Multi-scale Classification using Localized Spatial Depth

Subhajit Dutta[a] and Anil K. Ghosh[b]

[a]Department of Mathematics and Statistics, Indian Institute of Technology, Kanpur 208016, U.P., India.

[b]Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India.

Email: tijahbus@gmail.com[a], akghosh@isical.ac.in[b]

Abstract

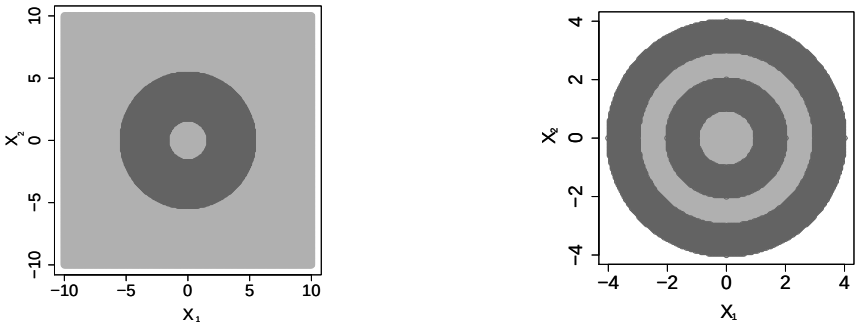
In this article, we develop and investigate a new classifier based on features extracted using spatial depth. Our construction is based on fitting a generalized additive model to the posterior probabilities of the different competing classes. To cope with possible multi-modal as well as non-elliptic population distributions, we develop a localized version of spatial depth and use that with varying degrees of localization to build the classifier. Final classification is done by aggregating several posterior probability estimates each of which is obtained using localized spatial depth with a fixed scale of localization. The proposed classifier can be conveniently used even when the dimension is larger than the sample size, and its good discriminatory power for such data has been established using theoretical as well as numerical results.

Keywords : Bayes classifier, elliptic and non-elliptic distributions, HDLSS asymptotics, uniform strong consistency, weighted aggregation of posteriors.

1 Introduction

In a classification problem with J classes, we usually have n_j labeled observations $\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jn_j}$ from the j-th class ($1 \leq j \leq J$), and we use these $n = \sum_{j=1}^{J} n_j$ observations to construct a decision rule for classifying a new unlabeled observation $\mathbf{x}$ to one of these J pre-defined classes. If π_j and f_j respectively denote the prior probability and the probability density function of the j-th class, and $p(j|\mathbf{x})$ denotes the corresponding posterior probability, the optimal *Bayes classifier* assigns $\mathbf{x}$ to the class j^*, where $j^* = \arg\max_{1 \leq j \leq J} p(j|\mathbf{x}) = \arg\max_{1 \leq j \leq J} \pi_j f_j(\mathbf{x})$. However, the $f_j(\mathbf{x})$'s (or, the $p(j|\mathbf{x})$'s) are unknown in practice, and one needs to estimate them from the training sample of labeled observations. Popular parametric classifiers like linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are motivated by parametric model assumptions on the underlying class distributions. So, they may lead to poor classification when the model assumptions fail to hold, and the class boundaries of the Bayes classifier have complex geometry. On the other hand, nonparametric classifiers like those based on k-nearest neighbors (k-NN) and kernel density estimates (KDE) are more flexible and free from such model assumptions. But, they suffer from the curse of dimensionality and are often not suitable for high-dimensional data.

Consider two examples denoted by **E1** and **E2**, respectively. **E1** involves a classification problem with two classes in $\mathbb{R}^d$, where the distribution of the first class is an equal mixture of $\mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $\mathrm{N}_d(\mathbf{0}_d, 10\mathbf{I}_d)$, and that for the second class is $\mathrm{N}_d(\mathbf{0}_d, 5\mathbf{I}_d)$. Here N_d denotes the d-variate normal distribution, $\mathbf{0}_d = (0, \ldots, 0)^T \in \mathbb{R}^d$ and $\mathbf{I}_d$ is the $d \times d$ identity matrix. In **E2**, each class distribution is an equal mixture of two uniform distributions. The distribution for the first (respectively, the second) class is a mixture of $\mathrm{U}_d(0, 1)$ and $\mathrm{U}_d(2, 3)$ (respectively, $\mathrm{U}_d(1, 2)$ and $\mathrm{U}_d(3, 4)$). Here $\mathrm{U}_d(r_1, r_2)$ denotes the uniform distribution over the region $\{\mathbf{x} \in \mathbb{R}^d : r_1 \leq \|\mathbf{x}\| \leq r_2\}$ with $0 \leq r_1 < r_2$. Figure 1 shows the class boundaries of the Bayes classifier for these two examples when $d = 2$, and $\pi_1 = \pi_2 = 1/2$. The regions colored grey (respectively, black) correspond to observations classified to the first (respectively, the

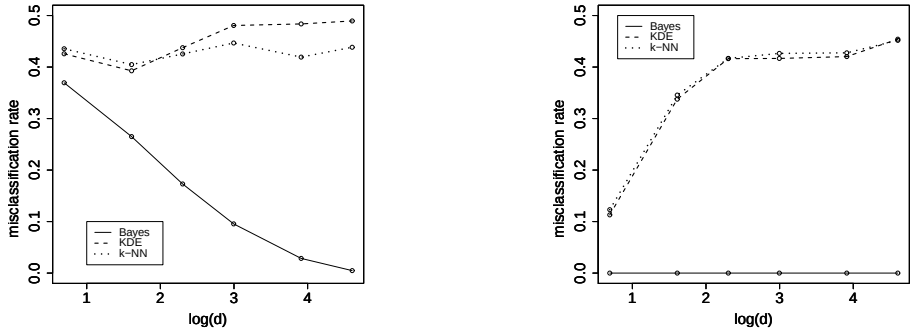


(a) Example **E1** (b) Example **E2**

Figure 1: `Bayes class boundaries in` $\mathbb{R}^2$.

second) class by the Bayes classifier. It is clear that classifiers like LDA and QDA or any other classifier with linear or quadratic class boundaries will deviate significantly from the Bayes classifier in both examples. A natural question then is how standard nonparametric classifiers like those based on k-NN and KDE perform in such examples. In Figure 2, we have plotted average misclassification rates of these two classifiers along with the Bayes risks for different values of d. These classifiers were trained on a sample of size 100 generated from each class distribution, and the misclassification rates were computed based on a sample of size 250 from each class. This procedure was repeated 500 times to calculate the average misclassification rate. Smoothing parameters associated with k-NN and KDE (i.e., the k in k-NN and the bandwidth in KDE) were chosen by minimizing leave-one-out cross-validation estimates of misclassification rates [17]. Figure 2 shows that in **E1**, the Bayes risk decreases to zero as d grows. Since the class distributions in **E2** have disjoint supports, the Bayes risk is zero irrespective of the value of d. But in both examples, the misclassification rates of these two nonparametric classifiers increased to almost 50% as d increased.

These two examples clearly show the necessity to develop new classifiers to cope with such situations. Over the last three decades, data depth (see, e.g., [29, 42]) has emerged as a powerful tool for data analysis with applications in many areas including supervised and unsupervised classification (see [20, 11, 12, 18, 39, 7, 25, 23, 33]). Spatial depth (also known as the L_1 depth) is a popular notion of data depth that was introduced and studied



(a) Example **E1** (b) Example **E2**

Figure 2: `Misclassification rates of nonparametric classifiers and the Bayes classifier for` $d = 2, 5, 10, 20, 50$ `and` 100.

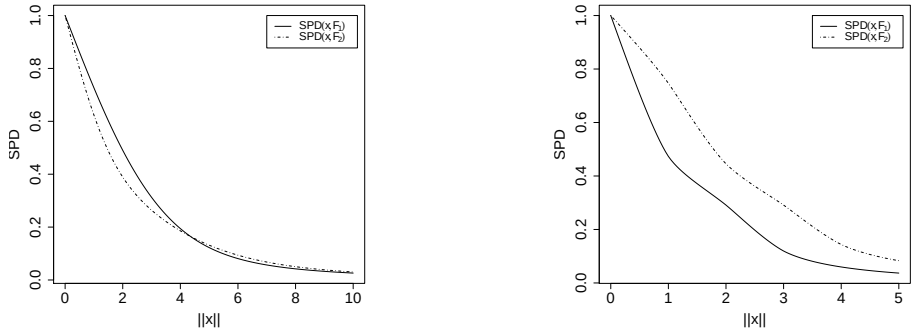
in [38] and [37]. The *spatial depth* (SPD) of an observation $\mathbf{x} \in \mathbb{R}^d$ w.r.t. a distribution function F on $\mathbb{R}^d$ is defined as $\text{SPD}(\mathbf{x}, F) = 1 - \big\|E_F\{u((\mathbf{x} - \mathbf{X}))\}\big\|$, where $\mathbf{X} \sim F$ and $u(\cdot)$ is the multivariate sign function given by $u(\mathbf{x}) = \|\mathbf{x}\|^{-1}\mathbf{x}$ if $\mathbf{x} \neq \mathbf{0}_d \in \mathbb{R}^d$, and $u(\mathbf{0}_d) = \mathbf{0}_d$. Henceforth, $\|\cdot\|$ will denote the Euclidean norm. Spatial depth is often computed on the standardized version of the data. In that case, SPD is defined as

$$\text{SPD}(\mathbf{x}, F) = 1 - \big\|E_F\{u(\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mathbf{X}))\}\big\|,$$

where $\mathbf{\Sigma}$ is a scatter matrix associated with F. If $\mathbf{\Sigma}$ has the affine equivariance property, this version of SPD is affine invariant.

Like other depth functions, SPD provides a centre-outward ordering of multivariate data. An observation has higher (respectively, lower) depth if it lies close to (respectively, away from) the centre of the distribution. In other words, given an observation $\mathbf{x}$ and a pair of probability distributions F_1 and F_2, if $\text{SPD}(\mathbf{x}, F_1)$ is larger than $\text{SPD}(\mathbf{x}, F_2)$, one would expect $\mathbf{x}$ to come from F_1 instead of F_2. Based on this simple idea, the *maximum depth classifier* was developed in [12, 20]. For a J-class problem involving distributions $F_1, \ldots, F_J$, it classifies an observation $\mathbf{x}$ to the j^*-th class, where $j^* = \arg\max_{1 \leq j \leq J} \text{SPD}(\mathbf{x}, F_j)$.

An important property of SPD (see Lemma 1 in Appendix) is that when the class distribution F is unimodal and spherically symmetric, the class density function turns out to be



(a) Example **E1** (b) Example **E2**

Figure 3: $\texttt{SPD}(\mathbf{x}, F_1)$ $\texttt{and}$ $\texttt{SPD}(\mathbf{x}, F_2)$ $\texttt{for different values of}$ $\|\mathbf{x}\|$ $\texttt{when}$ $\mathbf{x} \in \mathbb{R}^2$.

a monotonically increasing function of SPD. In both examples **E1** and **E2**, the class distributions are spherical. Consequently, $\text{SPD}(\mathbf{x}, F)$ is a function of $\|\mathbf{x}\|$ in view of the rotational invariance of $\text{SPD}(\mathbf{x}, F)$. In Figure 3, we have plotted $\text{SPD}(\mathbf{x}, F_1)$ and $\text{SPD}(\mathbf{x}, F_2)$ for different values of $\|\mathbf{x}\|$ for examples **E1** and **E2**, where F_1 and F_2 are the two class distributions and $\mathbf{x} \in \mathbb{R}^2$. It is transparent from the plots that the maximum depth classifier based on SPD will fail in both examples. In example **E1**, for all values of $\|\mathbf{x}\|$ smaller (respectively, greater) than a constant close to 4, the observations will be classified to the first (respectively, the second) class by the maximum SPD classifier. On the other hand, this classifier will classify all observations to the second class in example **E2**.

