

UNEXPECTED PROPERTIES OF BANDWIDTH CHOICE WHEN SMOOTHING DISCRETE DATA FOR CONSTRUCTING A FUNCTIONAL DATA CLASSIFIER

BY RAYMOND J. CARROLL¹, AURORE DELAIGLE² AND PETER HALL²

*Texas A&M University, University of Melbourne and
University of Melbourne*

The data functions that are studied in the course of functional data analysis are assembled from discrete data, and the level of smoothing that is used is generally that which is appropriate for accurate approximation of the conceptually smooth functions that were not actually observed. Existing literature shows that this approach is effective, and even optimal, when using functional data methods for prediction or hypothesis testing. However, in the present paper we show that this approach is not effective in classification problems. There a useful rule of thumb is that undersmoothing is often desirable, but there are several surprising qualifications to that approach. First, the effect of smoothing the training data can be more significant than that of smoothing the new data set to be classified; second, undersmoothing is not always the right approach, and in fact in some cases using a relatively large bandwidth can be more effective; and third, these perverse results are the consequence of very unusual properties of error rates, expressed as functions of smoothing parameters. For example, the orders of magnitude of optimal smoothing parameter choices depend on the signs and sizes of terms in an expansion of error rate, and those signs and sizes can vary dramatically from one setting to another, even for the same classifier.

1. Introduction. All supposedly “functional” data are actually observed discretely, sometimes on a grid and on other occasions at randomly scattered points. For example, in longitudinal data analysis the observation points are

Received March 2013; revised July 2013.

¹Supported by a Grant from the National Cancer Institute (R37-CA057030). This publication is based in part on work supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

²Supported by grants and fellowships from the Australian Research Council.

AMS 2000 subject classifications. 62G08.

Key words and phrases. Centroid method, discrimination, kernel smoothing, quadratic discrimination, smoothing parameter choice, training data.

<p>This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in <i>The Annals of Statistics</i>, 2013, Vol. 41, No. 6, 2739–2767. This reprint differs from the original in pagination and typographic detail.</p>
--

often widely spaced and irregularly placed, and substantial smoothing is commonly used to convert discrete data like these to functions. The impact of such smoothing has been addressed in the context of prediction or hypothesis testing for functional data; see, for example, Hall and Van Keilegom (2007), Panaretos, Kraus and Maddocks (2010), Wu and Müller (2011), Benhenni and Degras (2011), Cardot and Josserand (2011) and Cardot, Degras and Josserand (2013). The main conclusion of these papers has been that conventional rules for smoothing discrete data typically apply, and that smoothing parameters of standard size generally are appropriate.

In contrast, the present paper was motivated by numerical work indicating that, in the context of classifying functional data, smoothing parameters of highly nonstandard sizes are appropriate, and more generally that, even for a relatively simple classifier, there is no simple precept (even an asymptotic prescription of size) that leads to minimisation of error rate. If one had to give a rule, valid in some but by no means all cases, it would be to undersmooth, but even there unexpected caveats must be addressed.

For example, it turns out that the impact of smoothing the training data can be more significant than that of smoothing the new data to be classified. Indeed, the effect of smoothing the new data is characteristic of a parametric problem, rather than a nonparametric one. There, asymptotic arguments indicate that $(\text{sample size})^{-1/2}$ is an appropriate bandwidth size for reducing the impact of smoothing to parametric levels, whereas $(\text{sample size})^{-1/3}$ is the nearest analogue for smoothing the training data.

However, both these recommendations are incorrect in many cases. Depending on the signs and sizes of certain functionals of the data distributions, it can be optimal to use smoothing parameters that are an order of magnitude smaller, or an order of magnitude larger, than these. Using some viewpoints the need for a low level of smoothing is intuitively clear. Indeed, we expect that relatively minor features of a curve, of the sort that might disappear if we were too enthusiastic in the smoothing step, could have important information to convey in a classification analysis. On the other hand, our results show that very high levels of smoothing are sometimes advantageous.

We drew these conclusions after studying three different classifiers for functional data: the standard centroid-based method, the scale-adjusted form of that approach, and a version for functional data of quadratic discrimination. Our conclusions are valid for all three approaches, although they contradict conclusions which are well known for standard nonparametric approximations to the Bayes classifier in multivariate, rather than functional-data, settings. Specifically, for univariate and functional data, and nonparametric Bayes classifiers, conventional smoothing parameters, for example, those chosen using standard plug-in rules for function estimation, typically are of the correct order even though they do not quite minimise asymptotic

classification error; see, for example, Hall and Kang (2005). Moreover, there does not exist a version of our results in univariate or multivariate settings, since there is no analogue in such cases of the “lattice effect,” represented by the m_{kj} ’s.

To these comments, we should add that in practice there is relatively little difficulty in choosing smoothing parameters to minimise error rate; cross-validation is usually effective. Our aim in this paper is therefore not to develop methods for choosing the bandwidth optimally, or nearly optimally, in classification problems, but to provide an understanding of the many aspects of those problems that conspire together to determine the optimal choice.

2. Model and methodology.

2.1. *Model.* We consider n_0 (resp., n_1) unknown random functions $\{g_{0j}, 1 \leq j \leq n_0\}$ (resp., $\{g_{1j}, 1 \leq j \leq n_1\}$) coming from two populations. We observe a training sample of the data pairs $\mathcal{D}_{kj} = \{(X_{kji}, Y_{kji}), 1 \leq i \leq m_{kj}\}$, for $1 \leq j \leq n_k$ and $k = 0, 1$, corresponding to noisy versions of the g_{kj} ’s sampled at a discrete set of random points (i.e., X_{kji} ’s) and generated by the model

$$(2.1) \quad Y_{kji} = g_{kj}(X_{kji}) + \varepsilon_{kji},$$

where k indexes the population, Π_k , from which the data in \mathcal{D}_{kj} came, j denotes the index of an individual drawn from Π_k , and i is the index of a data pair (X_{kji}, Y_{kji}) for the j th individual from the k th population.

The g_{kj} are random functions defined on a compact interval \mathcal{I} , but observed only at m_{kj} points $X_{kj1}, \dots, X_{kjm_{kj}}$. These points may be fixed or random, and although we shall develop our arguments in the random case, they can easily be extended to the fixed case. We assume that each g_{kj} has two bounded derivatives on \mathcal{I} ; the respective sequences of X ’s and ε ’s are each identically distributed with distributions that do not depend on the g ’s; the g ’s, X ’s and ε ’s are all mutually independent; the X ’s are supported on \mathcal{I} ; and the ε ’s have zero mean and finite variance.

We also observe a new data set $\mathcal{D} = \{(X_i, Y_i), 1 \leq i \leq m\}$, similar to the \mathcal{D}_{kj} ’s except that in this case we do not know which population the data come from. Here,

$$(2.2) \quad Y_i = g(X_i) + \varepsilon_i,$$

where the function g , the X ’s and the ε ’s have the properties given in the previous paragraph. Using the training data, we wish to determine whether \mathcal{D} came from Π_0 or Π_1 .

In the functional data literature [see, e.g., Ramsay and Silverman (2005)], when the data are noisy, it is common to preprocess them prior to further

analysis. Typically, this is done by smoothing the data in some way, for example, through a spline or kernel smoother, thereby obtaining, from the data in \mathcal{D}_{kj} and \mathcal{D} , estimators \widehat{g}_{kj} and \widehat{g} of g_{kj} and g , respectively. In the classification context, once these estimators have been derived, they are plugged into functional data classifiers, replacing there the unobserved functions g and g_{kj} by their estimators \widehat{g} and \widehat{g}_{kj} . Our aim in this paper is to describe the application of estimators \widehat{g} and \widehat{g}_{kj} of g and g_{kj} , and in particular to describe the influence of tuning parameters used to construct them, when the aim is classification rather than just function estimation.

2.2. *Estimating g , g_{kj} and their mean and covariance functions.* There are several ways to obtain nonparametric estimators of the functions g and g_{kj} , but the most popular ones are spline and local linear methods. They have similar properties, but since local linear estimators are much more tractable theoretically, we shall use these in this work. For $x \in \mathcal{I}$, the local linear estimators of g and g_{kj} are defined by

$$(2.3) \quad \begin{aligned} \widehat{g}(x) &= \frac{U_2(x)V_0(x) - U_1(x)V_1(x)}{U_2(x)U_0(x) - U_1^2(x)}, \\ \widehat{g}_{kj}(x) &= \frac{U_{kj2}(x)V_{kj0}(x) - U_{kj1}(x)V_{kj1}(x)}{U_{kj2}(x)U_{kj0}(x) - U_{kj1}^2(x)}, \end{aligned}$$

where

$$(2.4) \quad U_\ell(x) = \frac{1}{m} \sum_{i=1}^m \left(\frac{x - X_i}{h} \right)^\ell K_h(x - X_i),$$

$$(2.5) \quad V_\ell(x) = \frac{1}{m} \sum_{i=1}^m Y_i \left(\frac{x - X_i}{h} \right)^\ell K_h(x - X_i),$$

$$(2.6) \quad U_{kj\ell}(x) = \frac{1}{m_{kj}} \sum_{i=1}^{m_{kj}} \left(\frac{x - X_{kji}}{h_1} \right)^\ell K_{h_1}(x - X_{kji}),$$

$$V_{kj\ell}(x) = \frac{1}{m_{kj}} \sum_{i=1}^{m_{kj}} Y_{kji} \left(\frac{x - X_{kji}}{h_1} \right)^\ell K_{h_1}(x - X_{kji}),$$

K is a kernel function, $h > 0$ and $h_1 > 0$ are bandwidths, and $K_h(x) = K(x/h)/h$. See, for example, Fan and Gijbels (1996). For simplicity, throughout we use the same bandwidth h_1 for each population and each individual, but we could have replaced h_1 by bandwidths that depended on k and j , as we do in our numerical work.

The classifiers we consider in this work require estimators of the population means and covariances. For $k = 0, 1$, let μ_k denote the mean function

$$(2.7) \quad \mu_k = E_k(g) = E_k(g_{kj}),$$

where E_k represents expectation under the assumption that the data come from Π_k . Also, let G_k be the covariance function, defined by $G_k(u, v) = \text{cov}_k\{g(u), g(v)\} = E_k\{g(u)g(v)\} - \mu_k(u)\mu_k(v)$, where cov_k denotes covariance when the data come from Π_k . Estimators $\hat{\mu}_k$ and \hat{G}_k of μ_k and G_k are defined in the standard way by the empirical mean and covariance functions, but replacing, in the definitions of these estimators, the unobserved g_{kj} by \hat{g}_{kj} :

$$(2.8) \quad \hat{\mu}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \hat{g}_{kj},$$

$$(2.9) \quad \hat{G}_k(u, v) = \frac{1}{n_k} \sum_{j=1}^{n_k} \{\hat{g}_{kj}(u) - \hat{\mu}_k(u)\} \{\hat{g}_{kj}(v) - \hat{\mu}_k(v)\}.$$

See, for example, Ramsay and Silverman (2005), Chapter 2.

Consider the spectral decomposition of the covariance function

$$(2.10) \quad G_k(u, v) = \sum_{\ell=1}^{\infty} \theta_{k\ell} \psi_{k\ell}(u) \psi_{k\ell}(v),$$

where $(\theta_{k\ell}, \psi_{k\ell})$ is an (eigenvalue, eigenfunction) pair for the linear operator G_k defined by $G_k(\psi)(u) = \int G_k(u, v)\psi(v) dv$, and where, following convention, we have used the notation G_k for both the operator and the covariance. The terms in (2.10) are ordered such that $\theta_{k1} \geq \theta_{k2} \geq \dots \geq 0$. If g is drawn from Π_k then we can write

$$(2.11) \quad g(x) = \mu_k(x) + \sum_{\ell=1}^{\infty} Z_{k\ell} \theta_{k\ell}^{1/2} \psi_{k\ell}(x),$$

where $\mu_k = E_k(g)$ denotes the mean of the random process of which g is a realisation, $Z_{k\ell} = \theta_{k\ell}^{-1/2} \int (g - \mu_k) \psi_{k\ell}$, and the $Z_{k\ell}$'s (for $\ell = 1, 2, \dots$) comprise a sequence of uncorrelated random variables with zero mean and unit variance. The quantities $\theta_{k\ell}$ and $\psi_{k\ell}$ can be estimated consistently by the eigenvalues and eigenfunctions $\hat{\theta}_{k\ell}$ and $\hat{\psi}_{k\ell}$ of the linear operator \hat{G}_k , defined by $\hat{G}_k(\psi)(u) = \int \hat{G}_k(u, v)\psi(v) dv$, with the covariance estimator \hat{G}_k defined as at (2.9):

$$(2.12) \quad \hat{G}_k(u, v) = \sum_{\ell=1}^{\infty} \hat{\theta}_{k\ell} \hat{\psi}_{k\ell}(u) \hat{\psi}_{k\ell}(v),$$

where $\hat{\theta}_{k1} \geq \hat{\theta}_{k2} \geq \dots \geq 0$, and, since $\hat{\theta}_{k\ell} = 0$ for all $\ell > n_k$, all but the first n_k terms in the series at (2.12) vanish. See Hall and Hosseini-Nasab (2006, 2009) for properties of these estimators in the case where g and g_{kj} are observed; see also Li and Hsing (2010a, 2010b) for other cases.

