Biometrika (2002), **89,** 1, *pp.* 129–143

Printed in Great Britain

The discrimination power of projection pursuit with different density estimators

BY OLIVIER RENAUD

Faculty of Psychology and Education, University of Geneva, Bd du Pont d'Arve 40, 1211 Geneva 4, Switzerland

olivier.renaud@pse.unige.ch

SUMMARY

We explore the properties of projection pursuit discriminant analysis. This discriminant method is very powerful but relies heavily on a univariate density estimate. We show that the procedure based on wavelets maintains the same rate of convergence as with univariate wavelet density estimation. We also show the Bayes risk strong consistency of both the kernel- and wavelet-based methods. Simulated data and real data concerning character recognition show that the method is effective and robust against the curse of dimensionality. The wavelet alternative seems more likely than the kernel counterpart to find an interesting projection. Wavelets are often criticised for giving too wiggly an estimate and for being too localised to give good global properties. In the above context, these potential drawbacks do not weaken the method but the use of wavelets seems to enhance it. A multiple projection generalisation is also considered.

Some key words: Bayes risk; Discriminant analysis; Kernel; Minimax; Strong consistency; Wavelet.

1. INTRODUCTION

Wavelet regression and density estimation in one dimension enjoy properties such as minimax and adaptive estimation not shared with other nonparametric methods. Much less is known about multivariate aspects of wavelet-based estimators. This paper discusses a multivariate application of wavelet density estimation and compares its performance with a kernel-based method.

For signal and image processing, there exist widely applied two- and three-dimensional generalisations of wavelets that preserve the good properties of the univariate case (Vetterli & Kovačević, 1995). For example, wavelet-based methods are allowed in the standard for the video coder MPEG 4. However, for statistical applications, one often needs methods that work in much higher dimensions. The curse of dimensionality implies that a wavelet basis will not capture the general behaviour of the underlying process. In fact, no local method, including kernels and splines, is well adapted to this problem. As a consequence, almost any statistical procedure involves some kind of dimension reduction.

The multivariate setting is however a good test for any nonparametric method. The procedure described in § 2·3, called projection pursuit discriminant analysis, is a projective method for discriminant analysis and needs a density estimator as a building block. In this paper, we will use and compare wavelet and convolution kernel density estimators in this framework. The projection pursuit discriminant analysis method is versatile but,

numerically, relies heavily on the properties of the density estimator. It should thus give insight about the strengths and weaknesses of wavelet and kernel methods. In § 3, we show that the procedure based on wavelets enjoys the same rate of convergence as wavelet density estimators in one dimension. The Bayes risk strong consistency of both the wavelet and the kernel-based methods is also established. The numerical aspects of the procedure are of great importance, since it provides only an implicit form, and the function to be optimised is complicated. These concerns are discussed in § 4, along with simulation studies. The method is then applied in § 5 in the context of character recognition. Section 6 treats an extension of the method that involves iterative selection of projections to improve discrimination power.

2. BACKGROUND

2·1. *Density estimation with wavelets*

For a complete coverage of wavelet theory in statistics, we refer to Härdle et al. (1998) and Vidakovic (1999). For the density-estimation setting let $(X_1, \ldots, X_N)$ be a vector of independent and identically distributed observations with density $f \in L_2$. As in regression, the estimator of f,

$$\hat{f}(x) = \sum_{k \in Z} \hat{\alpha}(j_0, k)\phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1} \sum_{k \in Z} \hat{\beta}(j, k)\psi_{jk}(x),$$

is expanded on the truncation of an orthonormal basis of L_2 based on $\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. In density estimation, $\hat{\beta}(j, k)$ estimates $E\{\psi_{jk}(X)\}$. The natural moment-based estimator is $\hat{\beta}(j, k) = \int \psi_{jk}(y)\, dF_N(y)$, where F_N is the empirical distribution function. A similar estimator is obtained for $\hat{\alpha}(j, k)$. However, no real noise reduction has yet been made. The first regularisation is a projection that sets all the β's to zero, leading to the linear wavelet density estimator. A more powerful regularisation that shrinks the $\hat{\beta}(j, k)$ towards zero leads to the thresholded wavelet density estimator. The amount of shrinkage has to be selected, either globally or as a function of the level j, and represents the trade-off between bias and variance.

In a series of papers by G. Kerkyacharian, D. Picard and I. M. Johnstone, summarised in Härdle et al. (1998, Ch. 10, 11), the properties of the linear and the thresholded density estimators are explored. In particular, their L_p risk can be bounded, as recalled here in the linear case.

THEOREM 1. *Let $\mathscr{B}(s, p, q, L)$ be the ball of densities whose Besov B^s_{pq} norms are smaller than L. Let ϕ be a scaling function such that, for some $M \in N^*$, $|\phi(x)| < \Omega(|x|)$, with $|x|^M \Omega(|x|) \in L_1$, and $\Omega \in L_p \cap L_1$ is bounded, symmetric around 0, and nonincreasing for positive values. If $0 < s < M$, $2 \leqslant p < \infty$, $1 \leqslant q < \infty$ and $2^{j_0} \backsimeq N^{1/(2s+1)}$, we have*

$$\sup_{f \in \mathscr{B}(s,p,q,L)} E\|\hat{f} - f\|^p_{L_p} < CN^{-sp/(2s+1)}$$

for some constant $C > 0$. This still holds for $1 \leqslant p < 2$ if in addition $f(x) < w(x)$, where $w \in L_{p/2}$, is symmetric around a point x_0 and is nonincreasing for $x > x_0$.

2·2. *Discriminant analysis*

The aim of discriminant analysis is to classify an observation into one of G different populations $\mathscr{G}_1, \ldots, \mathscr{G}_G$. If M continuous measurements $x = (x_1, \ldots, x_M)$ are recorded on

an individual, the assignment rule partitions R^M into G mutually exclusive regions $\mathscr{R}_1, \ldots, \mathscr{R}_G$ that define the group membership. We denote by $\mathscr{R}$ the set of the G regions. We suppose that each group possesses a density f_g, and we denote the a priori probability or proportion of group g by π_g. The quality of a rule is given by the global probability of misclassification of an individual X randomly selected from the entire population, which is given by

$$\sum_{g=1}^{G} \pi_g \, \mathrm{pr}(X \text{ not assigned to } \mathscr{G}_g \mid X \in \mathscr{G}_g) = 1 - \sum_{g=1}^{G} \int_{\mathscr{R}_g} \pi_g f_g(x) \, dx. \tag{2·1}$$

The best rule, Bayes' rule, corresponds to $\mathscr{R}_g = \{x \in R^M \mid \pi_g f_g(x) > \pi_i f_i(x), \text{ for all } i \neq g\}$. Various allocation rules try to imitate the best rule, using data from N individuals of known origins. The simplest is Fisher's discriminant function. If there are two groups, the only information about a point x used by this rule is a one-dimensional projection of x. Recent nonparametric methods in discriminant analysis also resort to some form of dimension reduction.