In Section 2, we develop a modified classifier based on SPD to overcome this limitation of the maximum depth classifier. Most of the existing modified depth based classifiers are developed mainly for two class problems (see, e.g., [12, 7, 25, 33, 23]). For classification problems involving $J(> 2)$ classes, one usually solves $\binom{J}{2}$ binary classification problems taking one pair of classes at a time and then uses majority votes to make the final classification. Our proposed classification method based on SPD addresses the J class problem directly.

Almost all depth based classifiers proposed in the literature require ellipticity of class distributions to achieve Bayes optimality. In order to cope with possible multimodal as well as non-elliptic population distributions, we construct a localized version of SPD (henceforth

referred to as LSPD) in Section 3. In Section 4, we develop a multiscale classifier based on LSPD. Relevant theoretical results on SPD, LSPD and the resulting classifiers have also been studied in these sections.

An advantage of SPD over other depth functions is its computational simplicity. Classifiers based on SPD and LSPD can be constructed even when the dimension of the data exceeds the sample size. We deal with such high-dimensional low sample size cases in Section 5, and show that both classifiers turn out to be optimal under a fairly general framework. In Sections 6 and 7, some simulated and benchmark data sets are analyzed to establish the usefulness of our classification methods. Section 8 contains a brief summary of the work and some concluding remarks. All proofs and mathematical details are given in the Appendix.

2 Bayes optimality of a classifier based on SPD

Let us assume that $f_1, \ldots, f_J$ are the density functions of J elliptically symmetric distributions on $\mathbb{R}^d$, where $f_j(\mathbf{x}) = |\mathbf{\Sigma}_j|^{-1/2} g_j(\|\mathbf{\Sigma}_j^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_j)\|)$ for $1 \leq j \leq J$. Here $\boldsymbol{\mu}_j \in \mathbb{R}^d$, $\mathbf{\Sigma}_j$ is a $d \times d$ positive definite matrix, and $g_j(\|\mathbf{t}\|)$ is a probability density function on $\mathbb{R}^d$ for $1 \leq j \leq J$. For such classification problems involving general elliptic populations with equal or unequal priors, the next theorem establishes the Bayes optimality of a classifier, which is based on $\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), \ldots, z_J(\mathbf{x}))^T = (\mathrm{SPD}(\mathbf{x}, F_1), \ldots, \mathrm{SPD}(\mathbf{x}, F_J))^T$, the vector of SPD. In particular, it follows from this theorem that for examples **E1** and **E2** discussed at the beginning of Section 1, the class boundaries (see Figure 1) of the Bayes classifiers are functions of $\mathbf{z}(\mathbf{x}) = (\mathrm{SPD}(\mathbf{x}, F_1), \mathrm{SPD}(\mathbf{x}, F_2))^T$.

Theorem 1 *If the densities of the J competing classes are elliptically symmetric, the posterior probabilities of these classes satisfy the logistic regression model given by*

$$p(j|\mathbf{x}) = p(j|\mathbf{z}(\mathbf{x})) = \frac{\exp(\Phi_j(\mathbf{z}(\mathbf{x})))}{[1 + \sum_{k=1}^{(J-1)} \exp(\Phi_k(\mathbf{z}(\mathbf{x})))]} \text{ for } 1 \leq j \leq (J-1) \tag{1}$$

$$\text{and } p(J|\mathbf{x}) = p(J|\mathbf{z}(\mathbf{x})) = \frac{1}{[1 + \sum_{k=1}^{(J-1)} \exp(\Phi_k(\mathbf{z}(\mathbf{x})))]}. \tag{2}$$

Here $\Phi_j(\mathbf{z}(\mathbf{x})) = \varphi_{j1}(z_1(\mathbf{x})) + \ldots + \varphi_{jJ}(z_J(\mathbf{x}))$*, and* φ_{ji}*s are appropriate real-valued functions of real variables. Consequently, the Bayes rule assigns an observation* $\mathbf{x}$ *to the class* j^**, where* $j^* = arg\ max_{1 \le j \le J}\ p(j|\mathbf{z}(\mathbf{x}))$.

Theorem 1 shows that the Bayes classifier is based on a nonparametric multinomial additive logistic regression model for the posterior probabilities, which is a special case of generalized additive models (GAM) [16]. If the prior probabilities of the J classes are equal, and $f_1, \ldots, f_J$ are all elliptic and unimodal differing only in their locations, this Bayes classifier reduces to the maximum SPD classifier [12, 20] (see Remark 1 after the proof of Theorem 1 in the Appendix).

For any fixed i and j, one can calculate the J-dimensional vector $\mathbf{z}(\mathbf{x}_{ji})$, where $\mathbf{x}_{ji}$ is the i-th training sample observation in the j-th class for $1 \le i \le n_j$ and $1 \le j \le J$. These $\mathbf{z}(\mathbf{x}_{ji})$s can be viewed as realizations of the vector of co-variates in a nonparametric multinomial additive logistic regression model, where the response corresponds to the class label that belongs to $\{1, \ldots, J\}$. So, a classifier based on SPD can be constructed by fitting a generalized additive model with the logistic link function. In practice, when we compute SPD of $\mathbf{x}$ from the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ generated from F, we use its empirical version as $\mathrm{SPD}(\mathbf{x}, F_n) = 1 - \left\| \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{x} - \mathbf{x}_i) \right\|$. For the standardized version of the data, it is defined as

$$\mathrm{SPD}(\mathbf{x}, F_n) = 1 - \left\| \frac{1}{n} \sum_{i=1}^{n} u(\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x} - \mathbf{x}_i)) \right\|,$$

where $\widehat{\mathbf{\Sigma}}$ is an estimate of $\mathbf{\Sigma}$, and F_n is the empirical distribution of the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The resulting classifier worked well in examples **E1** and **E2**, and we shall see it in Section 6.

3 Extraction of small scale distributional features by localization of spatial depth

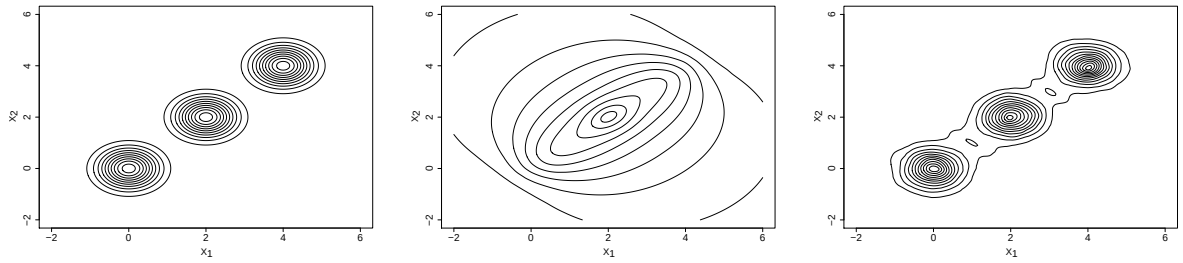
Under elliptic symmetry, the density function of a class can be expressed as a function of SPD, and hence the SPD contours coincide with the density contours. This is the main mathemat-

ical argument used in the proof of Theorem 1. Now, for certain non-elliptic distributions, where the density function cannot be expressed as a function of SPD, such mathematical arguments are no longer valid. For instance, consider an equal mixture of $\mathrm{N}_d(\mathbf{0}_d, 0.25\mathbf{I}_d)$, $\mathrm{N}_d(2\mathbf{1}_d, 0.25\mathbf{I}_d)$ and $\mathrm{N}_d(4\mathbf{1}_d, 0.25\mathbf{I}_d)$, where $\mathbf{1}_d = (1, \ldots, 1)^T$. We have plotted its SPD contours in Figure 4 when $d = 2$. For this trimodal distribution, the SPD contours fail to match the density contours. As a second example, we consider a d-dimensional distribution with independent components, where the i-th component is exponential with the scale parameter $d/(d - i + 1)$ for $1 \leq i \leq d$. We have plotted its SPD contours in Figure 5 when $d = 2$. Even in this example, the SPD contours differ significantly from the density contours. To cope with this issue, we suggest a *localization* of SPD (see the third contour plots (c) in Figures 4 and 5). As we shall see later, this localized SPD relates to the underlying density function, and the resulting classifier turns out to be the Bayes classifier (in a limiting sense) in a general nonparametric setup with arbitrary class densities.

Note that $\mathrm{SPD}(\mathbf{x}, F) = 1 - \|E_F\{u(\mathbf{x} - \mathbf{X})\}\|$ is constructed by assigning the same weight to each unit vector $u(\mathbf{x} - \mathbf{X})$ ignoring the significance of distance between $\mathbf{x}$ and $\mathbf{X}$. By introducing a weight function, which depends on this distance, one can extract important features related to the local geometry of the data. To capture these local features, we introduce a kernel function $K(\cdot)$ as a weight and define

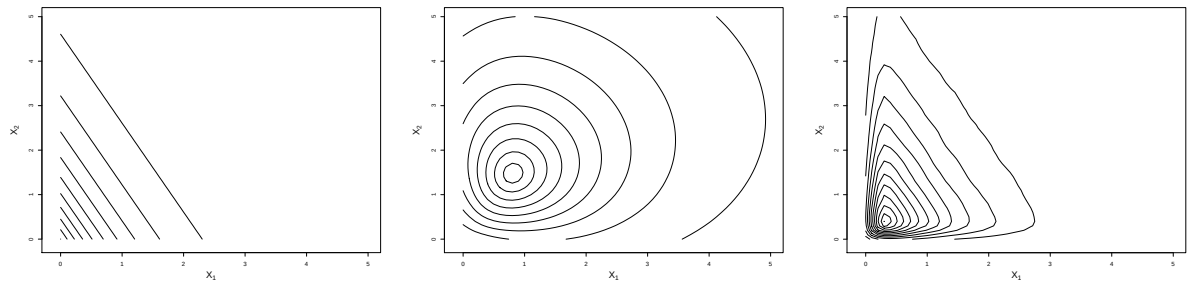
$$\Gamma_h(\mathbf{x}, F) = E_F[K_h(\mathbf{t})] - \|E_F[K_h(\mathbf{t})u(\mathbf{t})]\|,$$

where $\mathbf{t} = (\mathbf{x} - \mathbf{X})$ and $K_h(\mathbf{t}) = h^{-d}K(\mathbf{t}/h)$. Here K is chosen to be a bounded continuous density function on $\mathbb{R}^d$ such that $K(\mathbf{t})$ is a decreasing function of $\|\mathbf{t}\|$ and $K(\mathbf{t}) \to 0$ as $\|\mathbf{t}\| \to \infty$. The Gaussian kernel $K(\mathbf{t}) = (\sqrt{2\pi})^{-d} \exp\{-\|\mathbf{t}\|^2/2\}$ is a possible choice. It is desirable that the localized version of SPD approximates the class density or a monotone function of it for small values of h. This will ensure that the class densities and hence, the class posterior probabilities become functions of the local depth as $h \to 0$. On the other hand, one should expect that as $h \to \infty$, the localized version of SPD should tend to SPD



(a) Density (b) SPD (c) LSPD$_{h=.4}$

Figure 4: `Contours of density, SPD and LSPD`$_h$ `(with` $h = .4$`) functions for a symmetric, trimodal density function.`



(a) Density (b) SPD (c) LSPD$_{h=.25}$

Figure 5: `Contours of density, SPD and LSPD`$_h$ `(with` $h = .25$`) functions for the density function` $f(x_1, x_2) = .5 \exp\{-(x_1 + .5x_2)\} I\{x_1 > 0, x_2 > 0\}$.

or a monotone function of it. However, $\Gamma_h(\mathbf{x}, F) \to 0$ as $h \to \infty$. So, we re-scale $\Gamma_h(\mathbf{x}, F)$ by an appropriate factor to define the *localized spatial depth* (LSPD) function as follows:

$$\text{LSPD}_h(\mathbf{x}, F) = \begin{cases} \Gamma_h(\mathbf{x}, F) & \text{if } h \leq 1, \\ h^d \Gamma_h(\mathbf{x}, F) & \text{if } h > 1. \end{cases} \tag{3}$$

Using $\mathbf{t} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mathbf{X})$ in the definition of $\Gamma_h(\mathbf{x}, F)$, one gets LSPD on standardized data, which is affine invariant if $\mathbf{\Sigma}$ is affine equivariant. LSPD$_h$ defined this way is a continuous function of h, and $\mathbf{z}_h(\mathbf{x}) = (\text{LSPD}_h(\mathbf{x}, F_1), \ldots, \text{LSPD}_h(\mathbf{x}, F_J))^T$ has the desired behavior as shown in Theorem 2.