2.3. *Constructing classifiers.* Classifiers for functional data have received a great deal of attention in the literature. See, for example, Vilar and Pértiga (2004), Biau, Bunea and Wegkamp (2005), Fromont and Tuleau (2006), Leng and Müller (2006), López-Pintado and Romo (2006), Rossi and Villa (2006), Cuevas, Febrero and Fraiman (2007), Wang, Ray and Mallick (2007), Berlinet, Biau and Rouvière (2008), Epifanio (2008), Araki et al. (2009), Delaigle and Hall (2012) and Delaigle, Hall and Bathia (2012).

In those papers the authors suggest methods for constructing classifiers, but so far the theoretical impact of smoothing; that is, the impact of using \widehat{g} and \widehat{g}_{kj} instead of g and g_{kj} when constructing classifiers; has been largely ignored in the literature. In this paper, we study this impact of smoothing for three relatively simple functional classifiers: the centroid classifier, or Rocchio classifier [see, e.g., Manning, Raghavan and Schütze (2008)], commonly used for classifying high-dimensional data; a scaled version of this classifier, which we define below in a general way; and a version for functional data of Fisher's quadratic discriminant, studied, for example, by Leng and Müller (2006) and Delaigle and Hall (2012). These classifiers are usually defined in terms of the functions g and g_{kj} , and here we shall define them in terms of \widehat{g} and \widehat{g}_{kj} . The standard versions of these classifiers are obtained by replacing \widehat{g} and \widehat{g}_{kj} by g and g_{kj} . The functions \widehat{g}_{kj} appear only implicitly through the estimated means and covariance functions constructed in Section 2.2.

In the present setting, the centroid-based classifier assigns the curve g , observed through \mathcal{D} , to Π_0 if the statistic

$$(2.13) \quad S(\widehat{g}) = \int_{\mathcal{I}} \{\widehat{g}(t) - \widehat{\mu}_0(t)\}^2 dt - \int_{\mathcal{I}} \{\widehat{g}(t) - \widehat{\mu}_1(t)\}^2 dt$$

is negative, and to Π_1 if $S(\widehat{g}) > 0$.

A scaled version of the centroid classifier, which accommodates differences in scales between the two populations, can be defined by replacing S in (2.13) by

$$(2.14) \quad S_{\text{scale}}(\widehat{g}) = \frac{1}{s_0^2} \int_{\mathcal{I}} \{\widehat{g}(t) - \widehat{\mu}_0(t)\}^2 dt - \frac{1}{s_1^2} \int_{\mathcal{I}} \{\widehat{g}(t) - \widehat{\mu}_1(t)\}^2 dt + \log\left(\frac{s_0^2}{s_1^2}\right),$$

where s_k^2 is an estimator of the scale of population Π_k . For example, we might take s_k^2 to equal $n_k^{-1} \sum_{j=1}^{n_k} \int_{\mathcal{I}} (\widehat{g}_{kj} - \widehat{\mu}_k)^2$, the version we used in our numerical work, or $\int_{\mathcal{I}} \int_{\mathcal{I}} \widehat{G}_k(u, v) \psi(u) \psi(v) du dv$, where ψ is open to choice; or s_0^2 and s_1^2 could be selected empirically by minimising a cross-validation estimator of classification error. The definition at (2.14) should be compared with those at (2.15) and (2.16), below. The form of (2.14), and also of (2.15) and (2.16), is motivated by likelihood-ratio statistics for Gaussian data.

A version for functional data of Fisher's quadratic discriminant is based on

$$(2.15) \quad T(\widehat{g}) = \sum_{\ell=1}^p \left[\frac{1}{\widehat{\theta}_{0\ell}} \left\{ \int_{\mathcal{I}} (\widehat{g} - \widehat{\mu}_0) \widehat{\psi}_{0\ell} \right\}^2 - \frac{1}{\widehat{\theta}_{1\ell}} \left\{ \int_{\mathcal{I}} (\widehat{g} - \widehat{\mu}_1) \widehat{\psi}_{1\ell} \right\}^2 + \log \left(\frac{\widehat{\theta}_{0\ell}}{\widehat{\theta}_{1\ell}} \right) \right],$$

where \widehat{g} and $\widehat{\mu}_k$ are as at (2.3) and (2.8), $(\widehat{\theta}_{0\ell}, \widehat{\theta}_{1\ell})$ are at (2.12) and p is a positive truncation parameter. (Here we assume, as is often the case in practice, that the prior probabilities of each population are unknown and estimated by $1/2$. A more general version of the classifier can be used if these probabilities are estimated by other values, but this does not alter our main conclusions.) We assign the new data set \mathcal{D} to Π_0 if $T(\widehat{g}) \leq 0$, and to Π_1 otherwise. Of course, the statistic $T(\widehat{g})$, at (2.15), is just an empirical version of the quantity

$$(2.16) \quad T_0(g) = \sum_{\ell=1}^p \left[\frac{1}{\theta_{0\ell}} \left\{ \int_{\mathcal{I}} (g - \mu_0) \psi_{0\ell} \right\}^2 - \frac{1}{\theta_{1\ell}} \left\{ \int_{\mathcal{I}} (g - \mu_1) \psi_{1\ell} \right\}^2 + \log \left(\frac{\theta_{0\ell}}{\theta_{1\ell}} \right) \right].$$

If the functions g are Gaussian, and the first p eigenvalues, in versions of (2.10) and (2.12) for either population, are distinct and nonzero, and the remaining eigenvalues vanish, then the classifier based on $T_0(g)$, at (2.16), is optimal in the sense of having least classification error among all classifiers, since it is, after all, just a likelihood ratio statistic. When the eigenvalues and eigenfunctions are estimated from data, as at (2.15), the classifier is asymptotically optimal. Bearing in mind the effectiveness of Fisher's discriminant analysis in the case of vector-valued data, even when the data are not normal, the classifier based on $T(\widehat{g})$ is an attractive choice even in non-Gaussian cases.

3. Theoretical properties.

3.1. *Standard centroid-based classifier.* In this section, we derive properties of the centroid classifier based on the estimators \widehat{g} and \widehat{g}_{kj} , and in particular we examine the impact of smoothing. First, we introduce notation. Let $n = n_0 + n_1$ (hence n is a positive integer sequence diverging to infinity), let $m = m(n)$ be of the same size as m_{kj} [see (3.7) below], and write $\sigma_{\varepsilon_k}^2$ for the variance of the experimental errors ε_{kji} and ε_i , in (2.1) and

(2.2), when the data come from Π_k . Let

$$(3.1) \quad \bar{g}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} g_{kj}, \quad \nu_k = n_k^2 \left(\sum_{j=1}^{n_k} m_{kj}^{-1} \right)^{-1}$$

and define

$$(3.2) \quad b_{k0} = \int_{\mathcal{I}} (\mu_1 - \mu_0) \{2\mu_k - (\mu_0 + \mu_1)\},$$

$$\beta_{k0} = \int_{\mathcal{I}} (\bar{g}_1 - \bar{g}_0) \{2\mu_k - (\bar{g}_0 + \bar{g}_1)\},$$

$$(3.3) \quad \sigma_k^2 = 4\kappa_2 \int_{\mathcal{I}} \int_{\mathcal{I}} (\bar{g}_1 - \bar{g}_0)(x_1)(\bar{g}_1 - \bar{g}_0)(x_2) G_k(x_1, x_2) dx_1 dx_2,$$

$$(3.4) \quad \tau_k^2 = 4\kappa_2 \int_{\mathcal{I}} \int_{\mathcal{I}} (\mu_1 - \mu_0)(x_1)(\mu_1 - \mu_0)(x_2) G_k(x_1, x_2) dx_1 dx_2,$$

where $\kappa_2 = \int u^2 K(u) du$. Finally, put $\kappa = \int K^2$, and let \mathcal{I} be the compact interval that equals the support of the density f_X of the X_i 's and X_{kji} 's, and of the functions g and g_{kj} .

We make the following assumptions:

- (a) The distribution of the experimental errors ε_{kji} and ε_i , in (2.1) and (2.2), has zero mean and all moments finite, may depend on k , and has variance $\sigma_{\varepsilon k}^2$; (b) the density f_X of the variables X_{kji} and X_i does not depend on i, j or k ; (c) f_X has two bounded derivatives, $f_X(x) \geq C > 0$ for all $x \in \mathcal{I}$, and f_X'' is Hölder continuous on the support \mathcal{I} of f_X .

- (a) The functions g and g_{kj} associated with the populations Π_k , for $k = 0, 1$, are realisations of Gaussian processes, have uniformly bounded covariance functions G_k and mean functions μ_k , both depending only on k , and satisfy $\tau_k^2 > 0$ for $k = 0, 1$; and (b) with probability 1 the functions are uniformly bounded and have Hölder-continuous second derivatives, with the property that, for a constant $C > 0$, all moments of $\sup_{x_1, x_2} |g''(x_1) - g''(x_2)| / |x_1 - x_2|^C$ are finite when g is sampled from either Π_0 or Π_1 .

- (a) For a constant $C > 0$, the results $h^{(1)} = O(n^{-C})$ and $n^{1-C} h^{(1)} \rightarrow \infty$ hold for $h^{(1)} = h$ and $h^{(1)} = h_1$; (b) the kernel K is a symmetric, nonnegative, compactly supported and Hölder continuous probability density; and (c) for each k the values of $m^{-1} \min_j m_{kj}$, $m^{-1} \max_j m_{kj}$ and n_0/n_1 are bounded away from zero and infinity as $n \rightarrow \infty$, and, for constants C_1 and C_2 satisfying $0 < C_1 < C_2 < \infty$, m and n_0 lie between n^{C_1} and n^{C_2} .

The assumption in (3.5)(c) that f_X is bounded away from zero on its support is only a technical requirement, and is unnecessary in practice. To make this clear, in our numerical work we shall take f_X to be a normal density, and show that the conclusions of Theorem 1 are nevertheless reflected clearly.

Let $\text{err}_k = P_k\{(-1)^k S(\hat{g}) > 0\}$ denote the probability that the standard centroid-based classifier, based on the statistic $S(\hat{g})$ at (2.13), commits an error when the data set \mathcal{D} actually comes from Π_k . Theorem 1 below describes the asymptotic behaviour of err_k , and highlights the effect of the smoothing parameters h and h_1 , used to construct the estimators \hat{g} and \hat{g}_{kj} of g and g_{kj} , on the classifier. A proof is given in Appendix A.1.

