Biometrika (2017), **xx**, x, *pp.* 1–10
Printed in Great Britain

Supplementary Material for Optimal Bayes Classifiers for Functional Data and Density Ratios

By Xiongtao Dai, Hans-Georg Müller
Department of Statistics, University of California, Davis, California 95616, U.S.A.
dai@ucdavis.edu hgmueller@ucdavis.edu

and Fang Yao
Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada
fyao@utstat.toronto.edu

1. An Alternative Nonparametric Regression Estimate

An alternative approach for estimating the density ratios is via nonparametric regression. This is motivated by Bayes' theorem, as follows,

$$\begin{aligned}\frac{f_{j1}(u)}{f_{j0}(u)} &= \frac{\mathrm{pr}(Y=1 \mid \xi_j = u)p_j(u)/\mathrm{pr}(Y=1)}{\mathrm{pr}(Y=0 \mid \xi_j = u)p_j(u)/\mathrm{pr}(Y=0)} \\ &= \frac{\mathrm{pr}(Y=1 \mid \xi_j = u)/\pi_1}{\mathrm{pr}(Y=0 \mid \xi_j = u)/\pi_0} = \frac{\pi_0 \mathrm{pr}(Y=1 \mid \xi_j = u)}{\pi_1(1-\mathrm{pr}(Y=1 \mid \xi_j = u))}, \qquad (1)\end{aligned}$$

where $p_j(\cdot)$ is the marginal density of the jth projection. This reduces the construction of nonparametric Bayes classifiers to a sequence of nonparametric regressions $E(Y \mid \xi_j = u) = \mathrm{pr}(Y=1 \mid \xi_j = u)$. These again can be implemented by a kernel method (Nadaraya, 1964; Watson, 1964), smoothing the scatter plots of the pooled estimated scores $\hat{\xi}_{ijk}$ of group k, which leads to the nonparametric estimators

$$\hat{E}(Y \mid \hat{\xi}_j = u) = \frac{\sum_{k=0}^{1}\sum_{i=1}^{n_k} kK(\frac{u-\hat{\xi}_{ijk}}{h_j})}{\sum_{k=0}^{1}\sum_{i=1}^{n_k} K(\frac{u-\hat{\xi}_{ijk}}{h_j})},$$

where $h_j = h\{(\hat{\lambda}_{j0}+\hat{\lambda}_{j1})/2\}^{1/2}$ is the bandwidth. This results in estimates $\hat{E}(Y \mid \hat{\xi}_j = u) = \hat{\mathrm{pr}}(Y=1 \mid \hat{\xi}_j = u)$ that we plug-in at the right hand side of (1), which then yields an alternative estimate of the density ratio, replacing the two kernel density estimates $\hat{f}_{j1}(u)$, $\hat{f}_{j0}(u)$ by just one nonparametric regression estimate $\hat{E}(Y \mid \hat{\xi}_j = u)$.

The estimated criterion function based on kernel regression is

$$\hat{Q}_J^R(x) = \log\frac{\hat{\pi}_1}{\hat{\pi}_0} + \sum_{j\le J}\log\frac{\hat{\pi}_0\hat{E}(Y \mid \hat{\xi}_j = u)}{\hat{\pi}_1\{1-\hat{E}(Y \mid \hat{\xi}_j = u)\}}.$$

2. Perfect Classification when the Mean and the Covariance Functions are the Same

Let the projection scores ξ_j be independent random variables with mean 0 and variance ν_j that follow normal distributions under Π_1 and Laplace distributions under Π_0. Then

$$Q_J(X) = \sum_{j=1}^{J} \log \frac{\frac{1}{(2\pi\nu_j)^{1/2}} \exp(-\frac{\xi_j^2}{2\nu_j})}{\frac{1}{(2\nu_j)^{1/2}} \exp\{-\frac{|\xi_j|}{(\nu_j/2)^{1/2}}\}} = \sum_{j=1}^{J} \left(-\frac{1}{2} \log \pi - \frac{\xi_j^2}{2\nu_j} + \sqrt{2}|\xi_j|/\nu_j^{1/2} \right). \quad (2)$$

Since centred normal and Laplace distributions are in scale families, $\zeta_j = \xi_j/\nu_j^{1/2}$ have a common standard distribution ζ_{0k} under Π_k, irrespective of j. Denoting the summand of (2) by S_j, this implies $S_j = -(\log \pi + \zeta_j^2)/2 + \sqrt{2}|\zeta_j|$ are independent and identically distributed. Note that $E_{\Pi_0}(S_1) = (-\log \pi + 1)/2 + 1 < 0$, $E_{\Pi_1}(S_1) = -(\log \pi + 1)/2 + (\pi/2)^{-1/2} > 0$, and S_1 has finite variance under either population. So the misclassification error under Π_0 is