Theorem 2 *Consider a kernel function $K(\mathbf{t})$ that satisfy $\int_{\mathbb{R}^d} \|\mathbf{t}\| K(\mathbf{t}) d\mathbf{t} < \infty$. If $f_1, \ldots, f_J$ are continuous density functions with bounded first derivatives, and the scatter matrix $\mathbf{\Sigma}_j$ corresponding to $f_j(\mathbf{x})$ exists for all $1 \leq j \leq J$, then*

(a) $\mathbf{z}_h(\mathbf{x}) \to (|\mathbf{\Sigma}_1|^{1/2} f_1(\mathbf{x}), \ldots, |\mathbf{\Sigma}_J|^{1/2} f_J(\mathbf{x}))^T$ *as* $h \to 0$, *and*

(b) $\mathbf{z}_h(\mathbf{x}) \to (K(\mathbf{0}) SPD(\mathbf{x}, F_1), \ldots, K(\mathbf{0}) SPD(\mathbf{x}, F_J))^T$ *as* $h \to \infty$.

Now, we construct a classifier by plugging in LSPD_h instead of SPD in the GAM discussed in Section 2. So, we consider the following model for the posterior probabilities

$$p(j|\mathbf{z}_h(\mathbf{x})) = \frac{\exp(\Phi_j(\mathbf{z}_h(\mathbf{x})))}{[1 + \sum_{k=1}^{(J-1)} \exp(\Phi_k(\mathbf{z}_h(\mathbf{x})))]}, \text{ for } 1 \leq j < (J-1), \tag{4}$$

$$\text{and } \; p(J|\mathbf{z}_h(\mathbf{x})) = \frac{1}{[1 + \sum_{k=1}^{(J-1)} \exp(\Phi_k(\mathbf{z}_h(\mathbf{x})))]}. \tag{5}$$

The main implication of part (a) of Theorem 2 is that the classifier constructed using GAM and $\mathbf{z}_h(\mathbf{x})$ as the covariate tends to the Bayes classifier in a general nonparametric setup as $h \to 0$. On the other hand, part (b) of Theorem 2 implies that for elliptic class distributions, the same classifier tends to the Bayes classifier when $h \to \infty$. When we fit GAM, the functions Φ_js are estimated nonparametrically. Flexibility of such nonparametric estimates automatically takes care of the constants $|\mathbf{\Sigma}_j|^{1/2}$ for $1 \leq j \leq J$ and $K(\mathbf{0})$ in the expressions of the limiting values of $\mathbf{z}_h(\mathbf{x})$ in parts (a) and (b) of Theorem 2, respectively.

The empirical version of $\Gamma_h(\mathbf{x}, F)$, denoted by $\Gamma_h(\mathbf{x}, F_n)$, is defined as

$$\Gamma_h(\mathbf{x}, F_n) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{t}_i) - \left\| \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{t}_i) u(\mathbf{t}_i) \right\|,$$

where $\mathbf{t}_i = (\mathbf{x} - \mathbf{x}_i)$ (or, $\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x} - \mathbf{x}_i)$ if we use standardized version of the data) for $1 \leq i \leq n$. Then $\text{LSPD}_h(\mathbf{x}, F_n)$ is defined using (3) with $\Gamma_h(\mathbf{x}, F)$ replaced by $\Gamma_h(\mathbf{x}, F_n)$. Theorem 3 below shows the almost sure uniform convergence of $\text{LSPD}_h(\mathbf{x}, F_n)$ to its population counterpart $\text{LSPD}_h(\mathbf{x}, F)$. Similar convergence result for the empirical version of SPD has been proved in the literature (see, e.g., [10]).

Theorem 3 *Suppose that the density function* f *and the kernel* K *are bounded, and* K *has bounded first derivatives. Then, for any fixed* $h > 0$, $\sup_{\mathbf{X}} |LSPD_h(\mathbf{x}, F_n) - LSPD_h(\mathbf{x}, F)| \stackrel{a.s.}{\to} 0$ *as* $n \to \infty$.

From the proof of Theorem 3, it is easy to check that this almost sure uniform convergence also holds when $h \to \infty$. Under additional moment conditions on f and K, this holds for the $h \to 0$ case as well if $nh^{2d}/\log n \to \infty$ as $n \to \infty$ (see Remarks 2 and 3 after the proof of Theorem 3 in the Appendix). So, the result stated in parts (a) and (b) of Theorem 2 continue to hold for the empirical version of LSPD under appropriate assumptions.

Localization and kernelization of different notions of data depth have been considered in the literature [4, 1, 30, 19, 32]. The fact that LSPD_h tends to a constant multiple of the probability density function as $h \to 0$ is a crucial requirement for limiting Bayes optimality of classifiers based on this localized depth function. In [1], the authors proposed localized versions of simplicial depth and half-space depth, but the relationship between the local depth and the probability density function has been established only in the univariate case. A depth function based on inter-point distances has been developed in [30] to capture multimodality in a data set. Chen *et al.* [4] defined kernelized spatial depth in a reproducing kernel Hilbert space. In [19], the authors considered a generalized notion of Mahalanobis depth in reproducing kernel Hilbert spaces. However, there is no result connecting them to the probability density function. Infact, the kernelized spatial depth function becomes degenerate at the value $(1 - 1/\sqrt{2})$ as the tuning parameter goes to zero. Consequently, it becomes non-informative for small values of the tuning parameter. It will be appropriate to note here that none of the preceding authors used their proposed depth functions for constructing classifiers. Recently, in [33, 32], the authors proposed a notion of local depth and used it for supervised classification. But, their proposed version of local depth does not relate to the underlying density function either.

4 Multiscale classification based on LSPD

When the class distributions are elliptic, part (b) of Theorem 2 implies that LSPD_h with appropriately large choices of h lead to good classifiers. These large values may not be

appropriate for non-elliptic class distributions, but part (a) of Theorem 2 implies that LSPD_h with appropriately small choices of h lead to good classifiers for general nonparametric models for class densities. However, for small values of h, the empirical version of LSPD_h behaves like a nonparametric density estimate, and it suffers from the curse of dimensionality. So, the resulting classifier may have its statistical limitations for high-dimensional data.

We now consider two examples to demonstrate the above points. The first example (we call it **E3**) involves two multivariate normal distributions $\text{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $\text{N}_d(\mathbf{1}_d, 4\mathbf{I}_d)$. In the second example (we call it **E4**), both distributions are trimodal. The first class has the same density as in Figure 4 (i.e., an equal mixture of $\text{N}_d(\mathbf{0}_d, 0.25\mathbf{I}_d)$, $\text{N}_d(2\mathbf{1}_d, 0.25\mathbf{I}_d)$ and $\text{N}_d(4\mathbf{1}_d, 0.25\mathbf{I}_d)$), while the second class is an equal mixture of $\text{N}_d(\mathbf{1}_d, 0.25\mathbf{I}_d)$, $\text{N}_d(3\mathbf{1}_d, 0.25\mathbf{I}_d)$ and $\text{N}_d(5\mathbf{1}_d, 0.25\mathbf{I}_d)$. We consider the case $d = 10$ for **E3** and $d = 2$ for **E4**. For each of these two examples, we generated a training sample of size 100 from each class. The misclassification rate for the classifier based on LSPD_h was computed based on a test sample of size 500 (250 from each class). This procedure was repeated 100 times to calculate the average misclassification rate for different values of h. Figure 6 shows that the large (respectively, small) values of h yielded low misclassification rates in **E3** (respectively, **E4**). For small values of h, empirical LSPD_h behaved like a nonparametric density estimate that suffered from the curse of dimensionality in **E4**. Consequently, its performance deterioratesd. But,

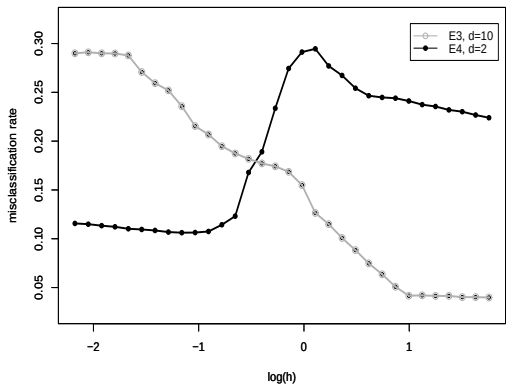


Figure 6: `Misclassification rates in examples` **E3** `and` **E4** `for the classifier based on` LSPD_h `for different values of` h.

for large h, the underlying elliptic structure was captured well by the proposed classifier. This provides a strong motivation for using a multi-scale approach in constructing the final classifier so that one can harness the strength of different classifiers corresponding to different levels of localization of SPD. One would expect that when aggregated, classifiers corresponding to different values of h will lead to improved misclassification rates. Usefulness of the multi-scale approach in combining different classifiers has been discussed in the classification literature by several authors including [8, 13, 14, 22].

A popular way of aggregation is to consider the weighted average of the estimated posterior probabilities computed for different values of h. There are various proposals for the choice of the weight function in the literature. Following [13], one can compute $\widehat{\Delta}_h$, the leave-one-out estimate of the misclassification rate of the classifier based on LSPD_h and use

$$W(h) \propto \exp\left[-\frac{1}{2}\frac{(\widehat{\Delta}_h - \widehat{\Delta}_0)^2}{\widehat{\Delta}_0(1-\widehat{\Delta}_0)/n}\right]$$

as the weight function, where $\widehat{\Delta}_0 = \min_h \widehat{\Delta}_h$. The exponential function helps to appropriately weighing up (respectively, down) the promising (respectively, poor) classifier resulting from different choices of the smoothing parameter h. However, $\int W(h)dh$ or $\int p(j|z_h(\mathbf{x}))W(h)dh$ may not be finite for some choices of $j \in \{1, 2, \ldots, J\}$. So, here we use a slightly modified weight function $W^*(h) = W(h)g(h)$, where g is a univariate Cauchy density with a large scale parameter and support restricted to have positive values only. Our final classifier, which we call the LSPD classifier, assigns an observation $\mathbf{x}$ to the j^*-th class, where

$$j^* = \arg\max_{1\le j\le J} \int_{h>0} W^*(h)\ p(j|\mathbf{z}_h(\mathbf{x}))dh = \arg\max_{1\le j\le J} \int_{h>0} W(h)g(h)\ p(j|\mathbf{z}_h(\mathbf{x}))dh.$$

Here $p(j|\mathbf{z}_h(\mathbf{x}))$ is as in equations (4) and (5) in Section 3. In practice, we first generate M independent observations $h_1, h_2, \ldots, h_M$ from g. For any given j and $\mathbf{x}$, we approximate $\int_{h>0} W(h)g(h)\ p(j|\mathbf{z}_h(\mathbf{x}))dh$ by $\sum_{i=1}^{M} W(h_i)\ p(j|\mathbf{z}_{h_i}(\mathbf{x}))/M$. The use of the Cauchy distribution with a large scale parameter (we use 100 in this article) helps us to generate small as well as large values of h. This is desirable in view of Theorem 2.