To our knowledge wavelets have been used in discriminant analysis only when the objects to be discriminated are one-dimensional curves and not multi-dimensional points; see Coifman & Saito (1994) and a Stanford University technical report by J. Buckheit and D. L. Donoho.

2·3. *Projection pursuit*

Projection pursuit (Friedman & Tukey, 1974) seeks interesting projections of high-dimensional data in one, two or three dimensions, as a tool for looking at the hidden structure of a dataset, such as clusters, outliers and so on. The value of a projection, a point θ in the unit sphere $\mathscr{S}^{M-1}$, is called the projection index and is denoted by $\Pi(\theta)$. In exploratory analysis, most projections of a high-dimensional dataset appear as a random sample drawn from a Gaussian density (Diaconis & Freedman, 1984). Thus, almost any index proposed in the literature is a measure of distance between the density of the projected points and a Gaussian density. Projection pursuit includes many other situations, such as regression (Friedman & Stuetzle, 1981), where dimension reduction is present.

Posse (1992) develops a technique for group discrimination, called projection pursuit discriminant analysis for estimating Bayes' rule. Since it is not manageable to estimate the rule in a nonparametric way from the dataset itself, its projections will be used. Suppose $f_g^\theta(x)$ is the marginal density of the group $\mathscr{G}_g$ on the projected space defined by θ. Then the projection index is its ability to separate the groups, i.e. the global probability of misclassification under projection on θ. If the sample is randomly drawn from the entire population, one can estimate the proportions π_g by the sample proportions, and the empirical projection index is

$$\hat{\Pi}(\theta) = 1 - \int_R \max_{g=1,\ldots,G} \{\hat{\pi}_g \hat{f}_g^\theta(x)\} \, dx, \tag{2·2}$$

where $\hat{f}_g^\theta$ is a suitable estimator of f_g^θ, based on the projection of the training sample. As we project on a one-dimensional space, a nonparametric density estimator can exhibit all its power without being hampered by the dimensionality problem. We will use and compare the wavelet and the kernel density estimators.

Even in moderate dimension this procedure poses a complex problem for any nonpara-

metric density estimator. On the one hand, the estimated projection index, as a function of the angle, must be as smooth as possible and not display many more local minima than does the true projection index. On the other hand, it has to take advantage of its nonparametric nature to find genuine differences between the groups.

In the following sections, we use the linear and the thresholded wavelet density estimator in the empirical projection index. We will prove the convergence of the estimator to the true optimiser of the projection index and we will compare the wavelet procedure with that obtained using a kernel density estimator.

Note that wavelets do not give bona-fide density estimates $\hat{f}$. They generally do not integrate to unity and may take negative values. Moreover, the usual wavelet-based projection index is neither affine nor scale invariant; see two unpublished reports by the author. This implies that densities obtained by projecting on θ or on $-\theta$ differ. However, these issues affect neither the optimisation algorithm nor the convergence properties given in the next section.

3. CONVERGENCE OF THE ESTIMATED RULES

Let

$$\Pi(\mathscr{R}, \theta) = 1 - \sum_{g=1}^{G} \int_{\mathscr{R}_g} \pi_g f_g^\theta(x)\, dx, \quad \hat{\Pi}(\mathscr{R}, \theta) = 1 - \sum_{g=1}^{G} \int_{\mathscr{R}_g} \hat{\pi}_g \hat{f}_g^\theta(x)\, dx$$

be the true and the estimated projection indices in the direction θ and for the set of univariate regions $\mathscr{R} = \{\mathscr{R}_1, \ldots, \mathscr{R}_G\}$. Let $\sup^\#$ denote the supremum over all $\theta \in \mathscr{S}^{M-1}$ and over all Borel partitions $\mathscr{R}$ of R in G groups.

The infimum over $\mathscr{R}$ and θ of both Π and $\hat{\Pi}$ is attained for fixed θ by the Bayes' rule in the univariate case, and minimisation with respect to θ is over a compact space. We will denote by $\mathscr{R}^*$ and θ^* the possibly not unique optimal choices, i.e. such that $\Pi(\mathscr{R}^*, \theta^*) = \min^\# \Pi(\mathscr{R}, \theta)$. Analogously, $\hat{\mathscr{R}}$ and $\hat{\theta}$ are such that $\hat{\Pi}(\hat{\mathscr{R}}, \hat{\theta}) = \min^\# \hat{\Pi}(\mathscr{R}, \theta)$. The following theorem gives the rate of convergence of the estimated projection index.

THEOREM 2. *Let $\{f_g\}$ be the M-dimensional densities of the G groups. Suppose that all the marginal densities f_g^θ are in the Besov ball $\mathscr{B}(s, 1, q, L)$ for some values s, q and L, and satisfy $f_g^\theta(x) < w(x)$ for $p = 1$; see Theorem 1. These restrictions are for example met if the marginal densities are compactly supported and of bounded variation, or in a Sobolev space H_1^s. Suppose also that the wavelet basis and the level j_0 satisfy all the conditions of Theorem 1. Then, for increasing values of N,*

$$\sup^\# E|\hat{\Pi}(\mathscr{R}, \theta) - \Pi(\mathscr{R}, \theta)| = O(N^{-s/(2s+1)}).$$

Proof. It is not difficult to show that

$$|\hat{\Pi}(\mathscr{R}, \theta) - \Pi(\mathscr{R}, \theta)| \leqslant \sum_{g=1}^{G} |\hat{\pi}_g - \pi_g| + \sum_{g=1}^{G} \int_R |\hat{f}_g^\theta(x) - f_g^\theta(x)|\, dx.$$

This bound does not depend on $\mathscr{R}$. This implies that the $\sup^\# E(.)$ of the left-hand side is dominated by two terms. The first term has the parametric rate of convergence $O(N^{-\frac{1}{2}})$ whereas Theorem 1 shows that the second is $O(N^{-s/(2s+1)})$. □

Stronger results can be obtained for the thresholded estimator if in the proof one uses Theorem 10.4 or Proposition 11.1 in Härdle et al. (1998) instead of Theorem 1.

A rule is said to be Bayes risk strongly consistent if its probability of misclassification converges almost surely to the probability of misclassification of the best rule. With very general assumptions on f_g, one can prove that the procedures based on wavelets or on convolution kernels are consistent. This is a consequence of the following theorem.