THEOREM 1. *Assume that (3.5)–(3.7) hold, and let $\Psi_0 = 1 - \Phi$ and $\Psi_1 = \Phi$, where Φ denotes the c.d.f. of a standard normal random variable. Then*

$$(3.8) \quad \begin{aligned} \text{err}_k &= \text{err}_{k0} + h^2 c_k + h_1^2 c_{k1} + \frac{d_{k0}}{\nu_0 h_1} + \frac{d_{k1}}{\nu_1 h_1} \\ &\quad + O\{m^{-1} + (mh)^{-2}\} + o\left(h^2 + h_1^2 + \frac{1}{\nu_0 h_1}\right), \end{aligned}$$

where $\text{err}_{k0} = E_k[\Psi_k\{-\beta_{k0}/\sigma_k\}]$, $c_k = \kappa_2 \alpha_k \int_{\mathcal{I}} (\mu_1 - \mu_0) \mu_k''$, $c_{k1} = -\kappa_2 \alpha_k \int_{\mathcal{I}} (\mu_1 - \mu_0) \mu_{1-k}''$, $d_{kj} = (-1)^j \alpha_k \sigma_{\varepsilon_j}^2 \kappa \int_{\mathcal{I}} f_X^{-1}$, with $\alpha_k = (-1)^k \tau_k^{-1} \phi(b_{k0}/\tau_k)$, and where ϕ denotes the standard normal density function.

The leading term err_{k0} on the right-hand side of (3.8) does not depend in any way on the bandwidths h and h_1 . It does involve the training sample sizes n_0 and n_1 , and in particular does not equal the asymptotic limit of err_k as n increases, since that limit is given by $\Psi_k(-b_{k0}/\tau_k)$, but the effects of the bandwidths are all confined to subsequent terms on the right-hand side of (3.8). The terms in h^2 and h_1^2 represent contributions to classification error arising from biases of the estimators \hat{g} and \hat{g}_{kj} , and the terms in $(\nu_0 h_1)^{-1}$ and $(\nu_1 h_1)^{-1}$ are contributions from the variances of the estimators \hat{g}_{kj} .

While a priori it might be thought that, since the total number of observations in the training sample, $\sum_j m_{kj}$, for $k = 0$ and 1 , is an order of magnitude larger than the number of observations, m , in the new data set \mathcal{D} , then h_1 should be chosen smaller than h , Theorem 1 shows that the influence of bandwidths on error rate is much more complex than this.

For one thing, there are no terms in $(mh)^{-1}$ on the right-hand side of (3.8). (Section 3.1.2 will explain the reason for this.) As a result, the terms on the right-hand side of (3.8) that depend on h can be rendered equal to $O(m^{-1})$ simply by taking h equal to a constant multiple of $m^{-1/2}$. As noted in Remark 1, below, this level of contribution to the error rate is generally impossible to remove, even in simple parametric problems. Therefore the

contribution of h to error rate cannot be rendered smaller than m^{-1} . However, in some instances choosing h to be an order of magnitude larger or smaller than $m^{-1/2}$ can be beneficial; see Section 3.1.1 below.

The terms in h_1 on the right-hand side of (3.8) are a different matter because each of c_{k1} and $d_{k0}\nu_0^{-1} + d_{k1}\nu_1^{-1}$ can be either positive or negative. Depending on the signs and sizes of c_{k1} and $d_{k0}\nu_0^{-1} + d_{k1}\nu_1^{-1}$, it can be optimal to take h_1 to be of order $\nu_k^{-1/3}$, which achieves a trade-off between terms in h_1^2 and $(\nu_k h_1)^{-1}$, or to take h_1 to decrease to zero more quickly or to converge to a positive constant, as n increases; see Section 3.1.1 below.

Therefore, the impact that smoothing has on classification performance is much more subtle than it might have appeared. We discuss these issues in more detail in the next sections.

3.1.1. *Sizes of h and h_1 that optimise overall error rate.* Using Theorem 1 we can deduce the orders of magnitudes of h and h_1 that minimise the error rate of the classifier, that is, that minimise the probability of misclassification,

$$(3.9) \quad \text{err} = \pi_0 \text{err}_0 + \pi_1 \text{err}_1,$$

where err_0 and err_1 are as in (3.8), π_k denotes the prior probability attached to population Π_k , and $\pi_0 + \pi_1 = 1$. Using (3.8) and (3.9), we can write

$$(3.10) \quad \begin{aligned} \text{err} = & \text{err}^0 + c^0 h^2 + c_1^0 h_1^2 + d^0 (\nu_0 h_1)^{-1} + O\{m^{-1} + (mh)^{-2}\} \\ & + o\{h^2 + h_1^2 + (\nu_0 h_1)^{-1}\}, \end{aligned}$$

where $\text{err}^0 = \pi_0 \text{err}_{00} + \pi_1 \text{err}_{10}$ (recall that err_{k0} does not depend on the bandwidths),

$$\begin{aligned} c^0 &= \kappa_2 \int (\mu_1 - \mu_0) \left\{ \pi_0 \frac{\mu_0''}{\tau_0} \phi\left(\frac{b_{00}}{\tau_0}\right) - \pi_1 \frac{\mu_1''}{\tau_1} \phi\left(\frac{b_{10}}{\tau_1}\right) \right\}, \\ c_1^0 &= \kappa_2 \int (\mu_0 - \mu_1) \left\{ \pi_0 \frac{\mu_1''}{\tau_0} \phi\left(\frac{b_{00}}{\tau_0}\right) - \pi_1 \frac{\mu_0''}{\tau_1} \phi\left(\frac{b_{10}}{\tau_1}\right) \right\}, \\ d^0 &= \kappa \left(\int_{\mathcal{I}} f_X^{-1} \right) \left\{ \frac{\pi_0}{\tau_0} \phi\left(\frac{b_{00}}{\tau_0}\right) - \frac{\pi_1}{\tau_1} \phi\left(\frac{b_{10}}{\tau_1}\right) \right\} \left(\sigma_{\varepsilon_0}^2 - \sigma_{\varepsilon_1}^2 \frac{\nu_0}{\nu_1} \right). \end{aligned}$$

Since the function ϕ is symmetric, and $b_{10} = -b_{00}$, then b_{10} can be replaced by b_{00} in the formula for d_0 without altering its veracity.

To appreciate the very wide range of optimal bandwidth choices that can arise in the problem of minimising error rate, let us consider minimising err , at (3.10). To help remove ambiguities, let us assume that as n increases the value of $\sigma_{\varepsilon_0}^2 - \sigma_{\varepsilon_1}^2 \nu_0 \nu_1^{-1}$ is of the same sign for all sufficiently large n , and its absolute value is bounded away from zero; assumption (3.7)(c) ensures that it is uniformly bounded. In this instance, and focusing just on the terms in h_1 , we see that four distinct cases can arise in practice:

(i) c_1^0 and d^0 are both positive. In this case, to minimise the contribution from h_1 , we should minimise $c_1^0 h_1^2 + d^0 (\nu_0 h_1)^{-1}$, which is achieved by taking h_1 to be of size $\nu_0^{-1/3}$.

(ii) c_1^0 and d^0 are both negative. In this case, the contribution made by h_1 behaves like $-\{|c_1^0| h_1^2 + |d^0| (\nu_0 h_1)^{-1}\}$ as sample size increases. The term within braces here is maximised by taking $h_1 = 0$, and analogously, in minimising err, it is optimal to take h_1 to be of strictly smaller order than $\nu_0^{-1/3}$.

(iii) $c_1^0 > 0$ and $d^0 \leq 0$. In this case, to minimise the error rate, we need to maximise the size of the negative term and minimise that of the positive term, which is achieved by taking h_1 to be of strictly smaller order than $\nu_0^{-1/3}$ (the precise order depends on the magnitude of second order terms, but deriving the latter precisely would require a lot of additional computation).

(iv) $c_1^0 < 0$ and $d^0 \geq 0$. Here, using arguments similar to those in case (iii), taking h_1 to be of strictly larger order than $\nu_0^{-1/3}$ is optimal.

The case $d^0 = 0$ occurs, for example, if the covariance G_k of the Gaussian process g , the experimental error variance $\sigma_{\varepsilon k}^2$, and the values of m_{kj} and n_k do not depend on k . Equal values of m_{kj} commonly arise when the data are observed on a grid; see Remark 4.

A similar analysis can be carried out in the case of optimisation over h rather than h_1 , although there the optimum is accessed from a comparison of terms in h and $(mh)^{-2}$, rather than h_1^2 and $(\nu_0 h_1)^{-1}$. [A tedious analysis of the term of size $(mh)^{-2}$, represented by the remainder $O\{(mh)^{-2}\}$ in (3.8), shows that it can be either positive or negative.] Depending on the relative signs of the terms in h^2 and $(mh)^{-2}$, it can be optimal to take $h \asymp m^{-1/2}$, or h of strictly larger, or strictly smaller order than $m^{-1/2}$.

Similar results are obtained if we investigate properties of err_k , in (3.8), instead of the overall error rate, err, at (3.9).

These results explain the very diverse patterns of behaviour that are seen in numerical work, and that motivated our research; see Section 1. In summary, in apparently similar problems and using the same type of classifier, it can be optimal to use a very small bandwidth, or a very large bandwidth, or a bandwidth of only moderate size, depending on the signs of certain constants. Therein lies the contradictory nature of the smoothing parameter choice problem for classification of functional data.

3.1.2. *Absence of terms in $(mh)^{-1}$.* The centroid-based classifier statistic $S(\hat{g})$, at (2.13), can be written equivalently as

$$(3.11) \quad S(\hat{g}) = \int_{\mathcal{I}} (\hat{\mu}_1 - \hat{\mu}_0)(2\hat{g} - \hat{\mu}_0 - \hat{\mu}_1) dt.$$

Importantly, there is no quadratic term in \hat{g}^2 in (3.11), and as a result the impact of the bandwidth h , although not h_1 , on properties of the classifier is greatly reduced. This reduction is brought about by the smoothing effect of the integral in (3.11), which results in the elimination of terms in $(mh)^{-1}$.

This property, to which we shall refer to as the “integration effect,” is known in other settings, for example, when integrating a kernel density estimator, computed from a sample of size m , to produce a distribution estimator. Integration results in the variance reducing from order $(mh)^{-1}$, for the density estimator, to order m^{-1} , for the distribution function estimator—just as it does in the setting above.

REMARK 1 (Order m^{-1} term in expansion of classification error). We assumed in (3.7)(c) that the values of m_{kj} , representing the number of pairs (X_{kji}, Y_{kji}) for a given population index k and given individual j , are all of roughly the same size. In this setting it is easy to see that, even in an elementary parametric setting, we must expect the operation of observing the functions g_{kj} at scattered points to affect error rate through a term of order m^{-1} , and no smaller. For example, consider the case where $g_{kj} = \psi(\cdot | \omega_{kj})$, with $\psi(\cdot | \omega)$ being a known function completely determined by the parameter ω , and $\omega_{kj} = \int_{\mathcal{I}} g_{kj} w$ where the weight function w is known. Using the data \mathcal{D}_{kj} on g_{kj} we can estimate ω_{kj} root- m consistently, but no faster, and as a result we incur a classification error of size m^{-1} , and no smaller, from not knowing the values ω_{kj} . It is for this reason that, when developing expansions of classification error, we do not explore the remainder of size m^{-1} ; it is stated simply as $O(m^{-1})$ on the right-hand side of (3.8).

3.1.3. *Other remarks.* We conclude our discussion of Theorem 1 with a number of remarks.

REMARK 2 (Definition of $\hat{\mu}_k$). The size of the fourth and fifth terms on the right-hand side of (3.8) is determined by the sizes of ν_0^{-1} and ν_1^{-1} , and those quantities can be made slightly smaller by using a slightly different definition of $\hat{\mu}_k$, at (2.8). In particular in (2.8), on account of the definition of \hat{g}_{kj} at (2.3), $\hat{\mu}_k$ is defined as an average of ratios of sums, whereas slightly better statistical performance is obtained by taking $\hat{\mu}_k$ to be simply a ratio of sums:

$$\hat{\mu}_k = \frac{\sum_j (U_{kj2} V_{kj0} - U_{kj1} V_{kj1})}{\sum_j (U_{kj2} U_{kj0} - U_{kj1}^2)},$$

compare (2.3). However, this approach departs from standard practice in working with functional data, and therefore, since convergence rates do not alter (only the constant multiples of rates are reduced), we have followed standard practice in the definition of $\hat{\mu}_k$.

REMARK 3 (Gaussian assumption). Of course, if m is sufficiently large then \hat{g} is itself approximately Gaussian, and so the assumption that g is a Gaussian process is reflected particularly well in properties of its estimator. More generally, our assumption that g is a Gaussian process is made for simplicity, and can be relaxed. For example, generalisations to chi-squared and other processes, where shape can be described in terms of a small number of fixed functions (mean and covariance in the Gaussian case), are straightforward.