5 Classification of high-dimensional data

A serious practical limitation of many existing depth based classifiers is their computational complexity in high dimensions, and this makes such classifiers impossible to use even for moderately large dimensional data. Besides, depth functions that are based on random simplices formed by the data points (see [29, 42]), cannot be defined in a meaningful way if dimension of the data exceeds the sample size. Projection depth and Tukey's half-space depth (see, e.g., [42]) both become degenerate at zero for such high-dimensional data. Classification of high-dimensional data presents a substantial challenge to many nonparametric classification tools as well. We have seen in examples **E1** and **E2** (see Figure 2) that nonparametric classifiers like those based on k-NN and KDE can yield poor performance when data dimension is large. Some limitations of support vector machines for classification of high-dimensional data have also been noted in [31].

One of our primary motivations behind using SPD is its computational tractability, especially when the dimension is large. We now investigate the behavior of classifiers based on SPD and LSPD for such high-dimensional data. For this investigation, we assume that the observations are all standardized by a common positive definite matrix $\mathbf{\Sigma}$ for all J classes, and the following conditions are stated for those standardized random vectors, which are written as $\mathbf{X}$s for notational convenience.

(C1) Consider a random vector $\mathbf{X}_1 = (X_1^{(1)}, \ldots, X_1^{(d)})^T \sim F_j$. Assume that $a_j = \lim_{d\to\infty} d^{-1} \sum_{k=1}^{d} E(X_1^{(k)})^2$ exists for $1 \le j \le J$, and $d^{-1}\sum_{k=1}^{d}(X_1^{(k)})^2 \xrightarrow{a.s.} a_j$ as $d \to \infty$.

(C2) Consider two independent random vectors $\mathbf{X}_1 = (X_1^{(1)}, \ldots, X_1^{(d)})^T \sim F_j$ and $\mathbf{X}_2 = (X_2^{(1)}, \ldots, X_2^{(d)})^T \sim F_i$. Assume that $b_{ji} = \lim_{d\to\infty} d^{-1}\sum_{k=1}^{d} E(X_1^{(k)}X_2^{(k)})$ exists, and $d^{-1}\sum_{k=1}^{d} X_1^{(k)}X_2^{(k)} \xrightarrow{a.s.} b_{ji}$ as $d \to \infty$ for all $1 \le j, i \le J$.

It is not difficult to verify that for $\mathbf{X}_1 \sim F_j$ $(1 \le j \le J)$, if we assume that the sequence of variables $\{X_1^{(k)} - E(X_1^{(k)}) : k = 1, 2, \ldots\}$ centered at their means are independent with

uniformly bounded eighth moments (see Theorem 1 (2) in [21], p. 4110), or if we assume that they are m-dependent random variables with some appropriate conditions (see Theorem 2 in [5], p. 350), then the almost sure convergence in (C1) as well as (C2) holds. As a matter of fact, the almost sure convergence stated in (C1) and (C2) holds if we assume that for all $1 \leq j, i \leq J$, the sequences $\{(X_1^{(k)})^2 - E(X_1^{(k)})^2 : k = 1, 2, \ldots\}$ and $\{X_1^{(k)} X_2^{(k)} - E(X_1^{(k)} X_2^{(k)}) : k = 1, 2, \ldots\}$, where $\mathbf{X}_1 \sim F_j$ and $\mathbf{X}_2 \sim F_i$, are *mixingales* satisfying some appropriate conditions (see, e.g., Theorem 2 in [5], p. 350). Define $\sigma_j^2 = a_j - b_{jj}$ and $\nu_{ji} = b_{jj} - 2b_{ji} + b_{ii}$. For the random vector $\mathbf{X}_1 \sim F_j$, σ_j^2 is the limit of $d^{-1} \sum_{k=1}^{d} V(X_1^{(k)})$ as $d \to \infty$, where $V(Z)$ denotes the variance of a random variable Z. If we consider a second independent random vector $\mathbf{X}_2 \sim F_i$ with $i \neq j$, then ν_{ji} is the limit of $d^{-1} \sum_{k=1}^{d} \{E(X_1^{(k)}) - E(X_2^{(k)})\}^2$ as $d \to \infty$. In [15], the authors assumed a similar set of conditions to study the performance of the classifier based on support vector machines (SVM) with a linear kernel and the k-NN classifier with $k = 1$ as the data dimension grows to infinity. Similar conditions on observation vectors were also considered in [21] to study the consistency of principal components of the sample dispersion matrix for high-dimensional data. Under (C1) and (C2), the following theorem describes the behavior of $\mathbf{z}(\mathbf{x})$ and $\mathbf{z}_h(\mathbf{x})$ as d grows to infinity.

Theorem 4 *Suppose that the conditions (C1)-(C2) hold, and* $\mathbf{X} \sim F_j$ $(1 \leq j \leq J)$.

(a) $\mathbf{z}(\mathbf{X}) \xrightarrow{a.s.} (c_{j1}, \ldots, c_{jJ})^T = \mathbf{c}_j$ *as* $d \to \infty$, *where* $c_{jj} = 1 - \sqrt{\frac{1}{2}}$ *and* $c_{ji} = 1 - \sqrt{\frac{\sigma_j^2 + \nu_{ji}}{\sigma_j^2 + \sigma_i^2 + \nu_{ji}}}$ *for* $1 \leq j \neq i \leq J$.

(b) Assume that $h \to \infty$ *and* $d \to \infty$ *in such a way that* $\sqrt{d}/h \to 0$ *or* $A(> 0)$. *Then,* $\mathbf{z}_h(\mathbf{X}) \xrightarrow{a.s.} g(0)\mathbf{c}_j$ *or* $\mathbf{c}_j' = (g(e_{j1}A)c_{j1}, \ldots, g(e_{jJ}A)c_{jJ})^T$ *depending on whether* $\sqrt{d}/h \to 0$ *or* A, *respectively. Here* $K(\mathbf{t}) = g(\|\mathbf{t}\|)$, $e_{jj} = \sqrt{2}\sigma_j$ *and* $e_{ji} = \sqrt{\sigma_j^2 + \sigma_i^2 + \nu_{ji}}$ *for* $j \neq i$.

(c) Assume that $h > 1$, *and* $\sqrt{d}/h \to \infty$ *as* $d \to \infty$. *Then,* $\mathbf{z}_h(\mathbf{X}) \xrightarrow{a.s.} \mathbf{0}_J$.

The $\mathbf{c}_j$s as well as the $\mathbf{c}_j'$s in the statement of Theorem 4 are distinct for all $1 \leq j \leq J$ whenever either $\sigma_j^2 \neq \sigma_i^2$ or $\nu_{ji} \neq 0$ for all $1 \leq j \neq i \leq J$ (see Lemma 2 in Appendix). In such a case, part (a) of Theorem 4 implies that for large d, $\mathbf{z}(\mathbf{x})$ has good discriminatory power,

and our classifier based on SPD can discriminate well among the J populations. Further, it follows from part (b) that when both d and h grow to infinity in such a way that $\sqrt{d}/h \to 0$ or a positive constant, $\mathbf{z}_h(\mathbf{x})$ has good discriminatory power as well, and our classifier based on LSPD_h can yield low misclassification probability. However, part (c) shows that if $\sqrt{d}$ grows at a rate faster than h, $\mathbf{z}_h(\mathbf{x})$ becomes non-informative. Consequently, the classifier based on LSPD_h lead to high misclassification probability in this case.

6 Analysis of simulated data sets

We analysed some data sets simulated from elliptic as well as non-elliptic distributions. In each example, taking an equal number of observations from each of the two classes, we generated 500 training and test sets, each of size 200 and 500, respectively. We considered examples in dimensions 5 and 100. For classifiers based on SPD and LSPD, we used the usual sample dispersion matrix of the j-th ($j = 1, 2$) class as $\hat{\mathbf{\Sigma}}_j$ when $d = 5$. For $d = 100$, due to statistical instability of the sample dispersion matrix, we standardized each variable in a class by its sample standard deviation. Average test set misclassification rates of different classifiers (over 500 test sets) are reported in Table 1 along with their corresponding standard errors. To facilitate comparison, the corresponding Bayes risks are reported as well.

We compared our proposed classifiers with a pool of classifiers that include parametric classifiers like LDA and QDA, and nonparametric classifiers like those based on k-NN (with the Euclidean metric as the distance function) and KDE (with the Gaussian kernel). For the implementation of LDA and QDA in dimension 100, we used diagonal estimates of dispersion matrices as in the cases of SPD and LSPD. For k-NN and KDE, we used pooled versions of the scatter matrix estimates, which were chosen to be diagonal for $d = 100$. In Table 1, we report results for the multiscale methods of k-NN [13] and KDE [14] using the same weight function as described in Section 4. To facilitate comparison, we also considered SVM having the linear kernel and the radial basis function (RBF) kernel (i.e., $K_\gamma(\mathbf{x}, \mathbf{y}) = \exp\{-\gamma\|\mathbf{x} - \mathbf{y}\|^2\}$ with

the default value $\gamma = 1/d$ as in http://www.csie.ntu.edu.tw/∼cjlin/libsvm/); the classifier based on classification and regression trees (CART) and a boosted version of CART known as random forest (RF). For the implementation of SVM, CART and RF, we used the R codes available in the libraries `e1071` [6], `tree` [35] and `randomForest` [27], respectively. For classifiers based on SPD and LSPD, we wrote our own R codes using the library `VGAM` [40], and the codes are available at `https://sites.google.com/site/tijahbus/home/lspd`.

In addition, we compared the performance of our classifiers with two depth based classification methods; the classifier based on depth-depth plots (DD) [25] and the classifier based on maximum local depth [33] (LD). The DD classifier uses a polynomial of class depths (usually, half-space depth or projection depth is used, and depth is computed based on several random projections) to construct the separating surface. We used polynomials of different degrees and reported the best result in Table 1. For the LD classifier, we used the R package `DepthProc` and considered the best result obtained over a range of values for the localization parameter. However, in almost all cases, the performance of the LD classifier was inferior to that of the DD classifier. So, we did not report its misclassification rates in Table 1.

6.1 Examples with elliptic distributions

Recall examples **E1** and **E2** in Section 2 and example **E3** in Section 4 involving elliptic class distributions. In **E1** with $d = 5$, the DD classifier led to the lowest misclassification rate closely followed by SPD and LSPD classifiers, but in the case of $d = 100$, SPD and LSPD classifiers significantly outperformed all other classifiers considered here (see Table 1). The superiority of these two classifiers was evident in **E2** as well. In the case of $d = 5$, the difference between their misclassification rates was statistically insignificant, though the former had an edge. Since the class distributions were elliptic, dominance of the SPD classifier over the LSPD classifier was quite expected. However, this difference was found to be statistically significant when $d = 100$. In view of the normality of the class distributions,

QDA was expected to have the best performance in **E3**, and we observed the same. For $d = 5$, the DD classifier ranked second here, while the performance of SPD and LSPD classifiers was satisfactory. However, in the case of $d = 100$, SPD and LSPD classifiers again outperformed the DD classifier, and they correctly classified all the test set observations.

Table 1: Misclassification rates (in %) of different classifiers in simulated data sets.