$$\begin{aligned} \mathrm{pr}_{\Pi_0}(Q_J(X) > 0) &= \mathrm{pr}_{\Pi_0} \left\{ \sum_{j=1}^{J} S_j - E_{\Pi_0}(\sum_{j=1}^{J} S_j) > -E_{\Pi_0}(\sum_{j=1}^{J} S_j) \right\} \\ &\le \frac{\mathrm{var}_{\Pi_0}(\sum_{j=1}^{J} S_j)}{E_{\Pi_0}(\sum_{j=1}^{J} S_j)^2} \\ &= \frac{J\, \mathrm{var}_{\Pi_0}(S_1)}{J^2 E_{\Pi_0}(S_1)^2} \to 0, \end{aligned}$$

as $J \to \infty$, where the inequality is due to Chebyshev's inequality and the last equality is due to S_j are independently and identically distributed. Similarly, the misclassification error under Π_1 also goes to zero as $J \to \infty$. Therefore perfect classification occurs under this non-Gaussian case where both the mean and the covariance functions are the same.

3. Simulation Results without Pre-smoothing

The misclassification results when using predictor functions sampled with noise that are not presmoothed are reported in Table 1. When the covariances are the same but the means differ, the centroid method is the overall best if we use the noisy predictors while the Gaussian implementation of the proposed Bayes classifiers has comparable performance. This is expected because our method estimates more parameters than the centroid method while both assume the correct model for the simulated data. All methods gain performance from pre-smoothing due to the presence of noise in the predictor functions. The logistic method benefits the most from pre-smoothing and becomes the winner when only a mean difference is present.

Table 1. *Misclassification rates (%), with standard errors in brackets for raw predictors*

n	μ	λ	Centroid	Gaussian	NPD	NPR	Logistic
Scenario A (Gaussian case)							
50	same	diff	49.3 (0.12)	23.8 (0.18)	24.5 (0.21)	26.7 (0.22)	49.4 (0.12)
	diff	same	40.2 (0.16)	41.5 (0.16)	43.4 (0.17)	42.4 (0.18)	40.7 (0.16)
	diff	diff	37.9 (0.17)	20.8 (0.18)	21.2 (0.20)	23.3 (0.22)	38.8 (0.17)
100	same	diff	49.1 (0.13)	17.2 (0.11)	18.6 (0.12)	20.0 (0.13)	49.3 (0.13)
	diff	same	37.8 (0.13)	39.2 (0.13)	41.4 (0.15)	40.2 (0.16)	38.3 (0.13)
	diff	diff	35.3 (0.14)	14.6 (0.1)	15.8 (0.10)	17.1 (0.12)	35.8 (0.15)
Scenario B (exponential case)							
50	same	diff	49.0 (0.13)	30.2 (0.19)	31.2 (0.22)	33.5 (0.23)	49.2 (0.13)
	diff	same	38.3 (0.21)	40.6 (0.21)	39.5 (0.22)	38.6 (0.21)	38.7 (0.23)
	diff	diff	35.0 (0.20)	23.3 (0.18)	23.5 (0.21)	24.3 (0.22)	35.7 (0.22)
100	same	diff	48.8 (0.14)	26.0 (0.13)	25.4 (0.14)	26.7 (0.16)	48.9 (0.13)
	diff	same	35.8 (0.16)	38.6 (0.19)	36.3 (0.18)	35.7 (0.16)	35.9 (0.16)
	diff	diff	32.4 (0.14)	18.7 (0.13)	16.7 (0.13)	17.0 (0.14)	32.7 (0.15)
Scenario C (dependent case)							
50	same	diff	48.9 (0.14)	33.3 (0.19)	35.3 (0.22)	37.3 (0.22)	49.1 (0.14)
	diff	same	39.3 (0.22)	42.1 (0.21)	41.0 (0.22)	40.1 (0.22)	39.2 (0.23)
	diff	diff	36.0 (0.21)	27.3 (0.20)	28.6 (0.21)	29.3 (0.23)	36.7 (0.23)
100	same	diff	49.1 (0.13)	29.8 (0.14)	30.6 (0.14)	31.8 (0.15)	49.0 (0.13)
	diff	same	36.4 (0.17)	39.8 (0.20)	37.9 (0.18)	37.1 (0.17)	36.3 (0.16)
	diff	diff	33.3 (0.16)	24.1 (0.15)	22.6 (0.15)	22.9 (0.16)	33.5 (0.16)

Centroid method; Gaussian, NPD, and NPR correspond to the Gaussian, nonparametric density, and nonparametric regression implementations of the proposed Bayes classifiers, respectively; Logistic, functional logistic regression.