THEOREM 3. *Let f be a uniformly continuous M-dimensional density that vanishes outside a ball. Let $\hat{f}^\theta$ denote either the linear wavelet or the convolution kernel estimator of the marginal density f^θ. Suppose for the wavelet estimator that the scaling function is of bounded variation and satisfies the condition in Theorem* 1 *for $M = p = 1$, and that the level $j_0(N)$ grows at least as rapidly as $2^{j_0} = O\{(N/\log^{1+\gamma} N)^{\frac{1}{2}}\}$ for some $\gamma > 0$. Suppose for the kernel estimator that the kernel K is continuous with bounded variation, and that the bandwidth h satisfies $h \to 0$ and $Nh^2/\log^{1+\gamma} N \to \infty$ when $N \to \infty$, for some $\gamma > 0$. Then it follows that*

$$\sup_{\theta \in \mathscr{S}^{M-1}} \sup_{x \in R} |\hat{f}^\theta(x) - f^\theta(x)| \to 0, \tag{3·1}$$

$$\sup_{\theta \in \mathscr{S}^{M-1}} \int_R |\hat{f}^\theta(x) - f^\theta(x)|\, dx \to 0, \tag{3·2}$$

almost surely, as $N \to \infty$. This implies that, if the M-dimensional densities f_g for all groups satisfy the above assumptions, then

$$\sup^{\#} |\hat{\Pi}(\mathscr{R}, \theta) - \Pi(\mathscr{R}, \theta)| \to 0, \tag{3·3}$$

almost surely, as $N \to \infty$.

The proof is in the Appendix. Note that, among the usual wavelet bases, only the sinc wavelet does not satisfy the assumptions of Theorem 3. Moreover, the minimal rate for j_0 is not a constraint since it is always satisfied for the usual choices. For compactly supported convolution kernels, E. Elguero, in a 1988 Ph.D. thesis from l'Université des Sciences et Techniques du Languedoc proves equation (3·1) and Posse (1992) proves equation (3·3). The compactness hypothesis simplifies the proof of (3·1) considerably, since it implies the compactness of the estimator. Furthermore, (3·2) follows immediately from (3·1) in this case. However, this hypothesis is necessary for neither convolution kernels nor wavelets.

Theorem 3 proves the desired Bayes risk strong consistency of the estimated rule, using Posse (1992) who shows the following corollary in the case of compact convolution kernel estimation. As his proof depends only on the almost sure convergence of the projection index, it applies to the estimators described in Theorem 3.

COROLLARY 1. *With the same assumptions as in Theorem* 3, *the rule defined by $\{\hat{\mathscr{R}}, \hat{\theta}\}$ is Bayes risk strongly consistent. Moreover, the estimated error rate converges almost surely to the ideal value, that is $\Pi(\hat{\mathscr{R}}, \hat{\theta}) \to \Pi(\mathscr{R}^*, \theta^*)$ and $\hat{\Pi}(\hat{\mathscr{R}}, \hat{\theta}) \to \Pi(\mathscr{R}^*, \theta^*)$, almost surely, as $N \to \infty$.*

Note that, by the boundedness of Π and $\hat{\Pi}$, the almost sure convergences imply not only convergence in probability but also L_p convergence for any $p > 0$.

4. NUMERICAL ASPECTS AND SIMULATIONS

4·1. *An optimisation algorithm*

As Friedman stated in his discussion of Jones & Sibson (1987), the power of a projection pursuit procedure depends crucially on the reliability and thoroughness of the numerical

optimiser. Very simple two-dimensional examples show that empirical indices are very likely to have additional local minima, a problem we might expect to be even worse in higher dimensions.

Algorithms like steepest descent are too strongly influenced by local minima. Huber, reported in Posse (1995), defined an algorithm that performs local and global searches at the same time: given the current direction of the projection θ, $\hat{\Pi}$ is computed for six neighbouring directions, i.e. local search, and two global directions. If one of them has a smaller $\hat{\Pi}$, it replaces θ. The procedure is repeated and, if no improvement is recorded during $\eta = 50$ steps, the size of the local neighbourhood is halved. The procedure is stopped when the neighbourhood is small enough. The procedure is repeated with 10 different starting values.

Concerning the wavelet estimator, any non-continuous thresholding rule, like the hard threshold, has to be avoided at all costs. It leads to many discontinuities in the index, when we move continuously from one projection to another; any raw coefficient will vary continuously with the projection, but its thresholded version with such a rule can become discontinuous, and each discontinuity will be transmitted to the empirical projection index, creating 'nuisance' local minima. Simulations shows that a huge number of such discontinuities are present even in very simple examples. The soft threshold behaves much better, being a continuous transformation.

4·2. *Simulation results*

The procedure of projection pursuit discriminant analysis can be divided into two sub-problems. The first is to select a direction θ that best separates the marginal densities, and the second is to define the univariate regions $\mathscr{R}_g$ on this projection so as to minimise the probability of misclassification. Of course, both steps can be done with the same estimators $\hat{f}_g$, in which case the angle is defined as the minimiser of (2·2) and $\mathscr{R}_g = \{x \in R | \hat{\pi}_g \hat{f}_g^{\theta}(x) > \hat{\pi}_i \hat{f}_i^{\theta}(x) \text{ for all } i \neq g\}$. However, one can gain by separating the two problems; the estimator for the first step takes advantage of being less smooth, being composed of integrals of densities. For the second step, a smoother estimator is not only more efficient but also more convenient for interpretation, since it has more chances to give rise to connected regions. Therefore, to obtain a correct comparison between estimators, one should separate the two problems. As the second sub-problem is one-dimensional, we will concentrate in this section on the first one. In the following simulation study, we compare wavelet density estimators with kernel estimators in their ability to select an interesting direction.

For the kernel density estimator, we use the two-window procedure of Sheather & Jones (1991) to estimate the bandwidth and we use the cosine kernel. For the wavelet density estimator, we use the basis of the Daubechies family, extremal phase, with four vanishing moments. The level J is selected as $2^J = \lfloor N/\log N \rfloor$, and we used universal thresholding with soft threshold.

Since our index is an integral form of the estimated densities, the choices for the smoothing parameters should typically be smaller. However, the numerical optimisation of the index would benefit from larger smoothing parameters, that smooth the function being optimised. It is not clear which is more important. For that reason, for the kernel estimate, we tried bandwidths that are twice as small as Sheather & Jones' selection rule, $h/2$, four times smaller, $h/4$ and twice as large, $2h$. We also tried more extreme values,

but the performances were poorer. For the wavelet estimate, we show the results for the universal threshold, w, and half its value, $w+$.

Details of the simulation are given in Table 1 and Fig. 1. To explain how our study results are presented, consider the example of Fig. 2(a). The number of groups, the number of observations per group and the densities of the groups are given in Table 1. In this case, it is a two-group four-dimensional example. The first direction is discriminatory as the marginal distributions are Gaussian with different means, whereas the directions in the orthogonal three-dimensional subspace are not discriminatory. The two weighted marginal densities are sketched in Fig. 1(a), showing how well the groups are separated in the discriminant direction. The global probability of misclassification for this direction, $\Pi(\theta)$, is the hatched zone and is equal to $\Phi(-1{\cdot}3/2) \simeq 0{\cdot}258$, which is the minimum value over all projections; Φ is the distribution function of the standard normal.

