\qquad\qquad \times \Big|r_k(\mathbf{X}^{(k)}, 1) \cdot \eta_k(\mathbf{X}^{(k)}) - r_k(\mathbf{X}^{(k)}, 0) \cdot \big(1 - \eta_k(\mathbf{X}^{(k)})\big)\Big|\Bigg] \qquad (22) \\ &\geq 0,\end{aligned}$$

where (22) follows from the definitions of Γ^{B} and Γ in conjunction with the expression in (21). This completes the proof of Part (i).

Part (ii).

First observe that the expression in (22) of the proof of Part (i) shows that, in view of (21), one has

$$\begin{aligned}&\mathbb{P}\left\{\Gamma(\mathbf{X}^{(\delta)}) \neq Y\right\} - \mathbb{P}\left\{\Gamma^{\mathrm{B}}(\mathbf{X}^{(\delta)}) \neq Y\right\} \\ &\quad \leq \quad \sum_{k=1}^{M}\mathbb{E}\Big(I\Big\{I\left\{\varphi_k(\mathbf{X}^{(k)}) > 0\right\} \neq I\left\{\phi_k(\mathbf{X}^{(k)}) > 0\right\}\Big\} \times \Big|\phi_k(\mathbf{X}^{(k)})\Big|\Big).\end{aligned}$$

Given the fact that $\varphi_k(\mathbf{X}^{(d,k)}) \in [-1,1]$, (see the statement of Part (ii) of Theorem 1), it is straightforward to see that on the set $\Big\{I\left\{\varphi_k(\mathbf{X}^{(k)})>0\right\} \neq I\left\{\phi_k(\mathbf{X}^{(k)})>0\right\}\Big\}$, one has

$$\mathbb{E}\bigg(I\Big\{I\Big\{\varphi_k(\mathbf{X}^{(k)})>0\Big\} \neq I\Big\{\phi_k(\mathbf{X}^{(k)})>0\Big\}\Big\} \times \Big|\phi_k(\mathbf{X}^{(k)})\Big|\bigg) \;\; \leq \;\; \mathbb{E}\Big|\phi_k(\mathbf{X}^{(k)}) - \varphi_k(\mathbf{X}^{(k)})\Big|,$$

which completes the proof of Part (ii).

□

Acknowledgements

This work is supported by the NSF Grant DMS-1916161 of Majid Mojirsheibani.

Data availability statement

The Share Price Increase data set used in Section 4.2, and a description of it, is available at `http://www.timeseriesclassification.com/dataset.php`

Additionally, a copy of the ‘R’ codes used to carry out the analysis in Section 4.2 is posted on the GitHub repository at
`https://github.com/mynhinguyen/Statistical-classification-with-incomplete-covariates-via-filtering`

References

[1] Abraham, C., Biau, G., Cadre, B. On the kernel rule for functional classification. *AISM*, 2006; 58:619-633.

[2] Alonso, A., Casado, D., Lopez-Pintado, S., Romo, J., (2014). Robust functional supervised classification for time series. *J. Classification*, 2014; 31:325-350.

[3] Berlinet, A., Biau, G., Rouviere, L. Functional classification with wavelets. *Annales de l'Institut de statistique de l'université de Paris*, 2008; 52:61-80.

[4] Biau, G., Bunea, F., Wegkamp, M.H. Functional classification in hilbert spaces. *IEEE T. Inform. Theory*, 2005; 51:2163-2172.

[5] Bugni, F. Specification test for missing functional data. *Economet. Theor.*, 2012; 28:959-1002.

[6] Cai, T., Hall, P. Prediction in functional linear regression. *Annals of Statistics*, 2006; 34:2159-2179.

[7] Cérou, F., Guyader, A. Nearest neighbor classification in infinite dimensions. *ESAIM-Probab. Stat.*, 2006; 10:340-355.

[8] Chang, C., Chen, Y., Ogden, R.T. Functional data classification: a wavelet approach. *Computation. Stat.*, 2014; 29:1497-1513.

[9] Cuevas, A., Febrero, M., Fraiman, R. Robust estimation and classification for functional data via projection-based depth notions. *Computation. Stat.*, 2007; 22:481-496.

[10] Dai, X., Müller, H.-G. Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, 2017; 104:545-560.

[11] Delaigle, A., Hall, P. Approximating fragmented functional data by segments of Markov chains. *Biometrika*, 2016; 103:779-799.