More generally we would require a model which described the properties of random functions relatively simply. The Gaussian model fills this need ideally; shape is described by mean and variance functions, on which we have imposed only smoothness, rather than parametric, conditions. Moreover, in the Gaussian case all moments of $g(x)$ are finite, for each x (we use this property repeatedly during our theoretical arguments), and the principal component scores are independent (this is used frequently during our proof of Theorem 2).

REMARK 4 (Case of regularly spaced design). Theorem 1 continues to hold if the m_{kj} design variables X_{kji} are regularly spaced on \mathcal{I} for each k and j . The only change necessary is to replace $\int_{\mathcal{I}} f_X^{-1}$, on the right-hand side of (3.8), by the square of the length of the interval \mathcal{I} .

3.2. *Scale-adjusted centroid-based classifier.* Recall that scale-adjusted centroid-based classifier is defined in terms of $S_{\text{scale}}(\hat{g})$, at (2.14). A decomposition similar to that of Theorem 1 can be derived for this classifier, as we shall prove in Theorem 2 below. For this classifier, it seems necessary to strengthen (3.7) by imposing conditions on the behaviour of the eigenvalues $\theta_{k\ell}$ as ℓ increases. However, since our aim in this section is only to corroborate the conclusions in Section 3.1, drawn there in the case of the standard centroid-based classifier, then we shall simplify our account by assuming that g is finite dimensional, and in particular taking the covariance expansion at (2.10) to have just q terms:

(3.12) For $k = 0$ and 1: (a) the first q eigenvalues in the sequence $\theta_{k1} \geq \theta_{k2} \cdots$, arising in the covariance expansion (2.10) of g when the data come from Π_k , are distinct; (b) $\theta_{k\ell} = 0$ for $\ell > q$; (c) for $1 \leq \ell \leq q$ the eigenfunctions $\psi_{k\ell}$ in (2.10) have two Hölder continuous derivatives on \mathcal{I} ; (d) $E_k(g)$ is a linear form in $\psi_{k1}, \dots, \psi_{kq}$; and (e) in the definition of $S_{\text{scale}}(\hat{g})$, $s_0^2 \neq s_1^2$.

Without (3.12)(a), separate conditions, valid uniformly in $j = 1, 2, \dots$, have to be imposed on remainders in Taylor expansions of “smoothed” versions of the eigenvalues θ_{kj} , depending on h .

The next theorem indicates that the results of Theorem 1 also apply for the scale-adjusted centroid-based classifier. Its proof is given in the supplementary material [Carroll, Delaigle and Hall (2013)].

THEOREM 2. *Assume that (3.5), (3.6) and (3.12) hold. Then the error rate of the scale-adjusted centroid-based classifier, when the data in \mathcal{D} are drawn from Π_k , admits the expansion at (3.8), but with different constants, where the various terms have the properties stated immediately below that formula.*

The diversity of possible signs of c_k , c_{k1} and $d_{k0}\nu_0^{-1} + d_{k1}\nu_1^{-1}$ in (3.8), discussed in Section 3.1.1, is also present in this case. Therefore the conclusions drawn in that section apply to the scale-adjusted centroid-based classifier. However, we have not derived explicitly the counterparts of the constants c_k , c_{k1} , d_{k0} and d_{k1} that appear in equation (3.8).

The integration effect discussed in Section 3.1.2 is also present here, although we had originally expected that the scale-adjusted centroid classifier would produce a term of size $(mh)^{-1}$ in an expansion of error rate. Indeed, the situation initially seems quite different in the case of the scale-adjusted version $S_{\text{scale}}(\hat{g})$ of $S(\hat{g})$, at (2.14), when $s_0^2 \neq s_1^2$. There the quadratic term in \hat{g} persists. The reason it still does not produce a term in $(mh)^{-1}$ is quite subtle. Define \bowtie_k to be $>$ or \leq according as $k = 0$ or $k = 1$, respectively. The probability $P_k\{S_{\text{scale}}(\hat{g}) \bowtie_k 0\}$ can be written as

$$P_k \left\{ \sum_{j=1}^{\infty} w_j (Z_j + V_j)^2 \bowtie_k W \right\} + \text{negligible terms,}$$

where the Z_j 's are independent $N(0,1)$ variables, conditional on the V_j 's and W ; the positive weights w_j are nonrandom; and critically, W does not involve the experimental errors ε_i in (2.2), from which any term in $(mh)^{-1}$ would arise. The terms V_j depend on the experimental errors only through integrals of the error process, and the integration effect at this point largely removes the impact of the error bandwidth h , with the result that there is no term of size $(mh)^{-1}$. However, terms in $(\nu_0 h_1)^{-1}$ remain; the integration effect only influences smoothing of the new data, not of the training data.

3.3. Quadratic discriminant. Finally, we show that similar smoothing effects are present in the case of the quadratic discriminant classifier defined through the statistic $T(\hat{g})$ at (2.15). Recall that, when the data in \mathcal{D} come from Π_k , the random function g has covariance function G_k . To derive the counterpart of Theorem 1 for this classifier, let r, r_1, r_2 take the values 0 and 1, let $1 \leq \ell, \ell_1, \ell_2 \leq p$, and define the covariances

$$\text{cov}_k[r_1, r_2; \ell_1, \ell_2] = \int_{\mathcal{I}} \int_{\mathcal{I}} G_k(x_1, x_2) \psi_{r_1 \ell_1}(x_1) \psi_{r_2 \ell_2}(x_2) dx_1 dx_2,$$

the variances $\text{var}_k[r, \ell] = \text{cov}_k[r, r; \ell, \ell]$, and the correlations

$$\rho_k[r_1, r_2; \ell_1, \ell_2] = \frac{\text{cov}[r_1, r_2; \ell_1, \ell_2]}{(\text{var}_k[r_1, \ell_1] \text{var}_k[r_2, \ell_2])^{1/2}}.$$

Let $p \geq 1$, a fixed number, be the number of principal components used to construct the quadratic discriminant statistic $T(\hat{g})$, defined at (2.15). Theorem 3 below addresses the error rate of the quadratic discriminant based on $T(\hat{g})$, and there we shall assume that:

- (3.13) (a) For $k = 0, 1$ the eigenvalues $\theta_{k1}, \dots, \theta_{k,p+1}$ are distinct; and
 (b) among the values taken by $\rho_k[r_1, r_2; \ell_1, \ell_2]$ for $k, r_1, r_2 = 0, 1$ and $1 \leq \ell_1, \ell_2 \leq p$, the absolute value of $\rho_k[r_1, r_2; \ell_1, \ell_2]$ equals 1 only when $r_1 = r_2$ and $\ell_1 = \ell_2$.

Condition (3.13)(a) ensures that the eigenfunctions $\psi_{k\ell}$ are well defined for $k = 0, 1$ and $\ell = 1, \dots, p$; and (3.13)(b) guarantees that the quantities $\int_{\mathcal{X}} (g - \mu_{r_1}) \psi_{r_1 \ell_1}$ and $\int_{\mathcal{X}} (g - \mu_{r_2}) \psi_{r_2 \ell_2}$, which appear in the definition of $T_0(g)$ at (2.16), cannot be identical, except for a difference in means, unless $r_1 = r_2$ and $\ell_1 = \ell_2$, thereby avoiding degeneracy.

The counterpart of Theorem 1 for the quadratic discriminant classifier is stated in the next theorem. Its proof is given in the supplementary material [Carroll, Delaigle and Hall (2013)].

THEOREM 3. *Assume that (3.5)–(3.7) and (3.13) hold. Then the error rate of the quadratic discriminant, when the data in \mathcal{D} come from Π_k , admits the expansion at (3.8), but with different constants, where the various terms have the properties stated immediately below that formula.*

Again the signs of c_k , c_{k1} and $d_{k0}\nu_0^{-1} + d_{k1}\nu_1^{-1}$, in (3.8), are particularly diverse, and so the conclusions reached in Section 3.1.1 apply. Likewise, the integration effect discussed in Section 3.1.2 is also observed. Here, as can be seen directly from (2.15), the estimator \hat{g} is integrated, and only the integral is squared, not \hat{g} itself. The resulting integration effect eliminates any term in $(mh)^{-1}$ from the analogue of the expansion (3.8) in this setting, although again this influence does not carry over to the training data.

4. Numerical illustrations.

4.1. *Simulated data.* To illustrate the impact of bandwidth on classification performance, we generated data from several instances of model (2.1), taking, in each case, $m_{kj} = 50$. Let $\phi_\sigma(x)$ denote the normal density function with mean zero and standard deviation σ . We considered the following cases, each with three different levels of errors, which we refer to as noise versions 1, 2 and 3:

(A): $g_{kj}(t) = \mu_k(t) + (3t + 100)^{1/2} \{\cos(t/50)\}^k Z_{kj}$, where $\mu_0(t) = \phi_{10}(t - 5)$, $\mu_1(t) = \mu_0(t) + 0.3 \cos(t/5) + 0.1$, $Z_{kj} \sim U[-1/(30 - 10k), 1/(30 - 10k)]$, and $\varepsilon_{kji} \sim N(0, 1/(4 - 2k)^2)$ (noise version 1), $\varepsilon_{kji} \sim N(0, 2/(4 - 2k)^2)$ (noise version 2) or $\varepsilon_{kji} \sim N(0, 4/(4 - 2k)^2)$ (noise version 3), and $\pi_0 = 1/3$, $\pi_1 = 2/3$. Moreover, $X_{kji} = 2i - 1$, for $i = 1, \dots, 50$.

(B): $g_{kj}(t) = \mu_k(t) + (3t + 100)^{1/2} Z_{kj}$, where $\mu_0(t) = 30\{0.2\phi_4(t - 5) + 0.1\phi_4(t - 10) + 0.4\phi_6(t - 20) + 0.4\phi_6(t - 35) + 0.6\phi_7(t - 55) + 0.6\phi_7(t - 80)\}$, $\mu_1(t) = \mu_0(t) + 4/\{(t - 50)^2 + 10\}$, $Z_{kj} \sim U[-1/(60 + 15k), 1/(60 + 15k)]$, $\varepsilon_{kji} \sim \{\text{Exp}(0.5) - 2\}/(2 + 2k)$ (noise version 1), $\varepsilon_{kji} \sim \sqrt{2}\{\text{Exp}(0.5) - 2\}/(2 + 2k)$ (noise version 2) or $\varepsilon_{kji} \sim \{\text{Exp}(0.5) - 2\}/(1 + k)$ (noise version 3), and $\pi_0 = 2/5$ and $\pi_1 = 3/5$. Moreover, X_{kji} was as in (A).

(C): $g_{0j}(t) = \mu_0(t) + (3t + 100)^{1/2} Z_{0j}$, $g_{1j}(t) = \mu_1(t) + (t + 5)Z_{1j}$, where $\mu_0(t) = 15\phi_{17}(t - 65) \cos(t/7)$, $\mu_1(t) = \mu_0(t) + 5\phi_{20}(t - 50)$, $Z_{kj} \sim U[-1/(50 - 10k), 1/(50 - 10k)]$, $\varepsilon_{kji} \sim N(0, (4 - k)^2/100)$ (noise version 1), $\varepsilon_{kji} \sim N(0, (4 - k)^2/50)$ (noise version 2), $\varepsilon_{kji} \sim N(0, (4 - k)^2/25)$ (noise version 3), and $\pi_0 = 2/3$ and $\pi_1 = 1/3$. Moreover, X_{kji} was as in (A).

(D)–(F): Same as (A) to (C) but with $X_{kji} = 2i - 1 + T_{kji}$, where $T_{kji} \sim N(0, 0.25)$.

We chose these examples to illustrate various features of the problems, namely that the impact of smoothing may differ among classifiers, and that in some cases, some classifiers perform better with more smoothing and in other cases, they might perform better with less smoothing.