Table 1. *Details of the scenarios for the simulation study*

Associated figure	Group	Sample size	X_1 distribution	Other variables	Distribution
Fig. 1(a)	1	50	$\mathcal{N}(0, 1)$	X_2, X_3, X_4	$\mathcal{N}(0, 1)$
	2	50	$\mathcal{N}(1{\cdot}3, 1)$	X_2, X_3, X_4	$\mathcal{N}(0, 1)$
Fig. 1(b)	1	200	$\frac{1}{2}\mathcal{N}(-2, 1)+\frac{1}{2}\mathcal{N}(2, 1)$	X_2, X_3, X_4	$\mathcal{N}(0, 1)$
	2	100	$\mathcal{N}(0, 1)$	X_2, X_3, X_4	$\mathcal{N}(0, 1)$
Fig. 1(c)	1	60	$\frac{1}{2}\mathcal{N}(-2, 1)+\frac{1}{2}\mathcal{N}(2, 1)$	X_2, X_3, X_4	$\mathcal{N}(0, 1)$
	2	60	$\mathcal{N}(0, 1)$	X_2, X_3, X_4	$\mathcal{N}(0, 1)$
Fig. 1(d)	1	260	$\frac{1}{2}\mathcal{N}(-2, 1)+\frac{1}{2}\mathcal{N}(2, 1)$	$X_2, \ldots, X_{10}$	$\mathcal{N}(0, 1)$
	2	260	$\mathcal{N}(0, 1)$	$X_2, \ldots, X_{10}$	$\mathcal{N}(0, 1)$
Fig. 1(e)	1	50	Be(1·5, 5)	X_2	Un(0, 1)
	2	50	Be(5, 5)	X_2	Un(0, 1)
	3	50	Be(5, 1·5)	X_2	Un(0, 1)

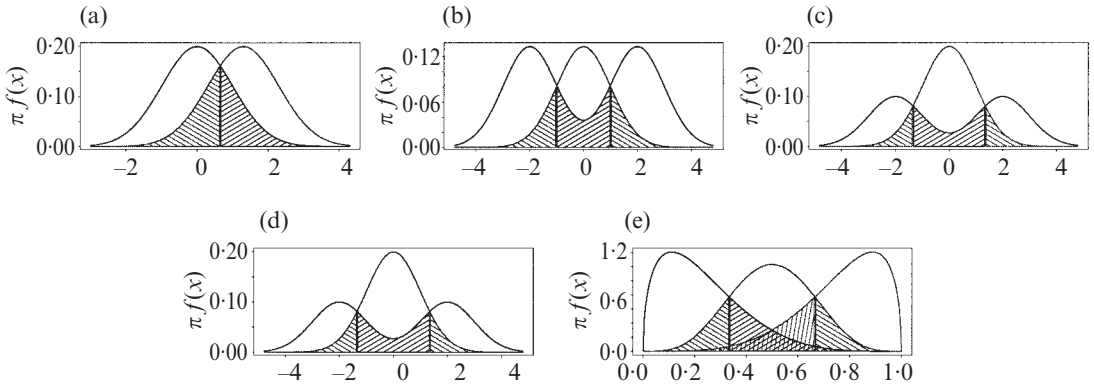

Fig. 1. Models for the simulation displayed in Fig. 2, showing weighted marginal densities for discriminant projections in the groups, with shaded areas denoting misclassification probabilities.

For each simulation, 50 observations of each group are generated. The kernel-based procedure returns the direction $\hat{\theta}$ that minimises its $\hat{\Pi}$. We do the same for the three other kernel-based methods and the two wavelet-based methods.

Since we know the underlying distributions, we can compute the capability of discrimination of the direction selected by any method by reporting $\Pi(\hat{\theta})$. In this example $\Pi(\theta) = \Phi\{-1{\cdot}3\cos(\rho)/2\}$, where ρ is the angle between θ and the discriminant direction.

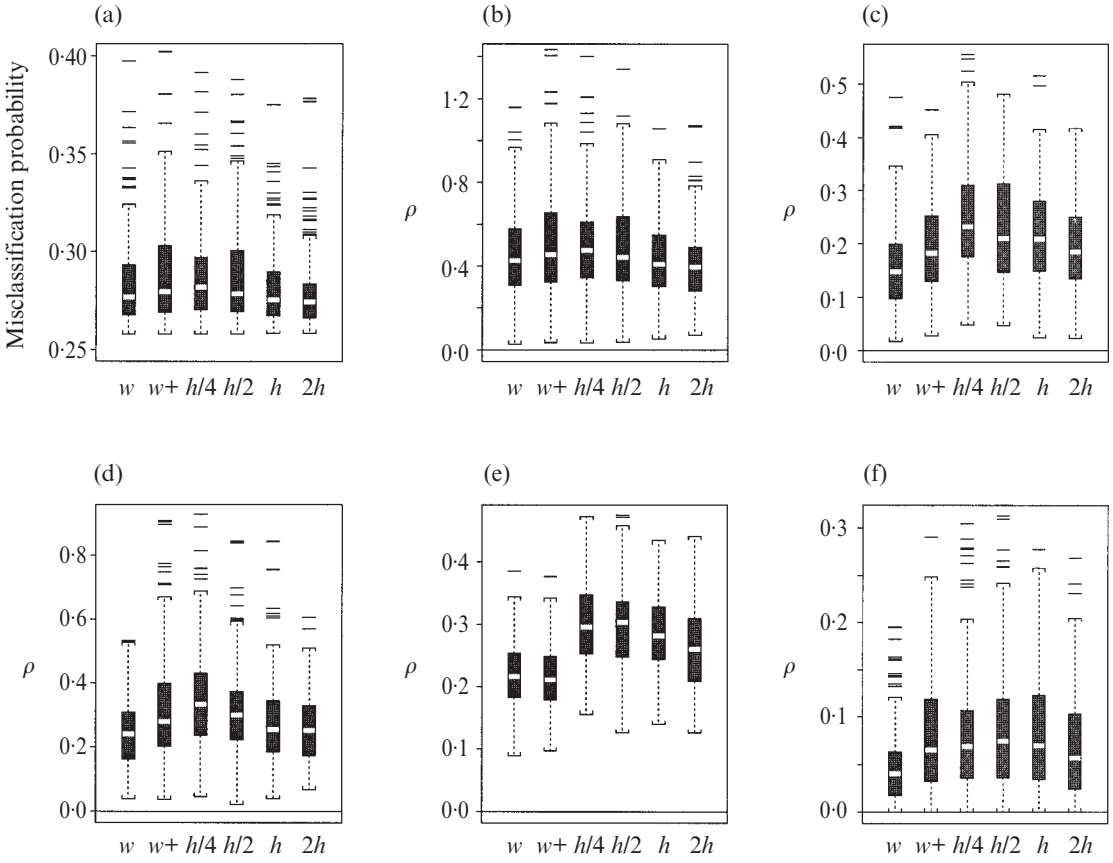

Fig. 2. Simulation study. Boxplots of the discriminant capability of the direction selected by different methods for projection pursuit discriminant analysis. Methods are wavelet-based for w and $w+$, and kernel-based for the rest. We used Sheather & Jones' (1991) rule for h and also bandwidth $h/4$, $h/2$ and $2h$. (a) shows global probability of misclassification for scenario in Fig. 1(a); (b)–(f) show angle ρ with true discriminant direction for scenarios in Figs 1(a)–(e).