4. Proofs

4·1. *Theorem A1 and Theorem A2*

Let $\mathcal{S}(c) = \{x : ||x|| \leq c\}$ be a bounded set of all square integrable functions for $c > 0$, where $||\cdot||$ is the L^2 norm. We will use the following lemma:

Lemma 1. *Under Conditions A1–A4, for any $j =\geq 1$, $k = 0, 1$,*

$$\sup_{x\in\mathcal{S}(c)} |\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)| = O_p\left\{h + \left(\frac{nh}{\log n}\right)^{-1/2}\right\}.$$

Proof. We prove the statement for $k=0$; the proof for $k=1$ is analogous. Let the sample mean of the jth estimated projection $\hat{\xi}_j$ be $\bar{\xi}_j$. Observe

$$\begin{aligned}
&\sup_{x\in\mathcal{S}(c)}\left|\hat{g}_{j0}\left(\frac{\hat{x}_j-\bar{\xi}_j}{\hat{\lambda}_{j0}^{1/2}}\right)-g_{j0}\left(\frac{x_j}{\lambda_{j0}^{1/2}}\right)\right| \\
&\quad\le \sup_{x\in\mathcal{S}(c)}\left|\hat{g}_{j0}\left(\frac{\hat{x}_j-\bar{\xi}_j}{\hat{\lambda}_{j0}^{1/2}}\right)-\bar{g}_{j0}\left(\frac{x_j}{\lambda_{j0}^{1/2}}\right)\right| + \sup_{x\in\mathcal{S}(c)}\left|\bar{g}_{j0}\left(\frac{x_j}{\lambda_{j0}^{1/2}}\right)-g_{j0}\left(\frac{x_j}{\lambda_{j0}^{1/2}}\right)\right| \\
&\quad= o_p\{(nh)^{-1/2}\}+O_p\left\{h+\left(\frac{nh}{\log n}\right)^{-1/2}\right\}=O_p\left\{h+\left(\frac{nh}{\log n}\right)^{-1/2}\right\}, \qquad (3)
\end{aligned}$$

where the first rate is due to Theorem 3.1 in Delaigle & Hall (2010), and the second to, for example, Theorem 2 in Stone (1983). Then

$$\begin{aligned}
\sup_{x\in\mathcal{S}(c)}|\hat{f}_{j0}(\hat{x}_j)-f_{j0}(x_j)| &= \sup_{x\in\mathcal{S}(c)}\left|\frac{1}{\hat{\lambda}_{j0}^{1/2}}\hat{g}_{j0}\left(\frac{\hat{x}_j-\bar{\xi}_j}{\hat{\lambda}_{j0}^{1/2}}\right)-\frac{1}{\lambda_{j0}^{1/2}}g_{j0}\left(\frac{x_j}{\lambda_{j0}^{1/2}}\right)\right| \\
&\le \sup_{x\in\mathcal{S}(c)}\left\{\frac{1}{\hat{\lambda}_{j0}^{1/2}}\left|\hat{g}_{j0}\left(\frac{\hat{x}_j-\bar{\xi}_j}{\hat{\lambda}_{j0}^{1/2}}\right)-g_{j0}\left(\frac{x_j}{\lambda_{j0}^{1/2}}\right)\right|+g_{j0}\left(\frac{x_j}{\lambda_{j0}^{1/2}}\right)\left|\frac{1}{\hat{\lambda}_{j0}^{1/2}}-\frac{1}{\lambda_{j0}^{1/2}}\right|\right\} \\
&= O_p\left\{\sup_{x\in\mathcal{S}(c)}\left|\hat{g}_{j0}\left(\frac{\hat{x}_j-\bar{\xi}_j}{\hat{\lambda}_{j0}^{1/2}}\right)-g_{j0}\left(\frac{x_j}{\lambda_{j0}^{1/2}}\right)\right|\right\}+O_p\left(\left|\frac{1}{\hat{\lambda}_{j0}^{1/2}}-\frac{1}{\lambda_{j0}^{1/2}}\right|\right) \\
&= O_p\left\{h+\left(\frac{nh}{\log n}\right)^{-1/2}\right\},
\end{aligned}$$

where the second equality follows from the consistency of $\hat{\lambda}_{j0}$ and Condition A4, and the third equality follows from (3) and the fact that $\hat{\lambda}_{j0}$ converges at a root-n rate. □

Proof of Theorem A1. For simplicity we consider the case where the supports of g_{j0} and g_{j1} are in common. The case where the supports differ can be proven in two step: First consider classifying elements x whose projections x_j are in the intersection of the supports of g_{j0} and g_{j1}; next consider classifying an element x for which a projection score x_j is not contained in the intersection of the supports, in which case $Q_J(x)$ will be $\pm\infty$, whence $\hat{Q}_J(x)$ will also diverge to $\pm\infty$, respectively, and thus consistency is obtained.

Now fix $\epsilon>0$. Set c be such that $\mathrm{pr}(||X||>c)=\mathrm{pr}\{X\notin\mathcal{S}(c)\}\le\epsilon/2$. First we prove there exists an event S such that $\hat{Q}_J(X)-Q_J(X)\to 0$ on S with $\mathrm{pr}(S)>1-\epsilon$. By Lemma 1 there exists $M_{jk}>0$ such that the event

$$S_{jk}=\left\{\sup_{x\in\mathcal{S}(c)}|\hat{f}_{jk}(\hat{x}_j)-f_{jk}(x_j)|\le M_{jk}\left\{h+\left(\frac{nh}{\log n}\right)^{-1/2}\right\}\right\}\quad(j=1,2,\ldots;\ k=0,1)$$

has probability $\mathrm{pr}(S_{jk})\ge 1-2^{-(j+2)}\epsilon$. Letting $S=\left(\bigcap_{j\ge1,k=0,1}S_{jk}\right)\cap\left[\bigcap_{j\ge1,k=0,1}\{\xi_j\in\mathrm{supp}(f_{jk})\}\right]\cap\{||X||\le c\}$, we have $\mathrm{pr}(S)\ge 1-\epsilon$, where supp means the support of a density. Let a_n be some increasing sequence such that $a_n\to\infty$ and