In each case, for $k = 0, 1$ and for several values of n_{tr} , we generated 100 (resp., n_{tr}) noisy test curves (resp., training curves) from model (2.1), each of which came with probability π_k from Π_k . We constructed each classifier from the training data, and applied it to the test data. To compute \hat{g} and \hat{g}_{kj} , we compared three approaches for selecting the bandwidths: no smoothing (NS), the standard plug-in (PI) bandwidths h_{PI} and $h_{\text{PI},kj}$ that estimate the optimal bandwidth for estimation of the regression functions g and g_{kj} , which we computed using the `dpill` function in the R package `KernSmooth`; see Ruppert, Sheather and Wand (1995); and the bandwidths γh_{PI} and $\gamma_1 h_{\text{PI},kj}$, where γ and γ_1 (and also the truncation parameter p in the case of the quadratic discriminant classifier) were chosen to minimise the following cross-validation (CV) estimator of classification error:

$$\widehat{\text{err}} = \frac{\hat{\pi}_0}{n_0} \sum_{i=1}^{n_0} I\{\hat{\mathcal{C}}_{i0,-i} = 1\} + \frac{\hat{\pi}_1}{n_1} \sum_{i=1}^{n_1} I\{\hat{\mathcal{C}}_{i1,-i} = 0\}$$

with $\hat{\pi}_0$ and $\hat{\pi}_1$ denoting estimators of π_0 and π_1 (we took $\hat{\pi}_k = 1/2$), and $\hat{\mathcal{C}}_{ik,-i}$ being the estimator of the class label of the i th training observation from group k , obtained from the classifier constructed without using this observation.

TABLE 1

Percentage of correctly classified observations for the simulated data of Section 4.1, using plug-in (PI) regression bandwidths, bandwidths that minimise a crossvalidation (CV) estimate of classification error, or without smoothing the noisy data (NS). The three noise versions, in increasing order, are described in cases (A)–(C) in Section 4.1. Here “Cent” is the centroid classifier (2.13), “Cent sc.” is the scaled centroid classifier (2.14) and “QDA” is the quadratic discriminant classifier (2.15)

	n_{tr}	Cent			Cent sc.			QDA		
		CV	PI	NS	CV	PI	NS	CV	PI	NS
Case (A)										
Noise version 1	50	82.9	74.1	84.0	91.8	73.2	92.0	95.1	94.1	53.5
Noise version 1	100	84.4	74.9	84.8	92.6	74.1	92.6	97.6	94.8	67.6
Noise version 2	50	77.7	69.6	78.1	94.3	70.4	94.4	91.0	89.3	49.1
Noise version 2	100	79.9	70.7	80.2	95.1	71.2	95.1	94.3	89.7	61.7
Noise version 3	50	71.1	65.6	71.1	97.1	69.0	97.1	85.4	84.3	46.2
Noise version 3	100	73.7	66.8	74.1	97.9	69.4	97.9	89.9	84.1	58.4
Case (B)										
Noise version 1	50	63.2	60.1	65.7	96.3	78.7	96.5	77.1	74.3	65.8
Noise version 1	100	65.5	61.5	66.8	96.8	80.0	96.8	81.8	76.3	73.0
Noise version 2	50	61.5	58.6	64.6	96.3	80.6	96.4	76.9	74.1	65.2
Noise version 2	100	62.6	58.7	64.4	96.7	81.3	96.7	81.3	75.0	72.4
Noise version 3	50	60.9	57.6	64.0	96.2	81.6	96.4	77.3	74.2	65.4
Noise version 3	100	60.7	56.8	63.3	96.7	82.3	96.7	81.6	75.2	72.3
Case (C)										
Noise version 1	50	61.5	60.8	60.8	88.7	89.2	87.4	84.8	83.7	82.0
Noise version 1	100	59.4	58.4	58.2	90.0	90.3	88.5	86.9	85.7	79.0
Noise version 2	50	61.3	60.2	60.3	87.3	87.9	82.8	81.9	81.2	82.4
Noise version 2	100	58.9	57.9	57.6	88.8	89.0	85.2	84.6	83.1	80.4
Noise version 3	50	61.0	59.7	59.3	85.2	85.4	71.2	80.5	79.9	79.6
Noise version 3	100	58.5	57.4	57.0	87.2	86.6	74.9	82.6	81.1	79.7

For each configuration, we generated $B = 200$ sets of training and test samples. In Tables 1 and 2, we report the percentage of correctly classified test curves, averaged over the B replicates. Depending on the model, the classifier, and the type of data (test or training), the cross-validation bandwidths were either smaller or larger than the PI regression bandwidths, illustrating the variety of settings already explained by our theory. See Table B.1 in Section B.3 in the supplementary material [Carroll, Delaigle and Hall (2013)], where we report the value of γ and γ_1 averaged over the B replicates. We can see from the table that in most cases, γ was smaller than γ_1 , and both were usually smaller than 1, except in cases (C) and (F).

As expected, we conclude from Tables 1 and 2, depending on the model and the classifier, the negative impact of smoothing with the standard PI

TABLE 2

Percentage of correctly classified observations for the simulated data of Section 4.1, using plug-in (PI) regression bandwidths, bandwidths that minimise a crossvalidation (CV) estimate of classification error, or without smoothing the noisy data (NS). The three noise versions, in increasing order, are described in cases (D)–(F) in Section 4.1. Here “Cent” is the centroid classifier (2.13), “Cent sc.” is the scaled centroid classifier (2.14) and “QDA” is the quadratic discriminant classifier (2.15)

	n_{tr}	Cent			Cent sc.			QDA		
		CV	PI	NS	CV	PI	NS	CV	PI	NS
Case (D)										
Noise version 1	50	80.2	69.5	80.6	85.7	68.5	86.3	93.9	92.7	69.2
Noise version 1	100	81.5	70.0	82.0	87.3	69.2	87.3	96.6	93.2	84.3
Noise version 2	50	74.7	65.9	75.6	90.0	65.6	90.3	88.5	86.7	60.9
Noise version 2	100	76.9	66.8	77.3	90.9	66.8	91.0	92.3	86.8	77.6
Noise version 3	50	69.2	62.3	69.6	94.2	65.4	94.4	82.9	80.5	55.6
Noise version 3	100	71.4	63.4	72.0	95.1	66.4	95.1	87.9	80.8	72.7
Case (E)										
Noise version 1	50	65.0	61.9	67.3	94.8	79.0	95.0	77.0	73.1	71.2
Noise version 1	100	65.8	62.5	67.6	95.4	79.6	95.4	84.3	69.7	82.8
Noise version 2	50	63.0	60.2	65.4	94.7	80.6	95.0	77.8	74.2	70.5
Noise version 2	100	63.4	59.9	64.9	95.4	81.4	95.5	84.8	69.8	82.4
Noise version 3	50	61.3	59.2	64.3	94.6	81.4	94.9	77.9	74.4	69.8
Noise version 3	100	61.8	58.2	63.1	95.4	82.5	95.5	84.5	71.5	81.8
Case (F)										
Noise version 1	50	60.2	59.1	59.4	88.0	88.7	87.9	83.5	82.6	80.4
Noise version 1	100	58.8	57.8	57.7	89.0	89.3	88.5	84.9	83.2	77.3
Noise version 2	50	59.8	58.7	59.0	86.5	87.2	84.6	80.8	80.2	80.0
Noise version 2	100	58.6	57.3	57.2	87.6	87.7	85.8	83.1	81.1	77.0
Noise version 3	50	59.2	58.3	58.2	84.5	84.1	76.7	79.4	78.5	78.9
Noise version 3	100	58.0	56.8	56.5	85.8	84.6	78.5	81.0	79.1	75.7

bandwidth can be quite significant, indeed sometimes reducing the percentage of correctly classified data by as much as 10%. In cases (A) and (D), it is the centroid classifier and its scaled version that are the most affected by this inappropriate level of smoothing, whereas the quadratic discriminant classifier is more robust against the level of smoothing. In cases (B) and (E), the scaled centroid classifier and the quadratic discriminant classifier are the most affected by inappropriate smoothing. Cases (C) and (F) are more robust against smoothing; there, all three versions (PI, CV and NS) of the data result in similar classification performance, although overall the data smoothed by CV result in slightly improved performance. Depending on the case, when the noise level increases the impact of inappropriate bandwidth choice can either increase or decrease.

TABLE 3

Percentage of correctly classified observations for the ovarian cancer data, using plug-in (PI) regression bandwidths or bandwidths that minimise a crossvalidation (CV) estimate of classification error. Here “Cent” is the centroid classifier (2.13), “Cent sc.” is the scaled centroid classifier (2.14) and “QDA” is the quadratic discriminant classifier (2.15)

Data	n_{tr}	Cent		Cent sc.		QDA	
		CV	PI	CV	PI	CV	PI
Original data	50	90.60	80.25	90.05	78.79	93.32	89.69
Original data	100	90.43	80.96	90.00	79.96	98.58	98.86
Noisy version 1	50	88.07	75.19	87.83	74.23	78.03	68.50
Noisy version 1	100	87.58	76.76	88.54	76.27	91.48	90.97
Noisy version 2	50	76.15	66.57	76.65	66.09	56.91	48.54
Noisy version 2	100	81.97	67.55	81.91	67.64	77.62	66.49

4.2. *Real data.* We illustrate our findings on the ovarian cancer data set 8-7-02, which concerns 253 patients (91 controls and 162 with ovarian cancer). The data, which were produced to study the effect of robotic sample handling, are available from <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>. In this example, the functions X_i represent proteomic mass spectra and $t \in [0, 20,000]$ is the mass over charge ratio, m/z . These raw curves are ideal for illustrating the negative impact that systematically smoothing by standard methods can have, because in some ranges of values of t , the spectra have considerable activity, and the impact of smoothing such data can be striking. We focus on one such ranges, namely $t \in [200, 500]$.

To assess the performance of classifiers on this data set, we randomly and uniformly created $B = 200$ pairs of (training sample, test sample), where we took the training sample to be of size n_{tr} and the test sample of size $253 - n_{\text{tr}}$, for $n_{\text{tr}} = 50$ and $n_{\text{tr}} = 100$. We also generated two more noise versions of the data, adding to the Y_{kji} ’s in both the test and training data, noise $\varepsilon'_{kji} \sim N(0, 0.04)$ (noise version 1) or $\varepsilon'_{kji} \sim N(0, 0.25)$ (noise version 2), where the ε'_{kji} ’s were totally independent.

For each version of the data (original data and noise versions 1 and 2), and for each pair of test and training sample, we constructed each classifier from the training sample, and applied the classifier to the test sample using either plug-in regression bandwidths to construct the estimators \hat{g} and \hat{g}_{kj} , or bandwidths obtained by minimising the CV estimator of classification error defined in Section 4.1, where we took $\hat{\pi}_k = 1/2$.

Table 3 reports the percentage, averaged over the B pairs of samples, of correctly classified observations from the test samples. The table indicates very clearly that smoothing the data using the plug-in regression bandwidths

degraded the quality of the two versions of the centroid classifier by about 10%, and a similar phenomenon was observed for the quadratic discriminant classifier when the training sample was small and when the data were noisy.