Boxplots of the discrimination capability of the directions selected by each method are reported based on 200 simulations. The best method is the one with a boxplot as close as possible to the minimal probability of misclassification, which is 0·258 in this case.

Kernels and wavelets give roughly equivalent boxplots, the kernel based on the largest bandwidth being slightly better.

In this and later simulations, the setting is such that the projection index of a direction is an increasing function of the angle ρ between the selected direction and the discriminant direction. We can therefore report the angle ρ instead of the projection index Π. The values on the y-axis are thus more demonstrative: if the angle ρ is close to zero, this means that the method essentially found the right direction, and, if it is close to $\pi/2 \simeq 1{\cdot}57$, the method 'lost itself' finding a direction that is orthogonal to the discriminant direction. Figure 2(b) shows boxplots for ρ obtained from the same simulations as in Fig. 2(a). Clearly all methods find directions that are close to, but rarely very close to, the discriminant axis.

Figures 2(c)–(f) show boxplots for ρ from simulations based on the settings in Figs 1(b)–(e) and Table 1. In Fig. 2(c), one group has a bimodal density and represents two-thirds of the population. Even though both groups have smooth densities, the wavelet

estimator is slightly better than the kernel. Figure 2(d) has the same setting as in Fig. 2(c), except that the two groups have equal probabilities. Here both methods give similar results. Figure 2(e), in which the non-discriminant subspace is of dimension 9, shows that even in 10 dimensions the projection index for the wavelet is not too oscillatory and a direction close to the discriminant one can be found. This result is surprising to us. Figure 2(f) provides an example where the densities are not everywhere smooth. The three groups come from different Beta densities for the discriminant marginal densities, and the non-discriminant direction is uniform. In fact, the conditions for applying the kernel are not met, and, as one might expect, the wavelet estimate with the right amount of smoothing is less perturbed by the irregularities in the densities; the extent of the improvement is substantial. However, the second wavelet estimate is not better than the kernel estimates.

Our experience from these and other simulations is that, for kernel estimates, bandwidths larger than Sheather & Jones' rule give better results than do smaller ones. In this application, reduction of variance seems more important than reduction of bias. However, it is an open question to find a criterion for the bandwidth that takes into account the integral form of the projection index and the choice of the optimisation algorithm. It is also unclear whether or not higher-order kernels can improve results over regular ones.

It seems that the index based on wavelets is smooth enough for the optimiser to find a very good discriminant projection. Note however that all the thresholding rules in the literature, including the universal threshold used here, are defined in terms of regression, not density estimation. It is unclear whether or not rules adapted to density estimation and to the integral form of the index would improve the method further.

5. Example

We applied the methods to character recognition, using a dataset created and tested in Frey & Slate (1991). The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20 000 unique datapoints, approximately 800 per letter. Each image was converted into 16 primitive numerical attributes, namely sample moments in the marginal x and y directions, and edge counts (Frey & Slate, 1991). We only compare pairs of letters. Hastie & Tibshirani (1998) show that a good discriminant rule for each pair of groups is the core of an efficient rule that predicts the letter using a majority vote from all pairwise comparisons. We compare pairs of letters that are somewhat similar, i.e. such that the two sets of samples of the letters cannot be separated by a hyperplane.

For each pair, we compare three methods: projection pursuit discriminant analysis, based on wavelet and kernel density estimation and linear discriminant analysis. For the first two methods, all the settings of § 4·2 are kept, except that we allow 20 different starting values.

The results are shown in Fig. 3(a), (c), (e) for the pair (B, E) and in Fig. 3(b), (d), (f) for the pair (G, Q). Each plot shows a kernel estimate of the marginal density of the projection points for each letter, with a bandwidth purposefully smaller than the usual guideline value in order to reduce the bias and better capture the structure of the projected points. The vertical lines on the plots define the two assignment regions, a half-line for each group or a middle segment for one group and the complementary two half-lines for the other. This division minimises the training error amongst all possible divisions of up to three connected components. The corresponding training error, noted on the top of each plot,

gives a good estimate of the probability of misclassification, given the fairly large group sizes of about 800.

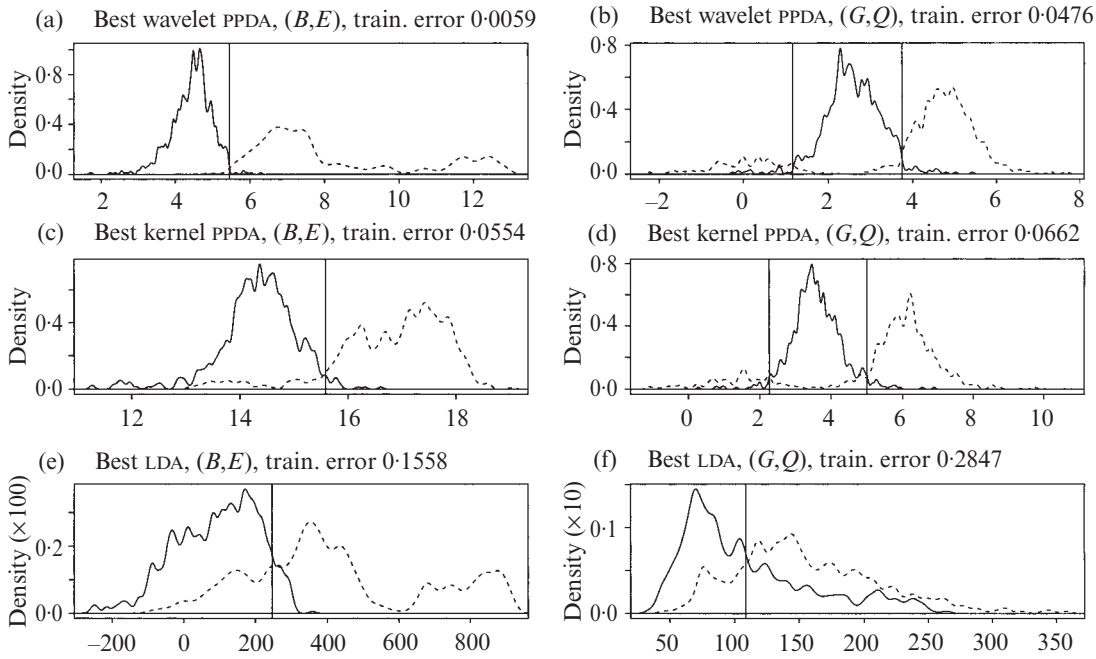

Fig. 3. Best projection according to wavelet, (a) and (b), and kernel, (c) and (d), projection pursuit discriminant analysis, PPDA, and linear discriminant analysis, LDA, given (e) and (f) in a 16-dimensional example of character recognition. (a), (c) and (e) show results of the methods for discrimination between about 800 instances of *B* and about 800 instances of *E*. (b), (d) and (f) show results as for (a), but for *G* and *Q*. Kernel density estimates for the projection points for each letter are shown and the assignment regions are separated by the vertical lines.