Data set	Bayes risk	LDA	QDA	SVM (LIN)	SVM (RBF)	k-NN	KDE	CART	RF	DD	SPD	LSPD
$d = 5$												
E1	26.50	50.00 (0.20)	52.53 (0.19)	45.46 (0.11)	30.03 (0.09)	40.65 (0.13)	39.16 (0.11)	36.90 (0.13)	31.32 (0.09)	**27.92**∗ (0.11)	28.32 (0.10)	28.54 (0.11)
E2	0.00	47.43 (0.15)	42.44 (0.06)	43.92 (0.12)	38.06 (0.09)	37.64 (0.16)	34.29 (0.11)	39.10 (0.11)	34.26 (0.08)	26.68 (0.09)	**9.26** ∗ (0.09)	**9.42** (0.10)
E3	10.14	21.56 (0.09)	**11.09** ∗ (0.07)	22.09 (0.09)	11.74 (0.07)	18.16 (0.09)	16.94 (0.08)	19.18 (0.13)	13.77 (0.08)	**11.17** (0.07)	11.49 (0.07)	11.64 (0.07)
E4	2.10	40.52 (0.09)	42.41 (0.08)	36.16 (0.10)	25.08 (0.13)	**2.42** ∗ (0.03)	**2.55** (0.03)	15.52 (0.09)	4.98 (0.06)	33.04 (0.12)	10.07 (0.07)	**2.58** (0.03)
E5	2.04	41.17 (0.15)	5.97 (0.05)	32.14 (0.34)	8.12 (0.07)	9.44 (0.08)	9.26 (0.07)	4.82 (0.08)	**2.84** ∗ (0.03)	5.82 (0.05)	5.65 (0.06)	5.52 (0.06)
$d = 100$												
E1	0.48	50.29 (0.10)	50.67 (0.13)	46.85 (0.11)	24.97 (0.06)	44.57 (0.08)	49.99 (0.10)	35.72 (0.12)	25.14 (0.12)	24.99 (0.10)	**1.60** ∗ (0.11)	2.34 (0.12)
E2	0.00	43.77 (0.09)	46.13 (0.04)	43.99 (0.09)	40.32 (0.06)	49.96 (0.02)	49.22 (0.06)	40.30 (0.11)	32.36 (0.10)	27.56 (0.09)	**2.90** ∗ (0.08)	3.18 (0.09)
E3	0.00	0.46 (0.01)	**0.00** ∗ (0.00)	3.21 (0.05)	**0.00** ∗ (0.00)	49.99 (0.00)	49.98 (0.00)	17.40 (0.12)	0.02 (0.00)	1.92 (0.02)	**0.00** ∗ (0.00)	**0.00** ∗ (0.00)
E4	0.00	33.40 (0.00)	33.40 (0.00)	46.28 (0.10)	19.43 (0.09)	**0.00** ∗ (0.00)	**0.00** ∗ (0.00)	17.28 (0.00)	**0.00** ∗ (0.09)	23.15 (0.10)	**0.00** ∗ (0.00)	**0.00** ∗ (0.00)
E5	0.00	46.74 (0.29)	**0.00** ∗ (0.00)	44.45 (0.31)	7.83 (0.15)	44.01 (0.21)	49.98 (0.04)	3.32 (0.11)	**0.00** ∗ (0.00)	3.12 (0.10)	**0.00** ∗ (0.00)	**0.00** ∗ (0.00)

The figure marked by '∗' is the best misclassification rate observed in an example. The other figures in bold (if any) are the misclassification rates whose differences with the best misclassification rate are statistically insignificant at the 5% level when the usual large sample test for proportion was used for comparison.

In all these examples, the Bayes classifier had non-linear class boundaries. So, LDA and SVM with linear kernel could not perform well. The performance of SVM with the RBF kernel was relatively better. In **E3**, it had competitive misclassification rates for both values of d. k-NN and KDE had comparable performance in the case of $d = 5$, but in the high-dimensional case ($d = 100$), they misclassified almost half of the test cases. In [15], the authors derived some conditions under which the k-NN classifier tends to classify all observations to a single class when the data dimension increases to infinity. These conditions

hold in this example. It can also be shown that the classifier based on KDE with equal prior probabilities have the same problem in high dimensions.

6.2 Examples with non-elliptic distributions

Recall the trimodal example **E4** discussed in Section 4. In this example, the LSPD classifier and the nonparametric classifiers based on k-NN and KDE significantly outperformed all other classifiers in the case of $d = 5$. The differences between the misclassification rates of these three classifiers was statistically insignificant. Interestingly, along with these classifiers, the SPD classifier also led to zero misclassification rate for $d = 100$. The DD classifier, LDA, QDA and SVM did not have satisfactory performance in this example.

The final example (we call it **E5**) is with exponential distributions, where the component variables are independently distributed in both classes. The i-th variable in the first (respectively, the second) class is exponential with scale parameter $d/(d - i + 1)$ (respectively, $d/2i$). Further, the second class has a location shift such that the difference between the mean vectors of the two classes is $\frac{1}{d}\mathbf{1}_d = (1/d, \ldots, 1/d)^T$. Recall that Figure 5 shows the density contours of the first class when $d = 2$. In this example, the RF classifier had the best performance followed by CART when $d = 5$. DD, SPD and LSPD classifiers also performed well, and their misclassification rates were significantly lower than all other classifiers. Linear classifiers like LDA and SVM with linear kernel failed to perform well. Note that as d increases, the difference between the locations of these two classes shrinks to zero. This results in high misclassification rates for these linear classifiers when $d = 100$. QDA performed reasonably well, and like SPD, LSPD and RF classifiers, it correctly classified all the test cases when $d = 100$. The DD classifier led to an average misclassification rate of 3.12%. Again, the classifiers based on k-NN and KDE had poor performance for $d = 100$. This is due to the same reason as in **E3** (see also [15]). Note that even in these examples with non-elliptic distributions, the SPD classifier performed well for high-dimensional data.

This can be explained using part (a) of Theorem 4. These examples also demonstrate that for non-elliptic or multimodal data, if not better, our LSPD classifier can perform as good as popular nonparametric classifiers. In fact, this adjustment of LSPD classifier is automatic in view of the multiscale approach developed in Section 4.

7 Analysis of benchmark data sets

We analyzed some benchmark data sets for further evaluation of our proposed classifiers. The biomedical data set is taken from the CMU data archive (http://lib.stat.cmu.edu/datasets/), the growth data set is obtained from [34], the colon data set is available in [2] (and also at the R-package 'rda'), and the lightning 2 data set is taken from the UCR time series classification archive (http://www.cs.ucr.edu/∼eamonn/time_series_data/). The remaining data sets are taken from the UCI machine learning repository (http://archive.ics.uci.edu/ml/). Descriptions of these data sets are available at these sources. In the case of biomedical data, we did not consider observations with missing values. Satellite image (satimage) data set has specific training and test samples. For this data set, we report the test set misclassification rates of different classifiers. If a classifier had misclassification rate ϵ, its standard error was computed as $\sqrt{\epsilon(1-\epsilon)/(\text{the size of the test set})}$. In all other data sets, we formed the training and the test sets by randomly partitioning the data, and this random partitioning was repeated 500 times to generate new training and test sets. The average test set misclassification rates of different classifiers are reported in Table 2 along with their corresponding standard errors. The sizes of the training and the test sets in each partition are also reported in this table. Since the codes for the DD classifier are available only for two class problems, we could use it only in cases of biomedical and Parkinson's data, where it yielded misclassification rates of 12.54% and 14.48%, respectively, with corresponding standard error of 0.18% and 0.15%. In the case of growth data, where training sample size from each class was smaller than the dimension, the values of randomized versions of half-space depth and

projection depth were zero for almost all observations. Due to this problem, the DD classifier could not be used. We used the maximum LD classifier on these real data sets, but in most of the cases, its performance was not satisfactory. So, we do not report them in Table 2.

Table 2: Misclassification rates (in %) of different classifiers in real data sets.

Data set	Biomed	Parkinson's	Wine	Waveform	Vehicle	Satimage	Growth	Lightning 2	Colon
Dimension (d)	4	22	13	21	18	36	31	637	2000
Classes (J)	2	2	3	3	4	6	2	2	2
Training size	100	97	100	300	423	4435	46	60	31
Test size	94	98	78	501	423	2000	47	61	31
LDA	15.66	30.93	2.00	19.90	22.49	16.06	29.15	32.51	**14.03** $*$
	(0.14)	(0.12)	(0.06)	(0.15)	(0.07)	(0.82)	(0.34)	(0.35)	(0.20)
QDA	**12.57**	xxxx	2.46	21.12	**16.38**	14.14	xxxx	xxxx	xxxx
	(0.13)	xxxx	(0.09)	(0.15)	(0.07)	(0.78)	xxxx	xxxx	xxxx
SVM (LIN)	22.03	15.31	3.64	18.88	21.20	15.18	5.16	35.64	16.38
	(0.13)	(0.12)	(0.09)	(0.07)	(0.07)	(0.80)	(0.12)	(0.35)	(0.23)
SVM (RBF)	**12.76**	13.69	1.86	16.08	25.57	30.99	**4.66** $*$	28.73	35.48
	(0.13)	(0.10)	(0.06)	(0.07)	(0.08)	(1.03)	(0.11)	(0.32)	(0.00)
k-NN	17.74	14.42	1.98	16.37	21.80	**9.23** $*$	**4.48**	30.09	22.47
	(0.15)	(0.16)	(0.06)	(0.08)	(0.08)	(0.65)	(0.10)	(0.20)	(0.27)
KDE	16.67	**11.01** $*$	**1.36** $*$	23.83	21.21	19.81	**4.79**	28.11	23.20
	(0.14)	(0.12)	(0.05)	(0.03)	(0.07)	(0.89)	(0.13)	(0.30)	(0.28)
CART	17.69	16.63	10.99	56.61	31.41	53.43	17.40	33.96	28.78
	(0.18)	(0.20)	(0.22)	(0.12)	(0.10)	(0.56)	(0.25)	(0.34)	(0.35)
RF	13.23	11.58	2.12	57.02	25.52	30.91	9.67	**22.08**$*$	19.10
	(0.14)	(0.15)	(0.06)	(0.12)	(0.07)	(0.48)	(0.25)	(0.34)	(0.30)
SPD	**12.53**	15.44	2.34	**15.12** $*$	**16.35** $*$	12.59	14.64	27.70	19.98
	(0.21)	(0.15)	(0.08)	(0.06)	(0.08)	(0.74)	(0.28)	(0.30)	(0.31)
LSPD	**12.49** $*$	**11.35**	1.85	**15.36**	17.15	12.59	5.10	27.46	20.51
	(0.15)	(0.11)	(0.07)	(0.06)	(0.08)	(0.74)	(0.12)	(0.30)	(0.33)

The figure marked by '$*$' is the best misclassification rate observed for a data set. The other figures in bold (if any) are the misclassification rates whose differences with the best misclassification rate are statistically insignificant at the 5% level. Because of the singularity of the estimated class dispersion matrices, QDA could note be used in some cases and those are marked by 'xxxx'.

In biomedical and vehicle data sets, scatter matrices of the competing classes were very different. So, QDA had significant improvement over LDA. In fact, its misclassification rates of QDA were close to the best ones. In both of these data sets, the class distributions were nearly elliptic (this can be verified using the diagnostic plots suggested in [26]). The SPD classifiers utilized the ellipticity of the class distributions to outperform the nonparametric classifiers. The LSPD classifier could compete with the SPD classifier in the biomedical data. But in the vehicle data, where the evidence of ellipticity was much stronger, it had a slightly

higher misclassification rate.