APPENDIX: PROOF OF THEOREM 1

A.1. Preliminary results. Define

$$\Delta_\ell(x) = \frac{1}{mh} \sum_{i=1}^m \varepsilon_i \left(\frac{x - X_i}{h} \right)^\ell K \left(\frac{x - X_i}{h} \right),$$

$$W_\ell(x) = \frac{1}{mh} \sum_{i=1}^m \left[\int_x^{X_i} \{g''(t) - g''(x)\} (X_i - t) dt \right] \left(\frac{x - X_i}{h} \right)^\ell K \left(\frac{x - X_i}{h} \right).$$

With U_ℓ and V_ℓ given by (2.4) and (2.5), and using the model at (2.2) and the exact form of the remainder in Taylor's theorem, we can write:

$$\begin{aligned} V_\ell(x) &= \frac{1}{mh} \sum_{i=1}^m \{g(X_i) + \varepsilon_i\} \left(\frac{x - X_i}{h} \right)^\ell K \left(\frac{x - X_i}{h} \right) \\ &= \frac{1}{mh} \sum_{i=1}^m \left[g(x) + (X_i - x)g'(x) + \frac{1}{2}(X_i - x)^2 g''(x) + \varepsilon_i \right] \\ &\quad \times \left(\frac{x - X_i}{h} \right)^\ell K \left(\frac{x - X_i}{h} \right) + W_\ell(x) \\ &= g(x)U_\ell(x) - hg'(x)U_{\ell+1}(x) + \frac{1}{2}h^2 g''(x)U_{\ell+2}(x) + \Delta_\ell(x) + W_\ell(x). \end{aligned}$$

Assuming, without loss of generality, that K is supported on $[-1, 1]$,

$$|W_\ell(x)| \leq h^2 \left\{ \sup_{t \in \mathcal{I}: |t-x| \leq h} |g''(t) - g''(x)| \right\} \frac{1}{mh} \sum_{i=1}^m K \left(\frac{x - X_i}{h} \right) \leq h^2 U_0(x) Q,$$

where $Q = \sup_{s, t \in \mathcal{I}: |s-t| \leq h} |g''(s) - g''(t)|$. Now,

$$\hat{g} = \frac{U_2 V_0 - U_1 V_1}{U_2 U_0 - U_1^2} = g + \frac{1}{2} h^2 g'' \frac{U_2^2 - U_1 U_3}{U_2 U_0 - U_1^2} + \Delta + \frac{U_2 W_0 - U_1 W_1}{U_2 U_0 - U_1^2},$$

where $\Delta = (U_2 \Delta_0 - U_1 \Delta_1) / (U_2 U_0 - U_1^2)$. Therefore, since $|U_\ell| \leq U_0$ for each $\ell \geq 0$,

$$(A.1) \quad \left| \hat{g} - \left(g + \frac{1}{2} h^2 g'' \frac{U_2^2 - U_1 U_3}{U_2 U_0 - U_1^2} + \Delta \right) \right| \leq \frac{2Qh^2 U_0^2}{U_2 U_0 - U_1^2},$$

uniformly on \mathcal{I} .

Similarly, defining $Q_{kj} = \sup_{s,t \in \mathcal{I}: |s-t| \leq h} |g''_{kj}(s) - g''_{kj}(t)|$, and

$$\begin{aligned} \Delta_{kj\ell}(x) &= \frac{1}{m_{kj}h_1} \sum_{i=1}^{m_{kj}} \varepsilon_{kji} \left(\frac{x - X_{kji}}{h_1} \right)^\ell K \left(\frac{x - X_{kji}}{h_1} \right), \\ \Delta_{kj} &= \frac{U_{kj2}\Delta_{kj0} - U_{kj1}\Delta_{kj1}}{U_{kj2}U_{kj0} - U_{kj1}^2}, \end{aligned}$$

where $U_{kj\ell}$ is as at (2.6), we have, uniformly on \mathcal{I}

$$(A.2) \quad \left| \hat{g}_{kj} - \left(g_{kj} + \frac{1}{2} h_1^2 g''_{kj} \frac{U_{kj2}^2 - U_{kj1}U_{kj3}}{U_{kj2}U_{kj0} - U_{kj1}^2} + \Delta_{kj} \right) \right| \leq \frac{2Q_{kj}h_1^2U_{kj0}^2}{U_{kj2}U_{kj0} - U_{kj1}^2}.$$

Define

$$(A.3) \quad \bar{\Delta}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \Delta_{kj}$$

and recall that $\kappa_2 = \int u^2 K(u) du$. We shall derive the following result in Section A.6:

LEMMA 1. *Under the conditions of Theorem 1, for some $C_1 > 0$, all $C_2 > 0$ and $k = 0, 1$,*

$$\begin{aligned} P_k \left(\sup_{\mathcal{I}} \left| \frac{U_2^2 - U_1U_3}{U_2U_0 - U_1^2} - \kappa_2 \right| > n^{-C_1} \right) &= O(n^{-C_2}), \\ P_k \left(\max_{j=1, \dots, n_k} \sup_{\mathcal{I}} \left| \frac{U_{kj2}^2 - U_{kj1}U_{kj3}}{U_{kj2}U_{kj0} - U_{kj1}^2} - \kappa_2 \right| > n^{-C_1} \right) &= O(n^{-C_2}) \end{aligned}$$

as $n \rightarrow \infty$, and for some $C_3 > 0$, all $C_2 > 0$ and $k = 0, 1$,

$$\begin{aligned} P_k \left(\sup_{\mathcal{I}} \frac{U_0^2}{U_2U_0 - U_1^2} > C_3 \right) &= O(n^{-C_2}), \\ P_k \left(\max_{j=1, \dots, n_k} \sup_{\mathcal{I}} \frac{U_{kj0}^2}{U_{kj2}U_{kj0} - U_{kj1}^2} > C_3 \right) &= O(n^{-C_2}). \end{aligned}$$

Furthermore, defining $M_{\text{sum}} = \min_{k=1,2} (\sum_j m_{kj})$, we have for all $C_2, C_4 > 0$,

$$(A.4) \quad \begin{aligned} &P_k \left\{ \sup_{\mathcal{I}} |\Delta| > n^{C_4} (mh)^{-1/2} \right\} \\ &+ \max_{k=0,1} P_k \left\{ \sup_{\mathcal{I}} |\bar{\Delta}_k| > n^{C_4} (M_{\text{sum}}h)^{-1/2} \right\} \\ &= O(n^{-C_2}). \end{aligned}$$

A.2. Initial calculation of err_k . Let \mathcal{G}_1 denote the sigma-field generated by the random variables introduced in Section 2, and the random functions g_{kj} , but excluding g . Specifically, \mathcal{G}_1 is the sigma-field generated by g_{kj} , X_{kji} and ε_{kji} for $1 \leq i \leq m_{kj}$, $1 \leq j \leq n_k$ and $k = 0, 1$, and by X_i and ε_i for $1 \leq i \leq m$. Recall that \bowtie_k is $>$ or \leq according as $k = 0$ or $k = 1$, respectively, and recall formula (3.11) for the statistic $S(\hat{g})$.

Under the assumption that the new data set \mathcal{D} comes from Π_k , and conditional on \mathcal{G}_1 , \hat{g} is a Gaussian process with mean $\hat{\alpha}_k = E_k(\hat{g} | \mathcal{G}_1)$ and covariance function $\hat{\Gamma}_k$, say. In this notation,

$$(A.5) \quad \text{err}_k \equiv E_k[P_k\{S(\hat{g}) \bowtie_k 0 | \mathcal{G}_1\}] = E_k\{\Psi_k(-\hat{\beta}_k/\hat{\sigma}_k)\},$$

where, by (3.11),

$$(A.6) \quad \hat{\beta}_k = E_k\{S(\hat{g}) | \mathcal{G}_1\} = \int_{\mathcal{I}} (\hat{\mu}_1 - \hat{\mu}_0) \{2\hat{\alpha}_k - (\hat{\mu}_0 + \hat{\mu}_1)\},$$

$$(A.7) \quad \begin{aligned} \hat{\sigma}_k^2 &= \text{var}\{S(\hat{g}) | \mathcal{G}_1\} \\ &= 4 \int_{\mathcal{I}} \int_{\mathcal{I}} \{\hat{\mu}_1(x_1) - \hat{\mu}_0(x_1)\} \{\hat{\mu}_1(x_2) - \hat{\mu}_0(x_2)\} \\ &\quad \times \hat{\Gamma}_k(x_1, x_2) dx_1 dx_2. \end{aligned}$$

The probability on the left-hand side of (A.5) equals the chance that, when \mathcal{D} comes from Π_k , the classifier based on $S(\hat{g})$ makes an error and assigns \mathcal{D} to the other population.

A.3. Approximations to $\hat{\alpha}_k$, $\hat{\beta}_k$ and $\hat{\sigma}_k$. In view of (A.1),

$$(A.8) \quad \left| \hat{\alpha}_k - \left(\mu_k + \frac{1}{2} h^2 \mu_k'' \frac{U_2^2 - U_1 U_3}{U_2 U_0 - U_1^2} + \Delta \right) \right| \leq \frac{2E_k(Q)h^2 U_0^2}{U_2 U_0 - U_1^2}.$$

Noting that, for random variables A_1, A_2, B_1 and B_2 , $|\text{cov}(A_1 + A_2, B_1 + B_2) - \text{cov}(A_1, A_2)| \leq |\text{cov}(B_1, B_2)| + |\text{cov}(A_1, B_2)| + |\text{cov}(B_1, A_2)|$ where the covariances are interpreted conditionally on \mathcal{G}_1 , we deduce from (A.1) that for a constant $C_4 > 0$,

$$(A.9) \quad \begin{aligned} \sup_{x_1, x_2 \in \mathcal{I}} \left| \hat{\Gamma}_k(x_1, x_2) - \left\{ G_k(x_1, x_2) + \frac{1}{2} h^2 G_k^{(0,2)}(x_1, x_2) \frac{U_2^2 - U_1 U_3}{U_2 U_0 - U_1^2}(x_2) \right. \right. \\ \left. \left. + \frac{1}{2} h^2 G_k^{(2,0)}(x_1, x_2) \frac{U_2^2 - U_1 U_3}{U_2 U_0 - U_1^2}(x_1) \right\} \right| \\ \leq C_4 h^2 \{h^2 + E_k(Q + Q^2)\} \sup_{\mathcal{I}} \left(1 + \frac{U_0^2}{U_2 U_0 - U_1^2} \right)^2, \end{aligned}$$

where we define $G_k^{(j_1, j_2)}(x_1, x_2) = \partial^{j_1+j_2} G_k(x_1, x_2) / \partial x_1^{j_1} \partial x_2^{j_2}$. (Recall that G_k denotes the covariance of the Gaussian process g when the data \mathcal{D} are drawn from Π_k .)

With \bar{g}_k defined as at (3.1), and defining $\bar{\Delta}_k$ as at (A.3), we have, in view of (A.2), Lemma 1 and (3.6)(b), the result

$$(A.10) \quad P_k \left\{ \sup_{\mathcal{I}} \left| \hat{\mu}_k - \left(\bar{g}_k + \frac{1}{2} h_1^2 \kappa_2 \bar{g}_k'' + \bar{\Delta}_k \right) \right| > n^{-C_1} h_1^2 \right\} = O(n^{-C_2})$$

for some $C_1 > 0$ and all $C_2 > 0$. Using Rosenthal's inequality, it can be proved from (3.6) and (3.7)(c) that, for some $C_1 > 0$ and all $C_2 > 0$,

$$(A.11) \quad P_k \left(\sup_{\mathcal{I}} |\bar{g}_k'' - \mu_k''| > n^{-C_1} \right) = O(n^{-C_2}).$$

Together, (A.10) and (A.11) imply that

$$(A.12) \quad P_k \left\{ \sup_{\mathcal{I}} \left| \hat{\mu}_k - \left(\bar{g}_k + \frac{1}{2} h_1^2 \kappa_2 \mu_k'' + \bar{\Delta}_k \right) \right| > n^{-C_1} h_1^2 \right\} = O(n^{-C_2}).$$

Define $H^2 = h^2 + h_1^2$,

$$(A.13) \quad \begin{aligned} \beta_k &= \int_{\mathcal{I}} \left\{ \bar{g}_1 - \bar{g}_0 + \frac{1}{2} h_1^2 \kappa_2 (\mu_1'' - \mu_0'') + \bar{\Delta}_1 - \bar{\Delta}_0 \right\} \\ &\quad \times \left\{ 2\mu_k - (\bar{g}_0 + \bar{g}_1) + h^2 \kappa_2 \mu_k'' \right. \\ &\quad \left. - \frac{1}{2} h_1^2 \kappa_2 (\mu_0'' + \mu_1'') + 2\Delta - (\bar{\Delta}_0 + \bar{\Delta}_1) \right\}, \\ \tilde{\sigma}_k^2 &= 4 \int_{\mathcal{I}} \int_{\mathcal{I}} \left\{ \bar{g}_1 - \bar{g}_0 + \frac{1}{2} h_1^2 \kappa_2 (\mu_1'' - \mu_0'') + \bar{\Delta}_0 - \bar{\Delta}_1 \right\} (x_1) \\ &\quad \times \left\{ \bar{g}_1 - \bar{g}_0 + \frac{1}{2} h_1^2 \kappa_2 (\mu_1'' - \mu_0'') + \bar{\Delta}_0 - \bar{\Delta}_1 \right\} (x_2) \\ &\quad \times \left[G_k(x_1, x_2) + \frac{1}{2} h^2 \kappa_2 \{ G_k^{(2,0)}(x_1, x_2) + G_k^{(0,2)}(x_1, x_2) \} \right]. \end{aligned}$$