For the pair (*B*, *E*), the wavelet-based procedure selected a projection that almost separates the two groups. This projection misclassifies only 9 instances out of 1534, representing a training error of 0·0059. The separation point is sharp, and the local sensitivity of the wavelet was essential to find it. The kernel-based procedure selected a projection that is quite different; the angle between the two projections, in a 16-dimensional space, is 62°. The training error is almost ten times bigger than that of wavelet projection pursuit discriminant analysis. The third method suffers from the obvious nonnormality of the dataset and selects a very poor projection.

For the pair (*G*, *Q*), it seems that the best discriminant projection is such that the instances of the letter *Q* are separated into two groups, with the instances of *G* in between. In this context again, the wavelet-based projection pursuit discriminant analysis does a slightly better job, compared to the kernel-based method. The angle between the two projections, 23°, as well as the difference in the training errors are smaller than in the previous example, but the projection in the top panel is clearly better. Once again, the linear discriminant analysis is not able to capture the structure of the groups and selects a projection with a high training error.

6. MULTIPLE PROJECTION PURSUIT DISCRIMINANT ANALYSIS

The idea of projection pursuit discriminant analysis can be generalised by using more than one projection. Friedman et al. (1984) define a nonparametric dimension reduction

tool called projection pursuit density estimation. To estimate the multivariate density f based on a random sample of size N, the first step is to create a parametric estimate $\hat{f}_0(x)$, such as a Gaussian density with parameters equal to the sample mean and sample variance matrix. Then, the direction where changes are most beneficial is selected. At step k, one selects θ_k and estimates a function h_k such that $\hat{f}_k(x) = \hat{f}_{k-1}(x)h_k(\theta_k' x)$ fits the sample better than for any other choice of θ. The measure of goodness of fit is the Kullback–Leibler divergence $\int \log\{\hat{f}_k(x)/f(x)\} f(x)\,dx$. Suppose a superscript θ on a density denotes its marginal density in the direction θ. For a fixed θ, the function h_k that maximises the divergence is given by $h_k(x) = f^{\theta}(x)/\hat{f}_{k-1}^{\theta}(x)$. The Kullback–Leibler distance associated with this function h_k is then the projection index $\Pi(\theta)$ that has to be maximised. The optimisation is slightly more difficult than in other projection pursuit methods, as one must calculate the univariate $\hat{f}_{k-1}^{\theta}(x)$.

The density estimator so defined can, of course, be used for discriminant analysis. E. Elguero uses this procedure in his thesis and applies it to pollution data. He also proves the uniform convergence of the estimator. Polzehl (1995) exploits this construction and adapts the choice of the direction to optimise the ability of the density to discriminate between the groups. The projection index $\Pi(\theta)$ is the global probability of misclassification given in equation (2·1), where the regions $\mathscr{R}_g$ are given by

$$\mathscr{R}_g = \{x \in R^M \,|\, \hat{\pi}_g \hat{f}_{g,k-1}(x) h_{g,k}(\theta' x) > \hat{\pi}_i \hat{f}_{i,k-1}(x) h_{i,k}(\theta' x) \text{ for all } i \neq g\}.$$

Replacing $f(x)\,dx$ by $dF_n(x)$ in equation (2·1) gives a discontinuous estimate $\hat{\Pi}(\theta)$ as a function of θ. The discontinuities arise when observations pass from one region to another. Writing $\int_{\mathscr{R}_g} \pi_g f_g(x)\,dx$ as $\int I(x \in \mathscr{R}_g)\pi_g f_g(x)\,dx$, one may replace the indicator function I by a smoother function. This leads to a continuous estimated projection index.

Wavelets are serious competitors. First, the convergence results of § 3 carry over without much difficulty. In Theorem 3 we generalised the uniform convergence results of Elguero's thesis for univariate projection. It is not difficult to carry over to the wavelets approach his results concerning the uniform convergence of the multivariate estimator $\hat{f}_k(x)$. Moreover, the simulations of § 4·2 show that wavelets behave quite well in such complicated settings. When we only need one one-dimensional estimator at a time, the conclusions of § 4·2 apply.

Acknowledgement

Most of this work was undertaken as part of my Ph.D. dissertation at the Swiss Federal Institute of Technology in Lausanne under the supervision of Prof. S. Morgenthaler, to whom I am very grateful. I would also like to thank C. Posse and E. Restle for their help, and the editor, the associate editor and a referee whose comments led to a better presentation.

Appendix

Proof of Theorem 3

For this proof, it is convenient to view the linear wavelet estimate as an approximation kernel. The kernel associated with a wavelet basis with scaling function ϕ is given by $K(x, y) = \sum_k \phi(x-k)\phi(y-k)$, and we define $K_j(x,y) = 2^j K(2^j x, 2^j y)$. Under a weaker condition on ϕ than in Theorem 1, $\int K_j(x, y)\,dy = 1$. Moreover, for a chosen level j_0 for the linear wavelet estimator, we

have

$$\hat{f}(x) = \int K_{j_0}(x, y)\, dF_N(y), \quad E\{\hat{f}(x)\} = \int K_{j_0}(x, y)\, dF(y). \tag{A·1}$$

For a unified notation, the operator Ξ is defined either as the supremum operator $\Xi_x b(x) = \sup_x b(x)$ or as the integration operator $\Xi_x b(x) = \int_R b(x)\, dx$. For the same reason we denote by $h = 1/2^{j_0}$ the parameter of the wavelet density estimator. We have to show that

$$\sup_\theta \Xi_x |\hat{f}^\theta(x) - f^\theta(x)| \leqslant \sup_\theta \Xi_x |\hat{f}^\theta(x) - E\{\hat{f}^\theta(x)\}| + \sup_\theta \Xi_x |E\{\hat{f}^\theta(x)\} - f^\theta(x)| \tag{A·2}$$

converges to 0 almost surely, as $N \to \infty$, and we consider the two terms separately.