[12] Delaigle, A., Hall, P. Classification using censored functional data. *J. Amer Stat Assoc*, 2013; 108:1269-1283.

[13] Delaigle, A., Hall, P. Achieving near perfect classification for functional data. *J Royal Stat Soc B*, 2012; 74:267-286.

[14] Delaigle, A., Hall, P., Bathia, N. Componentwise classification and clustering for functional data. *Biometrika*, 2012; 99:299-313.

[15] Demirdjian, L., Mojirsheibani, M. Kernel classification with missing data and the choice of smoothing parameters. *Statistical Papers*, 2019; 60:1487-1513.

[16] Devroye, L., Györfi, L., Lugosi, G. A Probabilistic Theory of Pattern Recognition. Springer, New York. 1996.

[17] Ferraty, F., Vieu, P. Nonparametric functional data analysis, Theory and practice. Springer, New York. 2006.

[18] Ferraty, F., Vieu, P.Curves discrimination: a nonparametric functional approach. *Comput. Stat. Data Anal.*, 2003; 4:161-173.

[19] Gromenko, O., Kokoszka, P., Sojka, J. Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Annals of Applied Statistics*, 2017; 11:898-918.

[20] Györfi, L., Kohler, M., Krzyzak, A., Walk, H. A distribution-free theory of non-parametric regression. Springer, New York. 2002.

[21] Hall, P., Heyde, C.C. Martingale limit theory and its application. Academic Press. 1980.

[22] Hall, P., Horowitz, J.L. Methodology and convergence rates for functional linear regression. *Annals of Statistics*, 2007; 35:70-91.

[23] Hall, P., Poskitt, D.S., Presnell, B. Functional data-analytic approach to signal discrimination. *Technometrics*, 2001; 43:1-9.

[24] Kraus, D. Components and completion of partially observed functional data. *J. R. Stat. Soc. B.*, 2015; 77:777-801.

[25] Kraus, D., Stefanucci, M, 2019. Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika*, 2019; 106:161-180.

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

[27] López-Pintado, S., Romo, J. Depth-based classification for functional data, in *DIMACS Ser. Discrete M.*, 2006; 72:103-120.

[28] Meister, A. Optimal classification and nonparametric regression for functional data. *Bernoulli*, 2016; 22:1729-1744.

[29] Mojirsheibani, M., Shaw, C. Classification with incomplete functional covariates. *Stat. & Probab. Lett.*, 2018; 139:40-46.

[30] Mojirsheibani, M. On the correct regression function (in L_2) and its applications when the dimension of the covariate vector is random. *J. Stat. Plan. Infer.*, 2012; 142:2586-2598.

[31] Mojirsheibani, M., Montazeri, Z. Statistical classification with missing covariates. *J. R. Stat. Soc. B.*, 2007; 69:839-857.

[32] Mosler, K., Mozharovskyi, P. Fast DD-classification of functional data. *Stat. Papers*, 2017; 58:1055-1089.

[33] Rachdi, M., Vieu, P. Nonparametric regression for functional data: Automatic smoothing parameter selection. *J. Stat. Plan. Infer.*, 2007; 137:2784-2801.

[34] Sansone, G. Orthogonal Functions. Interscience, New York. 1969.

[35] Song, J.J., Deng, W., Lee, H.-J., Kwon, D. Optimal classification for time-course gene expression data using functional data analysis. *Comput. Biol. Chem.*, 2008; 32:426-432.

[36] Yao, F., Müller, H.-G. Functional quadratic regression. *Biometrika*, 2010; 97:49-64.

[37] Zhou, R., Serban, N., Gebraeel, N., Müller, H.-G. A functional time warping approach to modeling and monitoring truncated degradation signals. *Technometrics*, 2014; 56:67-77.

[38] Zygmund, A. *Trigonometric Series I.* Cambridge Univ. Press. 1959.

[39] Fama, E. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 1970; 25:383-417.

[40] Khan, Z. H., Alin, T. S., and Hussain, A. Price prediction of share market using artificial neural network (ANN). *International Journal of Computer Applications*, 2011; 22(2):42-47.

[41] Bonde, G. and Khaled, R. Extracting the best features for predicting stock prices using machine learning. *Proceedings of the 2012 International Conference on Artificial Intelligence*, 2012; 1:222-229.