$a_n\{h + (nh/\log n)^{-1/2}\} = o(1)$. Define $\mathcal{U}_{jk} = \{x : x_j \in \mathrm{supp}(f_{jk})\}$, $\mathcal{U} = \bigcap_{j\geq 1, k=0,1} \mathcal{U}_{jk}$,

$$d_{jk} = \min\{1, \inf_{x\in\mathcal{S}(c)\cap\mathcal{U}} f_{jk}(x_j)\}, \quad J = \sup\left\{J' \geq 1 : \sum_{j\leq J',\, k=0,1} \frac{M_{jk}}{d_{jk}} \leq a_n\right\}.$$

The d_{jk} are bounded away from 0 by Condition A5, and J is nondecreasing and tends to infinity as $n \to \infty$. On S we have

$$\begin{aligned} \sum_{j=1}^{J} \frac{1}{d_{jk}} \sup_{x\in\mathcal{S}(c)} |\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)| &\leq \sum_{j=1}^{J} \frac{M_{jk}}{d_{jk}} \left\{h + \left(\frac{nh}{\log n}\right)^{-1/2}\right\} \\ &\leq a_n \left\{h + \left(\frac{nh}{\log n}\right)^{-1/2}\right\} = o(1), \end{aligned} \tag{4}$$

where the first and second inequalities are due to the property of S and J, respectively, and the last equality is by the definition of a_n.

From (4) we infer that on S,

$$\sup_{x\in\mathcal{S}(c)} |\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)| \leq d_{jk}/2 \tag{5}$$

eventually and uniformly for all $j \leq J$. Then on S it holds that

$$\begin{aligned} |\hat{Q}_J(X) - Q_J(X)| &\leq \sup_{x\in\mathcal{S}(c)\cap\mathcal{U}} |\hat{Q}_J(x) - Q_J(x)| \\ &\leq \sum_{j\leq J,\, k=0,1} \sup_{x\in\mathcal{S}(c)\cap\mathcal{U}} |\log \hat{f}_{jk}(\hat{x}_j) - \log f_{jk}(x_j)| \\ &\leq \sum_{j\leq J,\, k=0,1} \sup_{x\in\mathcal{S}(c)} |\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)| \frac{1}{\inf_{x\in\mathcal{S}(c)\cap\mathcal{U}} \eta_{3jk}} \\ &\leq \sum_{j\leq J,\, k=0,1} \sup_{x\in\mathcal{S}(c)} |\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)| \frac{2}{d_{jk}}, \\ &= o(1), \end{aligned}$$

where the third inequality is by Taylor's theorem, η_{3jk} is between $f_{jk}(x_j)$ and $\hat{f}_{jk}(\hat{x}_j)$, the last inequality is due to (5) which holds for large enough n, and the equality is due to (4). We conclude that $\mathrm{pr}(S \cap [I\{\hat{Q}_J(X) \geq 0\} \neq I\{Q_J(X) \geq 0\}]) \to 0$ as $n \to \infty$ by noting that $\hat{Q}_J(X)$ converges to $Q_J(X)$ and thus has the same sign as $Q_J(X)$ as $n \to \infty$. Notice that $Q_J(X)$ has a continuous density and thus $\mathrm{pr}\{Q_J(X) = 0\} = 0$ by Condition A4. $\square$

Proof of Theorem A2. Recall we assume $\hat\pi_1 = \hat\pi_0$. Then

$$\hat{E}(Y \mid \hat\xi_j = u) = \frac{\sum_{k=0}^{1}\sum_{i=1}^{n_k} kK(\frac{u-\hat\xi_{ijk}}{h_j})}{\sum_{k=0}^{1}\sum_{i=1}^{n_k} K(\frac{u-\hat\xi_{ijk}}{h_j})}$$

$$= \frac{\sum_{i=1}^{n_1} K(\frac{u-\hat\xi_{ij1}}{h_j})}{\sum_{i=1}^{n_1} K(\frac{u-\hat\xi_{ij1}}{h_j}) + \sum_{i=1}^{n_0} K(\frac{u-\hat\xi_{ij0}}{h_j})}$$

$$= \frac{\hat f_{j1}(u)}{\hat f_{j1}(u) + \hat f_{j0}(u)},$$

where $\hat f_{jk}$ are the kernel density estimators with bandwidth h_j, implying

$$\hat Q_J^R(x) = \sum_{j=1}^{J} \log\left[\frac{\hat E(Y \mid \hat\xi_j = \hat x_j)}{\{1 - \hat E(Y \mid \hat\xi_j = \hat x_j)\}}\right]$$

$$= \sum_{j=1}^{J} \log\left\{\frac{\hat f_{j1}(\hat x_j)}{\hat f_{j0}(\hat x_j)}\right\}.$$