First consider the second term, which is deterministic. Elguero's thesis shows that the family $f^\theta(x)$ is equi-uniformly continuous, equi in θ and uniformly in x, which means that, for all $\varepsilon > 0$, there exists an $\eta > 0$ such that, for all $\theta \in \mathscr{S}^{M-1}$ and for all $x_1, x_2 \in R$ with $|x_1 - x_2| < \eta$, we have $|f^\theta(x_1) - f^\theta(x_2)| < \varepsilon$. For the wavelet case, by Lemma 8.6 in Härdle et al. (1998, p. 85), the bounding condition on ϕ implies that $|K_h(x, y)| \leqslant C_1 C_2 h^{-1} \Omega(C_2 |x - y|/h)$, where C_1 and C_2 are positive constants. Denote $\int \Omega(|y|)\, dy$ by C_3. The possibly higher-order convolution kernel estimator is also trivially bounded. By equation (A·1), we have that

$$|E\{\hat{f}^\theta(x)\} - f^\theta(x)| \leqslant \int C_1 C_2 h^{-1} \Omega(C_2 |x - y|/h) |f^\theta(y) - f^\theta(x)|\, dy. \tag{A·3}$$

Denote by $w^\theta(x, y)$ the above integrand, which is symmetric in its arguments. Note that it is zero when both x and y are outside $[-r; r]$, where r is the radius of the ball containing the support of f. Thus

$$\int_R |E\{\hat{f}^\theta(x)\} - f^\theta(x)|\, dx \leqslant 2 \int_{-r}^{r} \int_R w^\theta(x, y)\, dy\, dx. \tag{A·4}$$

We can now show that $\Xi_x |E\{\hat{f}^\theta(x)\} - f^\theta(x)|$ converges to 0 uniformly over θ. For fixed $\delta > 0$ we show that there exists N_δ, independent of θ and x, such that (A·3) is bounded by δ for any $N \geqslant N_\delta$. For this, set $\varepsilon = \delta / \{2 C_1 C_3 \max(4r, 1)\}$ and pick the corresponding η, which, because of equicontinuity, does not depend on θ or x. Then separate the domain of integration in the integral (A·3) into two parts, $[x - \eta, x + \eta]$ and $(-\infty, x - \eta) \cup (x + \eta, \infty)$, thereby defining integrals $I_1(x)$ and $I_2(x)$. The first one gives

$$I_1(x) \leqslant \int_{x-\eta}^{x+\eta} C_1 C_2 h^{-1} \Omega(C_2 |x - y|/h) \varepsilon\, dy \leqslant C_1 C_3 \varepsilon.$$

This shows that $\sup_\theta \sup_x I_1(x) \leqslant \delta/2$. For the integral operator,

$$\sup_\theta \int_R I_1(x)\, dx \leqslant \sup_\theta 2 \int_{-r}^{r} I_1(x)\, dx \leqslant \delta/2,$$

by equation (A·4) and the choice of ε. The second integral $I_2(x)$ is

$$\begin{aligned} I_2(x) &= \left(\int_{-\infty}^{x-\eta} + \int_{x+\eta}^{\infty} \right) C_1 C_2 h^{-1} \Omega(C_2 |x - y|/h) |f^\theta(y) - f^\theta(x)|\, dy \\ &\leqslant 2 \max_\theta \max_v f^\theta(v) \left(\int_{-\infty}^{-\eta/h} + \int_{\eta/h}^{\infty} \right) C_1 C_2 \Omega(C_2 |z|)\, dz. \end{aligned}$$

Since this expression does not depend on x or θ, it is a bound for $\sup_\theta \sup_x I_2(x)$ and, by equation (A·4), the same expression multiplied by $4r$ is a bound for $\sup_\theta \int I_2(x)\, dx$. In addition, as Ω is in L_1, $I_2(x)$ tends to zero for decreasing values of h. There exists therefore an h, and thus an N_δ, such

that $I_2(x)$ is smaller than $\delta/2$ for any $N \geqslant N_\delta$. For the same values of N, the sum of the two integrals $I_1(x)$ and $I_2(x)$ is bounded by δ, which proves the claimed convergence.

We now turn to $\sup_\theta \Xi_x |\hat{f}^\theta(x) - E\{\hat{f}^\theta(x)\}|$, the first term in (A·2). This term is stochastic and we prove almost sure convergence. Denote by $F^\theta(y)$ and $F_N^\theta(y)$ the true and the empirical distributions on the margin defined by θ. We will show that, for both kernels and wavelets, and for both operators, we can find a constant C_4 such that

$$\sup_\theta \Xi_x |\hat{f}^\theta(x) - E\{\hat{f}^\theta(x)\}| \leqslant C_4 h^{-1} \sup_\theta \sup_x |F^\theta(x) - F_N^\theta(x)|. \tag{A·5}$$

From this, we can show that

$$\begin{aligned} \mathrm{pr}\left[\sup_\theta \Xi_x |\hat{f}^\theta(x) - E\{\hat{f}^\theta(x)\}| \geqslant \varepsilon\right] &\leqslant \mathrm{pr}\left\{\sup_\theta \sup_x |F^\theta(x) - F_N^\theta(x)| \geqslant \frac{\varepsilon}{C_4 h^{-1}}\right\} \\ &\leqslant C_5 \exp\left\{C_6 M \log(N/M) - \frac{\varepsilon^2}{8C_4^2} h^2 N\right\}, \end{aligned}$$

where the second inequality follows from Proposition 5.1 of Diaconis & Freedman (1984). Here, M is the dimension, and C_5 and C_6 are positive constants. By the Borel–Cantelli lemma, one has almost sure convergence if the series with its general term given by the last expression converges for any $\varepsilon > 0$. This happens if $h^2 N = O(\log^{1+\gamma} N)$ for some $\gamma > 0$, which is equivalent to the given conditions for 2^{j_0} and h.

We now prove equation (A·5), first for the kernel estimator and then for the wavelet estimator. For the kernel case, as K is continuous, of bounded variation and tends to zero for $x \to \pm\infty$, it can be written as the distribution of a finite signed measure μ: $K(x) = \int I(u < x)\, d\mu(u)$. We then have

$$\begin{aligned} |\hat{f}^\theta(x) - E\{\hat{f}^\theta(x)\}| &= \left|\frac{1}{h} \iint I\{u < (x-y)/h\}\, d\mu(u)\{dF_N^\theta(y) - dF^\theta(y)\}\right| \\ &= \left|\frac{1}{h} \int \{F^\theta(x - hu) - F_N^\theta(x - hu)\}\, d\mu(u)\right| \leqslant \frac{1}{h} \int |F^\theta(x - hu) - F_N^\theta(x - hu)|\, d|\mu|(u). \end{aligned}$$

For the supremum operator, this implies that

$$\sup_\theta \sup_x |\hat{f}^\theta(x) - E\{\hat{f}^\theta(x)\}| \leqslant \frac{\|\mu\|}{h} \sup_\theta \sup_x |F^\theta(x) - F_N^\theta(x)|,$$

where $\|\mu\| = |\mu|(R)$ is the total variation of μ. This proves equation (A·5) with $C_4 = \|\mu\|$. For the integration operator, one writes

$$\int_R |\hat{f}^\theta(x) - E\{\hat{f}^\theta(x)\}|\, dx \leqslant \frac{1}{h} \int_R |F^\theta(z) - F_N^\theta(z)|\, dz \int_R d|\mu|(u).$$

Since the true and empirical distributions coincide outside the ball of radius r, the domain of the first integral can be restricted to the interval $[-r, r]$. This shows that (A·5) holds in this case also with $C_4 = 2r\|\mu\|$.