Combining Lemma 1, (A.5)–(A.9) and (A.12), we deduce that, for some $C_1 > 0$ and all $C_2 > 0$,

$$(A.14) \quad \begin{aligned} P_k(|\hat{\beta}_k - \beta_k| > n^{-C_1} H^2) &= O(n^{-C_2}), \\ P_k(|\hat{\sigma}_k^2 - \tilde{\sigma}_k^2| > n^{-C_1} H^2) &= O(n^{-C_2}). \end{aligned}$$

Observe from (A.13) that $\beta_k = \beta_{k0} + b_{k1} + \beta_{k1} + \beta_{k2} + \bar{\Delta}_2$, where β_{k0} is as at (3.2),

$$(A.15) \quad b_{k1} = \kappa_2 \int_{\mathcal{I}} (\mu_1 - \mu_0) (h^2 \mu_k'' - h_1^2 \mu_{1-k}''),$$

$$\begin{aligned}\beta_{k1} &= \int_{\mathcal{I}} (\bar{g}_1 - \bar{g}_0) \{2\Delta - (\bar{\Delta}_0 + \bar{\Delta}_1)\} \\ &\quad + \int_{\mathcal{I}} \{2\mu_k - (\bar{g}_0 + \bar{g}_1)\} (\bar{\Delta}_1 - \bar{\Delta}_0), \\ \beta_{k2} &= \int_{\mathcal{I}} \{2\Delta - (\bar{\Delta}_0 + \bar{\Delta}_1)\} (\bar{\Delta}_1 - \bar{\Delta}_0)\end{aligned}$$

and $\bar{\Delta}_2 = \beta_k - (\beta_{k0} + b_{k1} + \beta_{k1} + \beta_{k2})$. Using (A.4) it can be shown that, for some $C_1 > 0$ and all $C_2 > 0$, and when $\ell = 2$,

$$(A.16) \quad P_k(|\bar{\Delta}_\ell| > n^{-C_1} H^2) = O(n^{-C_2}).$$

Hence, noting the first result in (A.14), we have:

$$(A.17) \quad P_k\{|\hat{\beta}_k - (\beta_{k0} + b_{k1} + \beta_{k1} + \beta_{k2})| > n^{-C_1} H^2\} = O(n^{-C_2}).$$

Recall the definitions of σ_k^2 and τ_k^2 at (3.3) and (3.4), and put

$$(A.18) \quad \begin{aligned}\sigma_{k0} &= 2h^2 \kappa_2 \int_{\mathcal{I}} \int_{\mathcal{I}} (\bar{g}_1 - \bar{g}_0)(x_1) (\bar{g}_1 - \bar{g}_0)(x_2) \\ &\quad \times \{G_k^{(2,0)}(x_1, x_2) + G_k^{(0,2)}(x_1, x_2)\} dx_1 dx_2,\end{aligned}$$

$$(A.19) \quad \sigma_{k1} = 4h_1^2 \kappa_2 \int_{\mathcal{I}} \int_{\mathcal{I}} (\bar{g}_1 - \bar{g}_0)(x_1) (\mu_1 - \mu_0)''(x_2) G_k(x_1, x_2) dx_1 dx_2$$

and $\bar{\Delta}_3 = \hat{\sigma}_k^2 - (\sigma_k^2 + \sigma_{k0} + \sigma_{k1})$. Thus, $\bar{\Delta}_3$ is the term in $\bar{\Delta}_0$ and $\bar{\Delta}_1$ that arises when $\hat{\sigma}_k^2$ is expanded. Using (A.4) it can be proved that (A.16) holds when $\ell = 3$. Moreover, $\hat{\sigma}_k^2$ can be written as

$$(A.20) \quad \hat{\sigma}_k^2 = \sigma_k^2 + \sigma_{k0} + \sigma_{k1} + \bar{\Delta}_3 + \bar{\Delta}_4,$$

where, in view of the second part of (A.14), (A.16) holds in the case $\ell = 4$ and for some $C_1 > 0$ and all $C_2 > 0$.

Define $\tau_{k\ell}$ to be equal to $\sigma_{k\ell}$, at (A.18) and (A.19), when \bar{g}_0 and \bar{g}_1 on the respective right-hand sides are replaced by μ_0 and μ_1 . Then for $k = 0, 1$ and $\ell = 0, 1$, noting property (3.7)(c) on the rates of increase of n_0 and n_1 , it can be shown that for some $C_1 > 0$,

$$(A.21) \quad P_k(|\sigma_{k\ell} - \tau_{k\ell}| > n^{-C_1} h_\ell^2) = O(n^{-C_2})$$

for all $C_2 > 0$, where we define $h_0 = h$. Therefore, if $C_1 > 0$ is sufficiently small,

$$(A.22) \quad \max_{k=0,1} \max_{\ell=0,1} P_k(|\sigma_{k\ell}| > n^{-C_1}) = O(n^{-C_2})$$

for all $C_2 > 0$.

A.4. Approximation to $\widehat{\sigma}_k^{-1}$. In the notation at (A.20),

$$\frac{1}{\widehat{\sigma}_k} = \frac{1}{\tau_k} \left(1 + \frac{\sigma_k^2 - \tau_k^2}{\tau_k^2} + \frac{\sigma_{k0} + \sigma_{k1} + \bar{\Delta}_3 + \bar{\Delta}_4}{\tau_k^2} \right)^{-1/2} = s_k(\infty),$$

where, for $0 \leq r \leq \infty$,

$$s_k(r) = \frac{1}{\tau_k} \sum_{j=0}^r \sum_{\ell=0}^j \binom{-\frac{1}{2}}{j} \binom{j}{\ell} \left(\frac{\sigma_k^2 - \tau_k^2}{\tau_k^2} \right)^{j-\ell} \left(\frac{\sigma_{k0} + \sigma_{k1} + \bar{\Delta}_3 + \bar{\Delta}_4}{\tau_k^2} \right)^\ell.$$

We claim that the infinite series defined by $s_k(\infty)$ converges with probability $1 - O(n^{-C_2})$ for all $C_2 > 0$. To appreciate why, note that, by (3.6) and (3.7)(c), there exists $C_1 > 0$ such that

$$P_k(|\sigma_k^2 - \tau_k^2| > n^{-C_1}) = O(n^{-C_2})$$

for all $C_2 > 0$. Combining this property, (A.16) for $\ell = 3$ and 4, and (A.22), we deduce that, for some $C_1 > 0$ and all $C_2 > 0$,

$$P_k \left(\left| \frac{\sigma_k^2 - \tau_k^2}{\tau_k^2} \right| + \left| \frac{\sigma_{k0} + \sigma_{k1} + \bar{\Delta}_3 + \bar{\Delta}_4}{\tau_k^2} \right| \leq n^{-C_1} \right) = 1 - O(n^{-C_2}).$$

Therefore, if $C_3 > 0$ is given then $r_0 = r_0(C_3) \geq 1$ can be chosen so large that, whenever $r_0 \leq r \leq \infty$, $P_k\{|\widehat{\sigma}_k^{-1} - s_k(r)| > n^{-C_3}\} = O(n^{-C_2})$ for all $C_2 > 0$. Using this property and (A.16), again for $\ell = 3$ and 4; and employing too (A.21); we see that for some $C_1 > 0$ and all $C_2 > 0$, if r_0 is chosen sufficiently large,

$$(A.23) \quad P_k\{|\widehat{\sigma}_k^{-1} - t_k(r)| > n^{-C_1} H^2\} = O(n^{-C_2})$$

for $r \geq r_0$, where

$$(A.24) \quad t_k(r) = \frac{1}{\tau_k} \sum_{j=0}^r \sum_{\ell=0}^{\min(j,1)} \binom{-\frac{1}{2}}{j} \binom{j}{\ell} \left(\frac{\sigma_k^2 - \tau_k^2}{\tau_k^2} \right)^{j-\ell} \left(\frac{\tau_{k0} + \tau_{k1}}{\tau_k^2} \right)^\ell.$$

A.5. Approximation to $E_k\{\Psi_k(-\widehat{\beta}_k/\widehat{\sigma}_k)\}$. Let $C_1 > 0$ and let $\ell_0 \geq 0$ be an integer. With $U_{kj\ell}$ defined as at (2.6), let \mathcal{E} denote the event

$$\mathcal{E} = \mathcal{E}(C_1, \ell_0) = \left\{ \max_{1 \leq \ell \leq \ell_0} \max_{j=1, \dots, n_k} \sup_{x \in \mathcal{I}} |U_{kj\ell}(x) - \kappa_\ell f_X(x)| \leq n^{-C_1} \right\},$$

where $\kappa_\ell = \int u^\ell K(u) du$ and hence vanishes for odd ℓ , since by (3.7)(b), K is symmetric. It will be proved in Section A.6 that, for some $C_1 > 0$ and each $\ell_0 \geq 0$,

$$(A.25) \quad P_k\{\mathcal{E}(C_1, \ell_0)\} = 1 - O(n^{-C_2}) \quad \text{for all } C_2 > 0.$$

If $\mathcal{E}(C_1, \ell_0)$ holds for an $\ell_0 \geq 2$ then, if $0 < C'_1 < C_1$, there exists a nonrandom integer $n_0 \geq 1$ such that the event $\mathcal{E}_1 = \mathcal{E}_1(C'_1)$, defined by

$$(A.26) \quad \mathcal{E}_1 = \left\{ \max_{j=1, \dots, n_k} \sup_{x \in \mathcal{I}} |U_{kj2}(x)U_{kj0}(x) - U_{kj1}(x)^2 - \kappa_2 f_X(x)^2| \leq n^{-C'_1} \right\}$$

holds for all $n \geq n_0$.

Let $I = I(\mathcal{E})$ denote the indicator of \mathcal{E} . In view of (A.25),

$$(A.27) \quad E_k\{\Psi_k(-\widehat{\beta}_k/\widehat{\sigma}_k)\} = E_k\{I\Psi_k(-\widehat{\beta}_k/\widehat{\sigma}_k)\} + O(n^{-C_2})$$

for all $C_2 > 0$, and so to approximate the term on the left-hand side of (A.27) we may develop an approximation to the first term on the right-hand side.

Let \mathcal{G}_2 denote the sigma-field generated by the random variables X_i for $1 \leq i \leq m$, and by X_{kji} and the functions g_{kji} for $1 \leq i \leq m_{kj}$, $1 \leq j \leq n_k$ and $k = 0, 1$ (i.e., generated by everything except g and the experimental errors ε_i and ε_{kji}). The quantities I , $t_k(r)$ at (A.24), β_{k0} at (3.2), and b_{k1} at (A.15) are all \mathcal{G}_2 -measurable. Therefore, using (A.17) and (A.23), and noting that Ψ_k is an analytic function with all derivatives uniformly bounded, we obtain

$$(A.28) \quad \begin{aligned} & E_k\{I\Psi_k(-\widehat{\beta}_k/\widehat{\sigma}_k)\} \\ &= E_k(E_k[I\Psi_k\{-(\beta_{k0} + b_{k1} + \beta_{k1} + \beta_{k2})t_k(r)\} \mid \mathcal{G}_2]) + o(H^2) \\ &= E_k[I\Psi_k\{-\beta_{k0}t_k(r)\}] - b_{k1}\tau_k^{-1}E_k[I\Psi'_k\{-\beta_{k0}t_k(r)\}] \\ &\quad - \tau_k^{-1}E_k[E_k(\beta_{k2} \mid \mathcal{G}_2)I\Psi'_k\{-\beta_{k0}t_k(r)\}] \\ &\quad + \frac{1}{2}\tau_k^{-2}E_k[E_k(\beta_{k1}^2 \mid \mathcal{G}_2)I\Psi''_k\{-\beta_{k0}t_k(r)\}] \\ (A.29) \quad & + O\{(mh)^{-2} + (M_{\text{sum}}h_1)^{-2}\} + o(H^2). \end{aligned}$$

Here we have used the properties $E_k(\beta_{k1} \mid \mathcal{G}_2) = 0$, $E_k|t_k(r) - \tau_k^{-1}| = O(n^{-C})$ for some $C > 0$,

$$E_k[E_k(\beta_{k\ell_1}^{\ell_2} \mid \mathcal{G}_2)I\Psi_k^{(\ell_2)}\{-\beta_{k0}t_k(r)\}] = O\{(mh)^{-2} + (M_{\text{sum}}h_1)^{-2}\}$$

for $\ell_2 \geq 3$ if $\ell_1 = 1$, and for $\ell_2 \geq 2$ if $\ell_1 = 2$, and

$$|E_k[E_k(\beta_{k1}\beta_{k2} \mid \mathcal{G}_2)I\Psi''_k\{-\beta_{k0}t_k(r)\}]| = O\{(mh)^{-2} + (M_{\text{sum}}h_1)^{-2}\}.$$

Further, we have used the fact that the event \mathcal{E}_1 , defined at (A.26), obtains whenever $I \neq 0$.