Observe that $\hat Q_J^R$ has the same form as $\hat Q_J$, so this result follows from Theorem A1. □

4·2. *Theorem 1*

The proof of Theorem 1 requires the following key lemma, which is an extension of Lemma 1, changing the rate from $h + (nh/\log n)^{-1/2}$ to $h + (nh/\log n)^{-1/2} + (m^{2/5}h^2)^{-1}$. The remainder of the proof is omitted, since it is analogous to that of Theorem A1.

Lemma 2. *Under Conditions A1–A4 and A6–A9, for any $j \geq 1$, $k = 0, 1$,*

$$\sup_{x\in\mathcal{S}(c)} |\tilde f_{jk}(\tilde x_j) - f_{jk}(x_j)| = O_p\left\{h + \left(\frac{nh}{\log n}\right)^{-1/2} + (m^{2/5}h^2)^{-1}\right\}.$$

Proof. Given $x \in \mathcal{S}(c)$, by triangle inequality

$$|\tilde f_{jk}(\tilde x_j) - f_{jk}(x_j)| \leq |\tilde f_{jk}(\tilde x_j) - \hat f_{jk}(\hat x_j)| + |\hat f_{jk}(\hat x_j) - f_{jk}(x_j)|.$$

The rate for the second term can be derived from Lemma 1, so we focus only on the first term. For fixed j, k and $h_{jk} = h\lambda_{jk}^{1/2}$,

$$|\tilde f_{jk}(\tilde x_j) - \hat f_{jk}(\hat x_j)| = \frac{1}{n_k h_{jk}}\left|\sum_{i=1}^{n_k} K\left[\frac{\int_{\mathcal T}\{\tilde X_i^{(k)}(t) - x(t)\}\tilde\phi_j(t)\,\mathrm{d}t}{h_{jk}}\right] - K\left[\frac{\int_{\mathcal T}\{X_i^{(k)}(t) - x(t)\}\hat\phi_j(t)\,\mathrm{d}t}{h_{jk}}\right]\right|$$

$$\leq \frac{1}{n_k h_{jk}^2}\sum_{i=1}^{n_k}\left|\int_{\mathcal T}\{\tilde X_i^{(k)}(t) - x(t)\}\tilde\phi_j(t) - \{X_i^{(k)}(t) - x(t)\}\hat\phi_j(t)\,\mathrm{d}t\right| \cdot |K'(\eta_{4jk})|$$

$$\leq \frac{c_3}{n_k h^2}\sum_{i=1}^{n_k}\left|\int_{\mathcal T}\{\tilde X_i^{(k)}(t) - x(t)\}\tilde\phi_j(t) - \{X_i^{(k)}(t) - x(t)\}\hat\phi_j(t)\,\mathrm{d}t\right|, \tag{6}$$

for a constant $c_3 > 0$, where the first inequality is by Taylor's theorem, η_{4jk} is a mean value, and the last inequality is by Condition A4. The summand in (6) is

$$\begin{aligned}
&\left|\int_{\mathcal{T}} \{\tilde{X}_i^{(k)}(t) - x(t)\}\tilde{\phi}_j(t) - \{X_i^{(k)}(t) - x(t)\}\hat{\phi}_j(t)\,\mathrm{d}t\right| \\
=&\left|\int_{\mathcal{T}} \{\tilde{X}_i^{(k)}(t) - X_i^{(k)}(t)\}\tilde{\phi}_j(t) + \{X_i^{(k)}(t) - x(t)\}\{\tilde{\phi}_j(t) - \hat{\phi}_j(t)\}\,\mathrm{d}t\right| \\
\leq&\left|\int_{\mathcal{T}} \{\tilde{X}_i^{(k)}(t) - X_i^{(k)}(t)\}\tilde{\phi}_j(t)\,\mathrm{d}t\right| + \left|\int_{\mathcal{T}} \{X_i^{(k)}(t) - x(t)\}\{\tilde{\phi}_j(t) - \hat{\phi}_j(t)\}\,\mathrm{d}t\right| \\
&\leq ||\tilde{X}_i^{(k)} - X_i^{(k)}|| \cdot ||\tilde{\phi}_j|| + ||X_i^{(k)} - x|| \cdot ||\tilde{\phi}_j - \hat{\phi}_j|| \\
&\leq ||\tilde{X}_i^{(k)} - X_i^{(k)}|| + (||X_i^{(k)}|| + c)||\tilde{\phi}_j - \hat{\phi}_j||,
\end{aligned}$$

where the second and third inequalities follow from Cauchy–Schwarz inequality and from $||x|| \leq c$, respectively. Plugging the previous result into (6),

$$|\tilde{f}_{jk}(\tilde{x}_j) - \hat{f}_{jk}(\hat{x}_j)| \leq \frac{c_3}{h^2}\left\{\frac{1}{n_k}\sum_{i=1}^{n_k} ||\tilde{X}_i^{(k)} - X_i^{(k)}|| + ||\tilde{\phi}_j - \hat{\phi}_j||(\frac{1}{n_k}\sum_{i=1}^{n_k} ||X_i^{(k)}|| + c)\right\}. \tag{7}$$