We now establish (A·5) in the wavelet case. As ϕ is of bounded variation and tends to zero for $x \to \pm\infty$, it can be written as the distribution of a bounded signed measure μ: $\phi(x) = \int I(u < x)\, d\mu(u)$. We thus have

$$|\hat{f}^{\theta}(x) - E\{\hat{f}^{\theta}(x)\}| = \left| \int \sum_{k \in Z} \phi_{j_0 k}(x)\phi_{j_0 k}(y)\{dF_N^{\theta}(y) - dF^{\theta}(y)\} \right|$$

$$= \left| \sum_{k \in Z} \phi_{j_0 k}(x) \iint 2^{j_0/2} I(u < 2^{j_0} y - k)\, d\mu(u)\{dF_N^{\theta}(y) - dF^{\theta}(y)\} \right|$$

$$\leqslant \sum_{k \in Z} |\phi_{j_0 k}(x)| \int 2^{j_0/2} |F^{\theta}\{(u+k)/2^{j_0}\} - F_N^{\theta}\{(u+k)/2^{j_0}\}|\, d|\mu|(u), \qquad (A{\cdot}6)$$

where the inversion of the integral and the sum is allowed because of the bounding condition on ϕ; see Proposition 8.4 in Härdle et al. (1998, p. 83). To bound the supremum operator, it suffices to note that

$$|\hat{f}^{\theta}(x) - E\{\hat{f}^{\theta}(x)\}| \leqslant \sum_{k \in Z} |\phi_{j_0 k}(x)| 2^{j_0/2} \|\mu\| \sup_v |F^{\theta}(v) - F_N^{\theta}(v)|.$$

By the condition on ϕ, $\sup_x \sum_k |\phi(x-k)| < C_7 < \infty$; see Proposition 8.5 in Härdle et al. (1998, p. 83). This allows us to establish equation (A·5) with $C_4 = C_7 \|\mu\|$, since $2^{j_0} = h^{-1}$. The proof for the integration operator is slightly trickier. Since the density vanishes outside the ball of radius r, the domain of the integral of (A·6) can be restricted to the interval $[-2^{j_0} r - k, 2^{j_0} r - k]$. We then have

$$\sup_{\theta} \int_R |\hat{f}^{\theta}(x) - E\{\hat{f}^{\theta}(x)\}|\, dx \leqslant 2^{j_0/2} \sum_{k \in Z} \int_R |\phi_{j_0 k}(x)|\, dx \int_{-2^{j_0} r - k}^{2^{j_0} r - k} d|\mu|(u) \sup_{\theta} \sup_v |F^{\theta}(v) - F_N^{\theta}(v)|$$

$$\leqslant C_3 \sum_{k \in Z} \int_{-2^{j_0}\lceil r \rceil - k}^{2^{j_0}\lceil r \rceil - k} d|\mu|(u) \sup_{\theta} \sup_v |F^{\theta}(v) - F_N^{\theta}(v)|$$

$$= 2^{j_0} \lceil r \rceil \|\mu\| \sup_{\theta} \sup_v |F^{\theta}(v) - F_N^{\theta}(v)|,$$

where the second inequality is a consequence of $\int |\phi(x)|\, dx \leqslant \int \Omega(|x|)\, dx = C_3$. This establishes (A·5) for the last case, since $2^{j_0} = h^{-1}$. This ends the proof of equations (3·1) and (3·2).

The last part of the theorem, equation (3·3), is a direct consequence of the uniform convergence of the integral in (3·2), since, from the proof of Theorem 2, we have for any partition $\mathscr{R}$ and any direction θ that

$$|\hat{\Pi}(\mathscr{R}, \theta) - \Pi(\mathscr{R}, \theta)| \leqslant \sum_{g=1}^{G} |\hat{\pi}_g - \pi_g| + \sum_{g=1}^{G} \sup_{\theta} \int_R |\hat{f}_g^{\theta}(x) - f_g^{\theta}(x)|\, dx.$$

As the bound does not depend on $\mathscr{R}$ or on θ, it also bounds $\sup^{\#}$.

REFERENCES

COIFMAN, R. R. & SAITO, N. (1994). Constructions of local orthonormal bases for classification and regression. *Comptes Rendus* **319**, 191–6.

DIACONIS, P. & FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815.

FREY, P. W. & SLATE, D. J. (1991). Letter recognition using holland-style adaptive classifiers. *Mach. Learn.* **6**, 161–82.

FRIEDMAN, J. H. & STUETZLE, W. (1981). Projection pursuit regression. *J. Am. Statist. Assoc.* **76**, 817–23.

FRIEDMAN, J. H. & TUKEY, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comp.* **23**, 881–90.

FRIEDMAN, J. H., STUETZLE, W. & SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Am. Statist. Assoc.* **79**, 599–608.

HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. & TSYBAKOV, A. (1998). *Wavelets, Approximation, and Statistical Applications*, Lecture Notes in Statistics, **129**. Berlin: Springer-Verlag.

HASTIE, T. & TIBSHIRANI, R. (1998). Classification by pairwise coupling. *Ann. Statist.* **26**, 451–71.

JONES, M. C. & SIBSON, R. (1987). What is projection pursuit? (with Discussion). *J. R. Statist. Soc.* A **150**, 1–36.

Polzehl, J. (1995). Projection pursuit discriminant analysis. *Comp. Statist. Data Anal.* **20**, 141–57.
Posse, C. (1992). Projection pursuit discriminant analysis for two groups. *Commun. Statist.* A **21**, 1–19.
Posse, C. (1995). Tools for two-dimensional exploratory projection pursuits. *J. Comp. Graph. Statist.* **4**, 83–100.
Sheather, S. J. & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc.* B **53**, 683–90.
Vetterli, M. & Kovačević, J. (1995). *Wavelets and Subband Coding*. Upper Saddle River, NJ: Prentice Hall.
Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. New York: Wiley.

[*Received June* 2000. *Revised March* 2001]