In addition,

$$(A.30) \quad \begin{aligned} \frac{1}{4}E_k[E_k(\beta_{k1}^2 \mid \mathcal{G}_2)I] &= E_k\left\{I \int_{\mathcal{I}} (\bar{g}_0 - \bar{g}_1)\Delta\right\}^2 \\ &+ E_k\left\{I \int_{\mathcal{I}} (\bar{g}_0 - \mu_k)\bar{\Delta}_0\right\}^2 + E_k\left\{I \int_{\mathcal{I}} (\bar{g}_1 - \mu_k)\bar{\Delta}_1\right\}^2 \\ &= O(m^{-1}), \end{aligned}$$

that

$$\begin{aligned}
& E_k[E_k(\beta_{k2} | \mathcal{G}_2)I\Psi'_k\{-\beta_{k0}t_k(r)\}] \\
&= (-1)^{k+1}\phi(b_{k0}/\tau_k) \int_{\mathcal{I}} E_k[I\{E_k(\bar{\Delta}_0^2 | \mathcal{G}_2) - E_k(\bar{\Delta}_1^2 | \mathcal{G}_2)\}] \\
\text{(A.31)} \quad &+ o\{(\nu_0 h_1)^{-1}\} \\
&= \frac{\kappa}{h_1}(\sigma_{\varepsilon_0}^2 \nu_0^{-1} - \sigma_{\varepsilon_1}^2 \nu_1^{-1})(-1)^{k+1}\phi(b_{k0}/\tau_k) \int_{\mathcal{I}} f_X^{-1} + o\{(\nu_0 h_1)^{-1}\}
\end{aligned}$$

and that

$$\text{(A.32)} \quad b_{k1}\tau_k^{-1}E_k[I\Psi'_k\{-\beta_{k0}t_k(r)\}] = b_{k1}\tau_k^{-1}(-1)^{k+1}\phi(b_{k0}/\tau_k) + o(H^2),$$

where b_{k0} and b_{k1} are as at (3.2) and (A.15), ϕ is the standard normal density, and we have used the fact that $\Psi'_k = (-1)^{k+1}\phi$. Combining (A.25) and (A.27)–(A.32), and taking r sufficiently large (but fixed), we deduce that

$$\begin{aligned}
\text{(A.33)} \quad E_k\{\Psi_k(-\widehat{\beta}_k/\widehat{\sigma}_k)\} &= E_k[\Psi_k\{-\beta_{k0}/\sigma_k\}] - b_{k1}\tau_k^{-1}(-1)^{k+1}\phi(b_{k0}/\tau_k) \\
&\quad - \frac{\kappa}{\tau_k h_1}(\sigma_{\varepsilon_0}^2 \nu_0^{-1} - \sigma_{\varepsilon_1}^2 \nu_1^{-1})(-1)^{k+1}\phi(b_{k0}/\tau_k) \int_{\mathcal{I}} f_X^{-1} \\
&\quad + O\{m^{-1} + (mh)^{-2}\} + o\{H^2 + (\nu_0 h_1)^{-1}\}.
\end{aligned}$$

Result (3.8) follows from (A.5) and (A.33).

A.6. Proof of Lemma 1 and (A.25). The results in Lemma 1, with the exception of (A.4); and also result (A.25); will follow if we show that for each $\ell \geq 1$, some $C_1 > 0$ and all $C_2 > 0$,

$$\text{(A.34)} \quad P_k \left\{ \sup_{x \in \mathcal{I}} |U_\ell(x) - \kappa_\ell f_X(x)| > n^{-C_1} \right\} = O(n^{-C_2}),$$

$$\text{(A.35)} \quad P_k \left\{ \max_{j=1, \dots, n_k} \sup_{x \in \mathcal{I}} |U_{kj\ell}(x) - \kappa_\ell f_X(x)| > n^{-C_1} \right\} = O(n^{-C_2}).$$

We shall derive (A.35); a proof of (A.34) is similar.

Markov's inequality can be used to prove that

$$\text{(A.36)} \quad \max_{j=1, \dots, n_k} \sup_{x \in \mathcal{I}} P_k \{ |U_{kj\ell}(x) - \kappa_\ell f_X(x)| > n^{-C_1} \} = O(n^{-C_2}).$$

It follows from (3.7)(c) that each n_k is increasing no faster than polynomially in n , and therefore, if we confine attention to x in a subset \mathcal{I}_n , say, of \mathcal{I} that contains only $O(n^C)$ points for some $C > 0$, we can place the maximum and supremum inside the probability statement at (A.36), provided that \mathcal{I} is replaced by \mathcal{I}_n : for some $C_1 > 0$ and all $C_2 > 0$,

$$\text{(A.37)} \quad P_k \left\{ \max_{j=1, \dots, n_k} \sup_{x \in \mathcal{I}_n} |U_{kj\ell}(x) - \kappa_\ell f_X(x)| > n^{-C_1} \right\} = O(n^{-C_2}).$$

The assumption, in (3.7)(b), that K is compactly supported and Hölder continuous, and the implication, in (3.5)(c), that f_X is also Hölder continuous, enable (A.35) to be derived directly from (A.37) by taking \mathcal{I}_n to be a sufficiently fine grid in \mathcal{I} .

A proof of (A.4) in Lemma 1 is similar. To illustrate the argument, we derive the following result part of (A.4): for all $C_2, C_4 > 0$,

$$(A.38) \quad P_k \left\{ \sup_{x \in \mathcal{I}} |\Delta(x)| > n^{C_4} (mh)^{-1/2} \right\} = O(n^{-C_2}).$$

Using Markov's and Rosenthal's inequalities, we first obtain the result when the supremum is outside the probability statement:

$$\sup_{x \in \mathcal{I}} P_k \{ |\Delta(x)| > n^{C_4} (mh)^{-1/2} \} = O(n^{-C_2}).$$

Taking \mathcal{I}_n to contain only $O(n^C)$ points, for any fixed $C > 0$, we deduce that

$$P_k \left\{ \sup_{x \in \mathcal{I}_n} |\Delta(x)| > n^{C_4} (mh)^{-1/2} \right\} = O(n^{-C_2}),$$

and taking \mathcal{I}_n to be a sufficiently fine grid in \mathcal{I} we obtain (A.38).

SUPPLEMENTARY MATERIAL

Supplement to “Unexpected properties of bandwidth choice when smoothing discrete data for constructing a functional data classifier”

(DOI: [10.1214/13-AOS1158SUPP](https://doi.org/10.1214/13-AOS1158SUPP); .pdf). The supplementary file contains the proof of Theorems 2 and 3, as well as additional simulation results.

REFERENCES

- ARAKI, Y., KONISHI, S., KAWANO, S. and MATSUI, H. (2009). Functional logistic discrimination via regularized basis expansions. *Comm. Statist. Theory Methods* **38** 2944–2957. [MR2568196](#)
- BENHENNA, K. and DEGRAS, D. (2011). Local polynomial regression based on functional data. Unpublished manuscript. Available at <http://arxiv.org/pdf/1107.4058v1>.
- BERLINET, A., BIAU, G. and ROUVIÈRE, L. (2008). Functional supervised classification with wavelets. *Ann. I.S.U.P.* **52** 61–80. [MR2435041](#)
- BIAU, G., BUNEA, F. and WEGKAMP, M. H. (2005). Functional classification in Hilbert spaces. *IEEE Trans. Inform. Theory* **51** 2163–2172. [MR2235289](#)
- CARDOT, H., DEGRAS, D. and JOSSEAND, E. (2013). Confidence bands for Horvitz–Thompson estimators using sampled noisy functional data. *Bernoulli* **19** 2067–2097.
- CARDOT, H. and JOSSEAND, E. (2011). Horvitz–Thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika* **98** 107–118. [MR2804213](#)
- CARROLL, R. J., DELAIGLE, A. and HALL, P. (2013). Supplement to “Unexpected properties of bandwidth choice when smoothing discrete data for constructing a functional data classifier.” DOI:[10.1214/13-AOS1158SUPP](https://doi.org/10.1214/13-AOS1158SUPP).

- CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Comput. Statist.* **22** 481–496. [MR2336349](#)
- DELAIGLE, A. and HALL, P. (2012). Achieving near perfect classification for functional data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74** 267–286. [MR2899863](#)
- DELAIGLE, A., HALL, P. and BATHIA, N. (2012). Componentwise classification and clustering of functional data. *Biometrika* **99** 299–313. [MR2931255](#)
- EPIFANIO, I. (2008). Shape descriptors for classification of functional data. *Technometrics* **50** 284–294. [MR2528652](#)
- FAN, J. and GJJBELS, I. (1996). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. [MR1383587](#)
- FROMONT, M. and TULEAU, C. (2006). Functional classification with margin conditions. In *Learning Theory—Proceedings of the 19th Annual Conference on Learning Theory, Pittsburgh, 2006* (J. G. CARBONELL and J. SIEKMANN, eds.). Springer, New York.
- HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 109–126. [MR2212577](#)
- HALL, P. and HOSSEINI-NASAB, M. (2009). Theory for high-order bounds in functional principal components analysis. *Math. Proc. Cambridge Philos. Soc.* **146** 225–256. [MR2461880](#)
- HALL, P. and KANG, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.* **33** 284–306. [MR2157804](#)
- HALL, P. and VAN KEILEGOM, I. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statist. Sinica* **17** 1511–1531. [MR2413533](#)
- LENG, X. and MÜLLER, H.-G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22** 68–76.
- LI, Y. and HSING, T. (2010a). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *Ann. Statist.* **38** 3028–3062. [MR2722463](#)
- LI, Y. and HSING, T. (2010b). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38** 3321–3351. [MR2766854](#)
- LÓPEZ-PINTADO, S. and ROMO, J. (2006). Depth-based classification for functional data. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **72** 103–119. Amer. Math. Soc., Providence, RI. [MR2343116](#)
- MANNING, C. D., RAGHAVAN, P. and SCHÜTZ, H. (2008). *Introduction to Information Retrieval*. Cambridge Univ. Press, Cambridge.
- PANARETOS, V. M., KRAUS, D. and MADDOCKS, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *J. Amer. Statist. Assoc.* **105** 670–682. [MR2724851](#)
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. [MR2168993](#)
- ROSSI, F. and VILLA, N. (2006). Support vector machine for functional data classification. *Neurocomputing* **69** 730–742.
- RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270. [MR1379468](#)
- VILAR, J. A. and PÉRTEGA, S. (2004). Discriminant and cluster analysis for Gaussian stationary processes: Local linear fitting approach. *J. Nonparametr. Stat.* **16** 443–462. [MR2073035](#)
- WANG, X., RAY, S. and MALLICK, B. K. (2007). Bayesian curve classification using wavelets. *J. Amer. Statist. Assoc.* **102** 962–973. [MR2354408](#)

WU, S. and MÜLLER, H.-G. (2011). Response-adaptive regression for longitudinal data.
Biometrics **67** 852–860. [MR2829259](#)

R. J. CARROLL
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843
USA
E-MAIL: carroll@stat.tamu.edu

A. DELAIGLE
P. HALL
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF MELBOURNE, PARKVILLE
VICTORIA 3010
AUSTRALIA
E-MAIL: A.Delaigle@ms.unimelb.edu.au
halpstat@ms.unimelb.edu.au