In the Parkinson's data set, we could not use QDA because of the singularity of the estimated class dispersion matrices. So, we used the estimated pooled dispersion matrix for standardization in our classifiers. In this data set, all nonparametric classifiers had significantly lower misclassification rates than LDA. Among them, the classifier based on KDE had the lowest misclassification rate. The performance of LSPD classifier was also competitive. Since the underlying distributions were non-elliptic, the LSPD classifier significantly outperformed the SPD classifier. We observed almost the same phenomenon in the wine data set as well, where the classifier based on KDE yielded the best misclassification rate followed by the LSPD classifier. In these two data sets, although the data dimension was quite high, all competing classes had low intrinsic dimensions (can be estimated using [24]). So, the nonparametric methods like KDE were not much affected by the curse of dimensionality. Recall that for small values of h, LSPD_h performs like KDE. Therefore, the difference between the misclassification rates of KDE and LSPD classifiers was statistically insignificant.

In the waveform data set, the SPD classifier had the best misclassification rate. In this data set, the class distributions were nearly elliptic. So, the SPD classifier was expected to perform well. As the LSPD classifier is quite flexible, it yielded competitive misclassification rates. Here, the class distributions were not normal (can be checked using the method in [36]), and they did not have low intrinsic dimensions. As a result, other parametric as well as nonparametric classifiers had relatively higher misclassification rates.

Recall that in the satimage data set, results are based on a single training and a single test set. So, the standard errors of the misclassification rates were high for all classifiers, and this makes it difficult to compare the performance of different classifiers. In this data set, k-NN classifiers led to the lowest misclassification rate, but SPD and LSPD classifiers performed better than all other classifiers. Nonlinear SVM, CART and RF had quite high misclassification rates.

We further analyzed some data sets, where the sample size was quite small compared to data dimension. In these data sets, we worked with unstandardized observations. Instead of using the estimated pooled dispersion matrix, we used the identity matrix for implementation of LDA. The growth data set contains growth curves of males and females, which are smooth and monotonically increasing functions. Because of high dependence among the measurement variables, the class distributions had low intrinsic dimensions, and they were non-elliptic. As a result, the nonparametric classifiers performed well. SVM with the RBF kernel had the best misclassification rate, but those of k-NN, KDE and LSPD classifiers were also comparable. Good performance of the linear SVM classifier indicates that there was a good linear separability between the two classes, but LDA failed to figure it out.

The lightning 2 data set consists of observations that are realizations of time series. In this data set, RF had the best performance followed by the LSPD classifier. Here also, we observed non-elliptic class distributions with low intrinsic dimensions [24]. This justifies the good performance of the classifiers based on k-NN and KDE. The SPD classifier also had competitive misclassification rates because of the flexibility of GAM. In fact, it yielded the third best performance in this data set.

Finally, we analyzed the colon data set, which contains micro-array expressions of 2000 genes for some 'normal' and 'colon cancer' tissues. In this data set, there was good linear separability among the observations from the two classes. So, LDA and linear SVM had lower misclassification rates than all other classifiers. Among the nonparametric classifiers, RF had the best performance closely followed by the SPD classifier. These two classifiers were less affected by the curse of dimensionality. Recall that LSPD_h with large bandwidth h approximates SPD. Because of this automatic adjustment, the LSPD classifier could nearly match the performance of the SPD classifier.

To compare the overall performance of different classifiers, following the idea of [3, 9], we computed their efficiency scores on different data sets. For a data set, if T classifiers

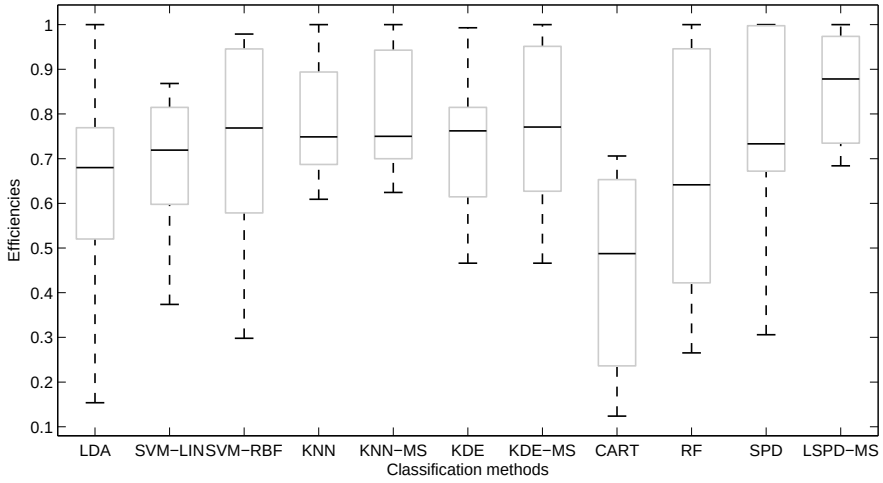


Figure 7: `Overall efficiencies of different classifiers.`

have misclassification rates $\epsilon_1, \ldots, \epsilon_T$, the efficiency of the t-th classifier (e_t) is defined as $e_t = \epsilon_0/\epsilon_t$, where $\epsilon_0 = \min_{1 \le t \le T} \epsilon_t$. Clearly, in any data, the best classifier has $e_t = 1$, while a lower value of e_t indicates the lack of efficiency of the t-th classifier. In each of these benchmark data sets, we computed this ratio for all classifiers, and they are graphically represented by box plots in Figure 7. This figure clearly shows the superiority of the LSPD classifier (with the highest median value of 0.88) over its competitors. We did not consider QDA for comparison because it could not be used for some of the data sets.

8 Concluding remarks

In this article, we develop and study classifiers constructed by fitting a nonparametric additive logistic regression model to features extracted from the data using SPD as well as its localized version, LSPD. The SPD classifier can be viewed as a generalization of parametric classifiers like LDA and QDA. When the underlying class distributions are elliptic, it has Bayes optimality. For large values of h, while LSPD_h behaves like SPD, for small values of h, it captures the underlying density. So, the multiscale classifier based on LSPD is flexible, and it overcomes several drawbacks associated with SPD and other existing depth based

classifiers. When the underlying class distributions are elliptic but not normal, both SPD and LSPD classifiers outperform popular parametric classifiers like LDA and QDA as well as nonparametric classifiers. In the case of non-elliptic or multi-modal distributions, while SPD may fail to extract meaningful discriminatory features, the LSPD classifier can compete with other nonparametric methods. Moreover, for high-dimensional data, while traditional nonparametric methods suffer from the curse of dimensionality, both SPD and LSPD classifiers can lead to low misclassification probabilities. Analyzing several simulated and benchmark data sets, we have amply demonstrated this. In high-dimensional benchmark data sets, the class distributions had low intrinsic dimensions due to high correlation among the the measurement variables [24]. Moreover, the competing classes differed mainly in their locations. As a consequence, though the proposed LSPD classifier had the best overall performance in benchmark data sets, its superiority over other nonparametric methods was not as prominent as it was in the simulated examples.

Appendix : Proofs and Mathematical Details

Lemma 1 : If F has a spherically symmetric density $f(\mathbf{x}) = g(\|\mathbf{x}\|)$ on $\mathbb{R}^d$ with $d > 1$, then $\|E_F[u(\mathbf{x} - \mathbf{X})]\|$ is a non-negative monotonically increasing function of $\|\mathbf{x}\|$.

Proof of Lemma 1 : In view of spherical symmetry of $f(\mathbf{x})$, $S(\mathbf{x}) = \|E_F[u(\mathbf{x} - \mathbf{X})]\|$ is invariant under orthogonal transformations of $\mathbf{x}$. Consequently, $S(\mathbf{x}) = \eta(\|\mathbf{x}\|)$ for some non-negative function η. Consider now $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\|\mathbf{x}_1\| < \|\mathbf{x}_2\|$. Using spherical symmetry of $f(\mathbf{x})$, without loss of generality, we can assume $\mathbf{x}_i = (t_i, 0, \ldots, 0)^T$ for $i = 1, 2$ such that $|t_1| < |t_2|$. For any $\mathbf{x} = (t, 0, \ldots, 0)^T$, we have

$$S(\mathbf{x}) = \left| E_F \left[\frac{(t - X_1)}{\sqrt{(t - X_1)^2 + X_2^2 + \ldots + X_d^2}} \right] \right|,$$

due to spherical symmetry of $f(\mathbf{x})$. Note also that for any $\mathbf{x} \in \mathbb{R}^d$ with $d > 1$, $E_F[\|\mathbf{x} - \mathbf{X}\|]$ is a strictly convex function of $\mathbf{x}$ in this case. Consequently, it is a strictly convex function

of t. Observe now that $S(\mathbf{x})$ with this choice of $\mathbf{x}$ is the absolute value of the derivative of $E_F[\|\mathbf{x}-\mathbf{X}\|]$ w.r.t. t. This derivative is a symmetric function of t that vanishes at $t=0$. Hence, $S(\mathbf{x})$ is an increasing function of $|t|$, and this proves that $\eta(\|\mathbf{x}_1\|) < \eta(\|\mathbf{x}_2\|)$. □

Proof of Theorem 1 : If the population distribution $f_j(\mathbf{x})$ $(1 \le j \le J)$ is elliptically symmetric, we have $f_j(\mathbf{x}) = |\mathbf{\Sigma}_j|^{-1/2} g_j(\delta(\mathbf{x}, F_j))$, where $\delta(\mathbf{x}, F_j) = \{(\mathbf{x}-\boldsymbol{\mu}_j)^T\ \mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\}^{1/2} = \|\mathbf{\Sigma}_j^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_j)\|$. Since $\mathrm{SPD}(\mathbf{x}, F_j) = 1 - \|E\{u(\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j))\}\|$ is affine invariant, it is a function of $\delta(\mathbf{x}, F_j)$, the Mahalanobis distance. Again, since $\mathbf{\Sigma}_j^{-1/2}(\mathbf{X}-\boldsymbol{\mu}_j)$ has a spherically symmetric distribution with its center at the origin, from Lemma 1 it follows that $\mathrm{SPD}(\mathbf{x}, F_j)$ is a monotonically decreasing function of $\delta(\mathbf{x}, F_j)$. So, $\delta(\mathbf{x}, F_j)$ is also a function of $\mathrm{SPD}(\mathbf{x}, F_j)$. Therefore, $f_j(\mathbf{x})$, which is a function of $\delta(\mathbf{x}, F_j)$, can also be expressed as

$$f_j(\mathbf{x}) = \psi_j(\mathrm{SPD}(\mathbf{x}, F_j)) \text{ for all } 1 \le j \le J,$$

where ψ_j is an appropriate real-valued function that depends on g_j. Now, one can check that

$$\log\{p(j|\mathbf{x})/p(J|\mathbf{x})\} = \log(\pi_j/\pi_J) + \log\psi_j(\mathrm{SPD}(\mathbf{x}, F_j)) - \log\psi_J(\mathrm{SPD}(\mathbf{x}, F_J)).$$

for $1 \le j \le (J-1)$. Now, for $1 \le j \ne i \le (J-1)$, define $\varphi_{jj}(z) = \log\pi_j + \log\psi_j(z)$ and $\varphi_{ij}(z) = 0$. So, if we define $\varphi_{1J}(z) = \ldots = \varphi_{(J-1)J}(z) = -\log\pi_J - \log\psi_J(z)$, the proof of the theorem is complete. □

Remark 1 : If $f_j(\mathbf{x})$ is unimodal, $\psi_j(z)$ is monotonically increasing for $1 \le j \le J$. Moreover, if the distributions differ only in their locations, the $\psi_j(z)$s are same for all class. In that case, $f_j(\mathbf{x}) > f_i(\mathbf{x}) \Leftrightarrow \delta(\mathbf{x}, F_j) > \delta(\mathbf{x}, F_i)$ for $1 \le i \ne j \le J$, and hence the classifier turns out to be the maximum depth classifier.