Since $(\tilde{X}_i^{(k)}, X_i^{(k)})$ are identically distributed ($i = 1, \ldots, n_k$), and by Condition A6 the first term in the brackets has expected value equal to

$$E(||\tilde{X}_i^{(k)} - X_i^{(k)}||) = E_{X_i^{(k)}}\{E_{\varepsilon_i}(||\tilde{X}_i^{(k)} - X_i^{(k)}|| \mid X_i^{(k)})\} = O\{(mw)^{-1/2} + w^2\} = O(m^{-2/5}),$$

where more details about the second equality can be found in the Supplementary Material of Kong et al. (2016). Also $E(n_k^{-1}\sum_{i=1}^{n_k} ||X_i^{(k)}|| + c) = O(1)$ by Condition A1. So

$$\frac{1}{n_k}\sum_{i=1}^{n_k} ||\tilde{X}_i^{(k)} - X_i^{(k)}|| = O_p(m^{-2/5}), \quad \frac{1}{n_k}\sum_{i=1}^{n_k} ||X_i^{(k)}|| + c = O_p(1). \tag{8}$$

It remains to be shown $||\tilde{\phi}_j - \hat{\phi}_j|| = O_p(m^{-2/5})$. Let $\tilde{\Delta}_k = \tilde{G}_k - \hat{G}_k$ and for a square-integrable function $A(s,t)$ denote $||A||_F = \{\int_{\mathcal{T}}\int_{\mathcal{T}} A(s,t)^2\,\mathrm{d}s\,\mathrm{d}t\}^{1/2}$ be the Frobenius norm. In their Supplementary Material, Kong et al. (2016) show that $||\tilde{\Delta}_k||_F = O_p(m^{-2/5})$, so $||\tilde{\Delta}||_F = ||\tilde{\Delta}_0 + \tilde{\Delta}_1||_F/2 = O_p(m^{-2/5})$. By standard perturbation theory for operators (Bosq, 2000), for a fixed j

$$||\tilde{\phi}_j - \hat{\phi}_j|| = O(||\tilde{\Delta}||_F / \sup_{k \neq j} |\hat{\lambda}_j - \hat{\lambda}_k|) = O_p(m^{-2/5}). \tag{9}$$

Inserting (8) and (9) into (7) we arrive at the conclusion. □

4·3. *Theorem 2*

Assuming X is Gaussian under $k = 0, 1$, whence the criterion function $Q_J(x)$ defined in (3) in the main text becomes

$$Q_J^G(x) = \frac{1}{2}\sum_{j=1}^{J}\left[(\log\lambda_{j0} - \log\lambda_{j1}) - \left\{\frac{1}{\lambda_{j1}}(x_j - \mu_j)^2 - \frac{1}{\lambda_{j0}}x_j^2\right\}\right] > 0. \tag{10}$$

Letting $\zeta_j = \xi_j/\lambda_{j0}^{1/2}$, then

$$\zeta_j \sim N(0,1) \text{ under } \Pi_0, \quad \zeta_j \sim N(m_j, r_j^{-1}) \text{ under } \Pi_1,$$

$$Q_J^G(X) = \frac{1}{2}\sum_{j=1}^{J}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\}.$$

Under Gaussian assumptions, our Bayes classifier is a special case of the quadratic discriminant, a non-Bayes classifier because it uses two different sets of projections. The perfect classification properties for the functional quadratic discriminant were discussed in Delaigle & Hall (2013) in the context of truncated functional observations or fragments. We use the following auxiliary result.

LEMMA 3. *Assume the predictors come from Gaussian processes. If $\sum_{j=1}^{\infty} m_j^2 < \infty$ and $\sum_{j=1}^{\infty}(r_j - 1)^2 < \infty$, then $Q_J^G(X)$ converges almost surely to a random variable as $J \to \infty$, in which case perfect classification does not occur. Otherwise perfect classification occurs.*

This lemma is similar to Theorem 1 of Delaigle & Hall (2013), but uses more transparent conditions and a proof technique based on the optimality property of Bayes classifiers which will be reused in the proof of Theorem 2. Lemma 3 states perfect classification occurs for Gaussian processes if and only if there are sufficient differences between the two groups in the mean or the covariance functions. This perfect classification phenomenon arises for the non-degenerate infinite dimensional case because we have infinitely many independent projection scores ξ_j for classification.