Proof of Theorem 2 (a) : Let $h < 1$. For any fixed $\mathbf{x} \in \mathbb{R}^d$ and the distribution function F_j, we have $\mathrm{LSPD}_h(\mathbf{x}, F_j) = E_{F_j}[K_h(\mathbf{t})] - \|E_{F_j}[K_h(\mathbf{t})u(\mathbf{t})]\|$, where $\mathbf{t} = {\mathbf{\Sigma}_j}^{-1/2}(\mathbf{x}-\mathbf{X})$ for $1 \le j \le J$. For the first term in the expression of $\mathrm{LSPD}_h(\mathbf{x}, F_j)$ above, we have

$$E_{F_j}[K_h(\mathbf{t})] = \int_{\mathbb{R}^d} h^{-d} K_h(\mathbf{\Sigma}_j^{-1/2}(\mathbf{x}-\mathbf{v})) f_j(\mathbf{v}) d\mathbf{v} = |\mathbf{\Sigma}_j|^{1/2} \int_{\mathbb{R}^d} K(\mathbf{y}) f_j(\mathbf{x} - h\mathbf{\Sigma}_j^{1/2}\mathbf{y}) d\mathbf{y},$$

where $\mathbf{y} = h^{-1}\mathbf{\Sigma}_j^{-1/2}(\mathbf{x}-\mathbf{v})$. So, using Taylor's expansion of $f_j(\mathbf{x})$, we get

$$E_{F_j}[K_h(\mathbf{t})] = |\mathbf{\Sigma}_j|^{1/2} f_j(\mathbf{x}) - h|\mathbf{\Sigma}_j|^{1/2} \int_{\mathbb{R}^d} K(\mathbf{y})\ (\mathbf{\Sigma}_j^{1/2}\mathbf{y})'\nabla f_j(\boldsymbol{\xi}) d\mathbf{y},$$

where $\boldsymbol{\xi}$ lies on the line joining $\mathbf{x}$ and $(\mathbf{x}-h\mathbf{\Sigma}_j^{1/2}\mathbf{v})$. So, using the Cauchy-Scawartz inequality, one gets $\left|E_{F_j}[K_h(\mathbf{t})] - |\mathbf{\Sigma}_j|^{1/2} f_j(\mathbf{x})\right| \leq h|\mathbf{\Sigma}_j|^{1/2}\lambda_j^{1/2} M_j^{\circ} M_K$, where $M_j^{\circ} = \sup_{\mathbf{x}\in\mathbb{R}^d} \|\nabla f_j(\mathbf{x})\|$, $M_K = \int \|\mathbf{y}\| K(\mathbf{y}) d\mathbf{y}$, and λ_j is the largest eigenvalue of $\mathbf{\Sigma}_j$ for $1 \leq j \leq J$. This implies $\left|E_{F_j}[K_h(\mathbf{t})] - |\mathbf{\Sigma}_j|^{1/2} f_j(\mathbf{x})\right| \to 0$ as $h \to 0$ for $1 \leq j \leq J$.

For the second term in the expression of $\mathrm{LSPD}_h(\mathbf{x}, F_j)$, a similar argument yields

$$\begin{aligned} E_{F_j}[K_h(\mathbf{t})u(\mathbf{t})] &= |\mathbf{\Sigma}_j|^{1/2} \int_{\mathbb{R}^d} K(\mathbf{y})u(\mathbf{y}) f_j(\mathbf{x} - h\mathbf{\Sigma}_j^{1/2}\mathbf{y}) d\mathbf{y} \\ &= -h|\mathbf{\Sigma}_j|^{1/2} \int_{\mathbb{R}^d} K(\mathbf{y})u(\mathbf{y})\ (\mathbf{\Sigma}_j^{1/2}\mathbf{y})'\nabla f_j(\boldsymbol{\xi}) d\mathbf{y} \quad (\text{since} \int K(\mathbf{y})u(\mathbf{y})d\mathbf{y} = \mathbf{0}). \end{aligned}$$

So, $\|E_{F_j}[K_h(\mathbf{t})u(\mathbf{t})]\| \leq h|\mathbf{\Sigma}_j|^{1/2}\lambda_j^{1/2} M_j^{\circ} M_K \to 0$ as $h \to 0$, and hence, $\mathrm{LSPD}_h(\mathbf{x}, F_j) \to |\mathbf{\Sigma}_j|^{1/2} f_j(\mathbf{x})$ as $h \to 0$. Consequently, $\mathbf{z}_h(\mathbf{x}) \to (|\mathbf{\Sigma}_1|^{1/2} f_1(\mathbf{x}), \ldots, |\mathbf{\Sigma}_J|^{1/2} f_J(\mathbf{x}))^T$ as $h \to 0$. □

Proof of Theorem 2 (b) : Here we consider the case $h > 1$. Consider any fixed $\mathbf{x} \in \mathbb{R}^d$ and any fixed j $(1 \leq j \leq J)$. For any fixed $\mathbf{t}$, since $K(\mathbf{t}/h) \to K(\mathbf{0})$ as $h \to \infty$, using Dominated Convergence Theorem (note that K is bounded), one can show that

$$\mathrm{LSPD}_h(\mathbf{x}, F_j) \to K(\mathbf{0})\mathrm{SPD}(\mathbf{x}, F_j) \text{ as } h \to \infty.$$

So, $\mathbf{z}_h(\mathbf{x}) \to (K(\mathbf{0})\mathrm{SPD}(\mathbf{x}, F_1), \ldots, K(\mathbf{0})\mathrm{SPD}(\mathbf{x}, F_J))^T$ as $h \to \infty$. □

Proof of Theorem 3 : Define the sets $B_n = \{\mathbf{x} = (x_1, \ldots, x_d) : \|\mathbf{x}\| \leq \sqrt{d}n\}$, and $A_n = \{\mathbf{x} :\ n^2 x_i \text{ is an integer and } |x_i| \leq n \text{ for all } 1 \leq i \leq d\}$. Clearly $A_n \subset B_n \subset \mathbb{R}^d$, the set B_n is a closed ball and the set A_n has cardinality $(2n^3+1)^d$. We will prove the almost sure (a.s.) uniform convergence on the three sets: (i) on A_n (ii) on $B_n \setminus A_n$, and (iii) on B_n^c.

Consider any fixed $h \in (0, 1]$. Recall that for this choice of h, $\mathrm{LSPD}_h(\mathbf{x}, F)$ (see equation (3)) and $\mathrm{LSPD}_h(\mathbf{x}, F_n)$ are defined as follows:

$$\mathrm{LSPD}_h(\mathbf{x}, F_n) = \frac{1}{nh^d}\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i)) - \left\| \frac{1}{nh^d}\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i))\ u(\mathbf{x}-\mathbf{X}_i) \right\|,$$

and $\text{LSPD}_h(\mathbf{x}, F) = h^{-d}E[K(h^{-1}(\mathbf{x}-\mathbf{X}))] - h^{-d}\|E[K(h^{-1}(\mathbf{x}-\mathbf{X}))\ u(\mathbf{x}-\mathbf{X})]\|$.

(i) Define $\mathbf{Z}_i = K(h^{-1}(\mathbf{x}-\mathbf{X}_i))u(\mathbf{x}-\mathbf{X}_i) - E[K(h^{-1}(\mathbf{x}-\mathbf{X}))u(\mathbf{x}-\mathbf{X})]$ for $1 \le i \le n$. Note that $\mathbf{Z}_i$s are independent and identically distributed (i.i.d.) with $E(\mathbf{Z}_i) = \mathbf{0}$ and $\|\mathbf{Z}_i\| \le 2K(\mathbf{0})$. Using the exponential inequality for sums of i.i.d. random vectors (see p. 491 of [41]), for any fixed $\epsilon > 0$, we get $P\Big(\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\| \ge \epsilon\Big) \le 2e^{-C_0 n\epsilon^2}$, where C_0 is a positive constant that depends on $K(\mathbf{0})$ and ϵ. This now implies that

$$
\begin{aligned}
&P\Bigg(\Big\|\tfrac{1}{nh^d}\textstyle\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i))u(\mathbf{x}-\mathbf{X}_i)\Big\| - \Big\|h^{-d}E[K(h^{-1}(\mathbf{x}-\mathbf{X}))u(\mathbf{x}-\mathbf{X})]\Big\| \ge \epsilon\Bigg) \\
&\le P\Bigg(\Big\|\tfrac{1}{nh^d}\textstyle\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i))u(\mathbf{x}-\mathbf{X}_i) - h^{-d}E[K(h^{-1}(\mathbf{x}-\mathbf{X}))u(\mathbf{x}-\mathbf{X})]\Big\| \ge \epsilon\Bigg) \\
&= P\Big(\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\| \ge h^d\epsilon\Big) \le 2e^{-C_0 nh^{2d}\epsilon^2}. \qquad (6)
\end{aligned}
$$

For a fixed value of h, since $\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i))$ is a sum of i.i.d. bounded random variables, using Bernstein's inequality, we also have

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i)) - E[K(h^{-1}(\mathbf{x}-\mathbf{X}))]\Big| \ge \epsilon\Big) \le 2e^{-C_1 n\epsilon^2}$$

for some suitable positive constant C_1. This implies

$$P\Big(\Big|\frac{1}{nh^d}\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i)) - h^d E[K(h^{-1}(\mathbf{x}-\mathbf{X}))]\Big| \ge \epsilon\Big) \le 2e^{-C_1 nh^{2d}\epsilon^2}. \qquad (7)$$

Combining equations (6) and (7), we get $P(|\text{LSPD}(\mathbf{x}, F_n) - \text{LSPD}(\mathbf{x}, F)| \ge \epsilon) \le C_3 e^{-C_4 nh^{2d}\epsilon^2}$ for some suitable constants C_3 and C_4. Since the cardinality of A_n is $(n^3+1)^d$, we have

$$P(\sup_{\mathbf{x}\in A_n} |\text{LSPD}(\mathbf{x}, F_n) - \text{LSPD}(\mathbf{x}, F)| \ge \epsilon) \le C_3(n^3+1)^d e^{-C_4 nh^{2d}\epsilon^2}. \qquad (8)$$

Now, $\sum_{n\ge 1}(n^3+1)^d e^{-C_4 nh^{2d}\epsilon^2} < \infty$. So, a simple application of Borel-Cantelli lemma implies that $\sup_{\mathbf{x}\in A_n} |\text{LSPD}_h(\mathbf{x}, F_n) - \text{LSPD}_h(\mathbf{x}, F)| \overset{a.s.}{\to} 0$ as $n \to \infty$.