Proof of Lemma 3. Case 1: Assume $\sum_{j=1}^{\infty}(r_j - 1)^2 = \infty$ and that there exists a subsequence r_{j_l} of r_j that goes to ∞ or 0 as $l \to \infty$. Correspondingly take a subsequence $r_{j_l} \to \infty, r_{j_l} > 1$ or $r_{j_l} \to 0, r_{j_l} < 1$ for all $l = 1, 2, \ldots$. Denoting the summand $(\log \lambda_{j0} - \log \lambda_{j1}) - \{(\xi_j - \mu_{j1})^2/\lambda_{j1} - \xi_j^2/\lambda_{j0}\}$ of (10) as S_j^G, for any $j \le J$ the misclassification rate $\mathrm{pr}[I\{Q_J^G(X) \ge 0\} \ne Y]$ is smaller than or equal to $\mathrm{pr}\{I(S_j^G \ge 0) \ne Y\}$, since the former is the Bayes classifier using the first J projections, which minimizes the misclassification error among the class. Thus the misclassification rate of $Q_J^G(X)$ is bounded above by that of the classifier $I\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2 \ge 0\}$ for any $j \le J$. Let pr_{Π_k} denote the conditional probability measure under group k. If there exists $r_{j_l} \to 0$,

$$\mathrm{pr}_{\Pi_0}\{\log r_{j_l} - r_{j_l}(\zeta_{j_l} - m_{j_l})^2 + \zeta_{j_l}^2 \ge 0\} \le \mathrm{pr}_{\Pi_0}(\log r_{j_l} + \zeta_{j_l}^2 \ge 0) \to 0,$$

observing $\zeta_{j_l}^2 \sim \chi_1^2$ under Π_0 and $\log r_{j_l} \to -\infty$.

If there exist $r_{j_l} \to \infty$, then there exists a sequence $M \to \infty$ such that $(\log r_{j_l} + M)/r_{j_l} \to 0$. For any $j = 1, 2, \ldots,$

$$\begin{aligned}\mathrm{pr}_{\Pi_0}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2 \ge 0\} &\le \mathrm{pr}_{\Pi_0}\{\log r_j - r_j(\zeta_j - m_j)^2 + M \ge 0\} + \mathrm{pr}_{\Pi_0}(\zeta_j^2 > M) \\ &= \mathrm{pr}_{\Pi_0}\{(\zeta_j - m_j)^2 \le \frac{\log r_j + M}{r_j}\} + o(1) \\ &= \mathrm{pr}_{\Pi_0}\left\{|\zeta_j - m_j| \le \left(\frac{\log r_j + M}{r_j}\right)^{1/2}\right\} + o(1).\end{aligned} \tag{11}$$

Plugging the sequence r_{j_l} for r_j into (11) we have $\{(\log r_j + M)/r_j\}^{1/2} \to 0$ as $l \to \infty$ and $M \to \infty$. Since ζ_j are standard normal and thus have uniformly bounded densities, (11) goes to zero and we have $\mathrm{pr}_{\Pi_0}\{\log r_{j_l} - r_{j_l}(\zeta_{j_l} - m_{j_l})^2 + \zeta_{j_l}^2 \geq 0\} \to 0$ as $l \to \infty$ and $M \to \infty$. Using similar arguments we can also prove $\mathrm{pr}_{\Pi_1}\{\log r_{j_l} - r_{j_l}(\zeta_{j_l} - m_{j_l})^2 + \zeta_{j_l}^2 < 0\} \to 0$ as $l \to \infty$. By Bayes' theorem $\mathrm{pr}\{I(S_{j_l}^G \geq 0) \neq Y\} = \mathrm{pr}(Y = 0)\mathrm{pr}(S_{j_l}^G \geq 0 \mid Y = 0) + \mathrm{pr}(Y = 1)\mathrm{pr}(S_{j_l}^G < 0 \mid Y = 1) \to 0$ as $l \to \infty$. Therefore

$$\mathrm{pr}[I\{Q_J^G(X) \geq 0\} \neq Y] \leq \mathrm{pr}\{I(S_{j_l}^G \geq 0) \neq Y\} \to 0, \quad J \to \infty,$$

which means perfect classification occurs.

Case 2: Assume $\sum_{j=1}^{\infty}(r_j - 1)^2 = \infty$, and there exists M_1 and M_2 such that $0 < M_1 \leq r_j \leq M_2 < \infty$ for all $j \geq 1$. Letting E_{Π_k} and var_{Π_k} be the conditional expectation and variance under group k, respectively, we have

$$\begin{aligned}
E_{\Pi_0}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} &= \log r_j - (r_j - 1) - m_j^2 r_j, \\
E_{\Pi_1}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} &= -\log r_j^{-1} + (r_j^{-1} - 1) + m_j^2, \\
\mathrm{var}_{\Pi_0}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} &= 2(1 - r_j)^2 + 4m_j^2 r_j^2, \\
\mathrm{var}_{\Pi_1}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} &= 2(r_j^{-1} - 1)^2 + 4m_j^2 r_j^{-1}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{pr}_{\Pi_0}\left(\sum_{j=1}^{J}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} \geq 0\right) &\leq \frac{\sum_{j=1}^{J}\{2(1 - r_j)^2 + 4m_j^2 r_j^2\}}{\{-\sum_{j=1}^{J}(r_j - 1 - \log r_j + m_j^2 r_j)\}^2} \\
&\leq \frac{\sum_{j=1}^{J}\{2(1 - r_j)^2 + 4M_2^2 m_j^2\}}{[\sum_{j=1}^{J}\{\frac{1}{M_2}(r_j - 1)^2 + M_1 m_j^2\}]^2} \\
&= \frac{4M_2^2/M_1}{\sum_{j=1}^{J}\{\frac{1}{M_2}(r_j - 1)^2 + M_1 m_j^2\}} \times \frac{\sum_{j=1}^{J}\{2(1 - r_j)^2 + 4M_2^2 m_j^2\}}{\sum_{j=1}^{J}\{4\frac{M_2}{M_1}(r_j - 1)^2 + 4M_2^2 m_j^2\}} \\
&\leq \frac{4M_2^2/M_1}{\sum_{j=1}^{J}\{\frac{1}{M_2}(r_j - 1)^2 + M_1 m_j^2\}} \to 0, \quad J \to 0,
\end{aligned}$$

where Chebyshev's inequality is used for the first inequality, and Taylor expansion in the second inequality. Analogously the misclassification rate under Π_1 also can be proven to go to zero.

Case 3: Assume $\sum_{j=1}^{\infty}(r_j - 1)^2 < \infty$ and $\sum_{j=1}^{\infty} m_j^2 = \infty$. The proof is essentially the same as in Case 2.

Case 4: Assume $\sum_{j=1}^{\infty}(r_j - 1)^2 < \infty$ and $\sum_{j=1}^{\infty} m_j^2 < \infty$. Then the mean and variance of $Q_J^G(X)$ converges, so $Q_J^G(X)$ converges to a random variable under either population by Billingsley (1995). Therefore misclassification does not occur. □

We can then proceed to prove Theorem 2, which does not assume Gaussianity.

Proof of Theorem 2. Case 1: Assume $\sum_{j=1}^{\infty}(r_j - 1)^2 = \infty$ and there exists a subsequence r_{j_l} of r_j that goes to 0 or ∞ as $l \to \infty$. By the optimality of Bayes classifiers, the Bayes classifier $I\{Q_J(X) \geq 0\}$ using the first J components has smaller misclassification error than that of $I(S_j \geq 0)$, where S_j is the jth component in the summand of (3) in the main text, for all $j \leq J$. Since $I(S_j \geq 0)$ is the Bayes classifier using only the jth projection, it has a smaller misclassification error than the non-Bayes classifier $I(S_j^G \geq 0)$, where $S_j^G = \log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2$ is the jth summand in (10). Under Conditions A10–A11, we prove that the misclassification error converges to zeros by adopting the same argument as in Lemma 3 Case 1.

Case 2: Assume $\sum_{j=1}^{\infty}(r_j - 1)^2 = \infty$, and there exists M_1 and M_2 such that $0 < M_1 \leq r_j \leq M_2 < \infty$ for all $j \geq 1$. By some algebra,

$$
\begin{aligned}
E_{\Pi_0}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} &= \log r_j - (r_j - 1) - m_j^2 r_j, \\
E_{\Pi_1}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} &= -\log r_j^{-1} + (r_j^{-1} - 1) + m_j^2, \\
\mathrm{var}_{\Pi_0}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} &\leq (2C_M - 1)(1 - r_j)^2 + 4(C_M + 1)m_j^2 r_j^2, \\
\mathrm{var}_{\Pi_1}\{\log r_j - r_j(\zeta_j - m_j)^2 + \zeta_j^2\} &\leq (2C_M - 1)(r_j^{-1} - 1)^2 + 4(C_M + 1)m_j^2 r_j^{-1}.
\end{aligned}
$$

The expectations are the same as in the Gaussian case because the first two moments of ζ_j do not depend on distributional assumptions. The inequalities in the variance calculation are due to $2ab \leq a^2 + b^2$ for all $a, b \in \mathbb{R}$. The same Chebyshev's inequality argument can be applied as for Theorem A1.

Case 3: Assume $\sum_{j=1}^{\infty}(r_j - 1)^2 < \infty$ and $\sum_{j=1}^{\infty} m_j^2 = \infty$. The proof is essentially the same as that for Case 2. □

REFERENCES

BILLINGSLEY, P. (1995). *Probability and Measure*. New York: John Wiley & Sons Inc., 3rd ed.

BOSQ, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. New York: Springer-Verlag.

DELAIGLE, A. & HALL, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics* **38**, 1171–1193.

DELAIGLE, A. & HALL, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association* **108**, 1269–1283.

KONG, D., XUE, K., YAO, F. & ZHANG, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103**, 147–159.

NADARAYA, E. (1964). On estimating regression. *Theory of Probability and Its Applications* **9**, 141–142.

STONE, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent Advances in Statistics*, M. H. Rizvi, J. S. Rustagi & D. Siegmund, eds. pp. 393–406.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Series A* **26**, 359–372.