(ii) Consider the set $B_n \setminus A_n$. Note that given any $\mathbf{x}$ in $B_n \setminus A_n$, there exists $\mathbf{y} \in A_n$ such that $\|\mathbf{x}-\mathbf{y}\| \le \sqrt{2}/n^2$. First we will show that $|\text{LSPD}(\mathbf{y}, F_n) - \text{LSPD}(\mathbf{x}, F_n)| \overset{a.s.}{\to} 0$ as $n \to \infty$. Using the mid-value theorem, one gets

$$\Big|\frac{1}{nh^d}\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i)) - \frac{1}{nh^d}\sum_{i=1}^n K(h^{-1}(\mathbf{y}-\mathbf{X}_i))\Big| \le \frac{1}{nh^{d+1}}\sum_{i=1}^n \big|(\mathbf{x}-\mathbf{y})^T\nabla K[(\boldsymbol{\xi}-\mathbf{X}_i)/h]\big|,$$

where $\boldsymbol{\xi}$ lies on the line joining $\mathbf{x}$ and $\mathbf{y}$. Note that the right hand side is less than $\frac{M_K'}{h^{d+1}}\frac{\sqrt{2}}{n^2}$, where $M_K' = \sup_{\mathbf{t}} \|\nabla K(\mathbf{t})\|$. This upper bound is free of $\mathbf{x}$, and goes to 0 as $n \to \infty$. Now,

$$\left\|\frac{1}{nh^d}\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i))u(\mathbf{x}-\mathbf{X}_i)\right\| - \left\|\frac{1}{nh^d}\sum_{i=1}^n K(h^{-1}(\mathbf{y}-\mathbf{X}_i))u(\mathbf{y}-\mathbf{X}_i)\right\|$$

$$\leq \left\|\frac{1}{nh^d}\sum_{i=1}^n [K(h^{-1}(\mathbf{x}-\mathbf{X}_i))u(\mathbf{x}-\mathbf{X}_i) - K(h^{-1}(\mathbf{y}-\mathbf{X}_i))u(\mathbf{y}-\mathbf{X}_i)]\right\|$$

$$\leq \left|\frac{1}{nh^d}\sum_{i=1}^n [K(h^{-1}(\mathbf{x}-\mathbf{X}_i)) - K(h^{-1}(\mathbf{y}-\mathbf{X}_i))]\right| + K(\mathbf{0})\left\|\frac{1}{nh^d}\sum_{i=1}^n \{u(\mathbf{x}-\mathbf{X}_i) - u(\mathbf{y}-\mathbf{X}_i)\}\right\|.$$

We have already proved that the first part converges to 0 in a.s. sense. For the second part, consider a ball of radius $1/n$ around $\mathbf{x}$ (say, $B(\mathbf{x}, 1/n)$). Now,

$$K(\mathbf{0})\left\|\frac{1}{nh^d}\sum_{i=1}^n \{u(\mathbf{x}-\mathbf{X}_i) - u(\mathbf{y}-\mathbf{X}_i)\}\right\| \leq \left|\frac{2K(\mathbf{0})}{nh^d}\sum_{i=1}^n I\{\mathbf{X}_i \in B(\mathbf{x}, 1/n)\}\right| + \frac{2nK(\mathbf{0})}{h^d}\|\mathbf{x}-\mathbf{y}\|$$

$$\leq \frac{2K(\mathbf{0})}{h^d}\left|\frac{1}{n}\sum_{i=1}^n I\{\mathbf{X}_i \in B(\mathbf{x}, 1/n)\} - P\{\mathbf{X}_1 \in B(\mathbf{x}, 1/n)\}\right| + \frac{2K(\mathbf{0})}{h^d}P\{\mathbf{X}_1 \in B(\mathbf{x}, 1/n)\} + \frac{2nK(\mathbf{0})\sqrt{2}}{n^2h^d}.$$

Note that $I\{\mathbf{X}_i \in B(\mathbf{x}, 1/n)\}$ are i.i.d. bounded random variables with expectation $P\{\mathbf{X}_1 \in B(\mathbf{x}, 1/n)\}$. So, the a.s. convergence of the first term follows from Bernstein's inequality. Since $P\{\mathbf{X}_1 \in B(\mathbf{x}, 1/n)\} \leq M_f n^{-d}$ (where $M_f = \sup_{\mathbf{x}} f(\mathbf{x}) < \infty$), the second term converges to 0. The third term also converges to 0 as $n \to \infty$. Therefore, we have $|\text{LSPD}(\mathbf{x}, F_n) - \text{LSPD}(\mathbf{y}, F_n)| \overset{a.s}{\to} 0$ as $n \to \infty$.

Similarly, one can prove that $|\text{LSPD}(\mathbf{x}, F) - \text{LSPD}(\mathbf{y}, F)| \overset{a.s}{\to} 0$ as $n \to \infty$. Note that in the arguments above, all bounds are free from $\mathbf{x}$ and $\mathbf{y}$. We have also proved that $\sup_{\mathbf{y}\in A_n} |\text{LSPD}(\mathbf{y}, F_n) - \text{LSPD}(\mathbf{y}, F)| \overset{a.s.}{\to} 0$ as $n \to \infty$. So, combining these results, we have $\sup_{\mathbf{x}\in B_n \setminus A_n} |\text{LSPD}_h(\mathbf{x}, F_n) - \text{LSPD}_h(\mathbf{x}, F)| \overset{a.s}{\to} 0$ as $n \to \infty$.

(iii) Now, consider the region outside B_n (i.e., B_n^c). First note that

$$\sup_{\mathbf{x}\in B_n^c} |\text{LSPD}_h(\mathbf{x}, F_n) - \text{LSPD}(\mathbf{x}, F)| \leq \sup_{\mathbf{x}\in B_n^c} \frac{1}{nh^d}\sum_{i=1}^n K(h^{-1}(\mathbf{x}-\mathbf{X}_i)) + \sup_{\mathbf{x}\in B_n^c} h^{-d}E\left[K(h^{-1}(\mathbf{x}-\mathbf{X}))\right].$$

We will show that both of these terms become sufficiently small as $n \to \infty$.

Fix any $\epsilon > 0$. We can choose two constants M_1 and M_2 such that $P(\|\mathbf{X}\| \geq M_1) \leq h^d\epsilon/2K(\mathbf{0})$ and $K(\mathbf{t}) \leq h^d\epsilon/2$ when $\|\mathbf{t}\| \geq M_2$. Now, one can check that

$$h^{-d}E\left[K(h^{-1}(\mathbf{x}-\mathbf{X}))\right] \leq h^{-d}E\left[K(h^{-1}(\mathbf{x}-\mathbf{X}))I(\|\mathbf{X}\| \leq M_1)\right] + h^{-d}K(\mathbf{0})P(\|\mathbf{X}\| > M_1).$$

Note that if $\mathbf{x} \in B_n^c$ and $\|\mathbf{X}\| \leq M_1$, $h^{-1}\|\mathbf{x}-\mathbf{X}\| \geq h^{-1}|\sqrt{d}n - M_1|$. Now, choose n large enough such that $|\sqrt{d}n - M_1| \geq M_2 h$, and this implies $K(h^{-1}(\mathbf{x}-\mathbf{X})) \leq h^d\epsilon/2$. So, we get

$$\begin{aligned} h^{-d}E\left[K(h^{-1}(\mathbf{x}-\mathbf{X}))\right] &\leq \epsilon/2 + h^{-d}K(\mathbf{0})P(\|\mathbf{X}\| > M_1) \leq \epsilon, \text{ and} \\ \frac{1}{nh^d}\sum_{i=1}^{n} K(h^{-1}(\mathbf{x}-\mathbf{X}_i)) &\leq \epsilon/2 + h^{-d}K(\mathbf{0})\frac{1}{n}\sum_{i=1}^{n} I(\|\mathbf{X}_i\| > M_1) \\ &\leq \epsilon + h^{-d}K(\mathbf{0})\left|\frac{1}{n}\sum_{i=1}^{n} I(\|\mathbf{X}_i\| > M_1) - P(\|\mathbf{X}\| > M_1)\right|. \end{aligned}$$

The Glivenko-Cantelli theorem implies that the last term on the right hand side converges to 0 as $n \to \infty$. So, we have $\sup_{\mathbf{X}\in B_n^c} |\text{LSPD}_h(\mathbf{x}, F_n) - \text{LSPD}_h(\mathbf{x}, F)| \overset{a.s}{\to} 0$ as $n \to \infty$.

Combining (i), (ii) and (iii), we now have $\sup_{\mathbf{X}} |\text{LSPD}_h(\mathbf{x}, F_n) - \text{LSPD}_h(\mathbf{x}, F)| \overset{a.s.}{\to} 0$ for any $h \in (0, 1]$.

For any fixed $h > 1$, this a.s. convergence can be proved in a similar way. In that case, recall that the definition of LSPD does not involve the h^d term in the denominator. □

Remark 2: Following the proof of Theorem 3, it is easy to check that the a.s. convergence holds when h diverges to infinity at any rate with n.

Remark 3: The result continues to hold when $h \to 0$ as well. However, for the a.s. convergence in part (i), (more specifically, to use the Borel-Cantelli lemma), we require $nh^{2d}/\log n \to \infty$ as $n \to \infty$. In part (iii), we need M_1 and M_2 to vary with n. Assume the first moment of f to be finite, and $\int \|\mathbf{t}\|K(\mathbf{t})d\mathbf{t} < \infty$ (which implies $\|\mathbf{t}\|K(\mathbf{t}) \to 0$ as $\|\mathbf{t}\| \to \infty$). Also assume that $nh^{2d}/\log n \to \infty$ as $n \to \infty$. We can now choose $M_1 = M_2 = \sqrt{n}$ to ensure that both $P(\|\mathbf{X}\| \geq M_1) \leq h^d\epsilon/2K(\mathbf{0})$ and $K(\mathbf{t}) \leq h^d\epsilon/2$ for $\|\mathbf{t}\| \geq M_2$ hold for sufficiently large n.

Proof of Theorem 4 (a) : Consider two independent random vectors $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})^T \sim F_j$ and $\mathbf{X}_1 = (X_1^{(1)}, \ldots, X_1^{(d)})^T \sim F_j$, where $1 \leq j \leq J$. It follows from (C1) and (C2) that $\|\mathbf{X} - \mathbf{X}_1\|/\sqrt{d} \xrightarrow{a.s.} \sqrt{2\sigma_j^2}$ as $d \to \infty$. So, for almost every realization $\mathbf{x}$ of $\mathbf{X} \sim F_j$,

$$\|\mathbf{x} - \mathbf{X}_1\|/\sqrt{d} \xrightarrow{a.s.} \sqrt{2\sigma_i^2} \text{ as } d \to \infty. \tag{9}$$

Next, consider two independent random vectors $\mathbf{X} \sim F_j$ and $\mathbf{X}_1 \sim F_i$ for $1 \leq i \neq j \leq J$. Using (C1) and (C2), we get $\|\mathbf{X} - \mathbf{X}_1\|/\sqrt{d} \xrightarrow{a.s.} \sqrt{\sigma_j^2 + \sigma_i^2 + \nu_{ji}}$ as $d \to \infty$. Consequently, for almost every realization $\mathbf{x}$ of $\mathbf{X} \sim F_j$

$$\|\mathbf{x} - \mathbf{X}_1\|/\sqrt{d} \xrightarrow{a.s.} \sqrt{\sigma_j^2 + \sigma_i^2 + \nu_{ji}} \text{ as } d \to \infty, \tag{10}$$

Let us next consider $\langle \mathbf{x} - \mathbf{X}_1, \mathbf{x} - \mathbf{X}_2 \rangle$, where $\mathbf{X} \sim F_j$, $\mathbf{X}_1, \mathbf{X}_2 \sim F_i$ are independent random vectors, and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^d$. Therefore, for almost every realization $\mathbf{x}$ of $\mathbf{X}$, arguments similar to those used in (8) and (9) yield

$$\frac{\langle \mathbf{x} - \mathbf{X}_1, \mathbf{x} - \mathbf{X}_2 \rangle}{d} \xrightarrow{a.s.} \sigma_j^2 \text{ as } d \to \infty \text{ if } 1 \leq i = j \leq J, \text{ and} \tag{11}$$

$$\frac{\langle \mathbf{x} - \mathbf{X}_1, \mathbf{x} - \mathbf{X}_2 \rangle}{d} \xrightarrow{a.s.} \sigma_j^2 + \nu_{ji} \text{ as } d \to \infty \text{ if } 1 \leq i \neq j \leq J. \tag{12}$$

Observe now that $\|E_{F_j}[u(\mathbf{x} - \mathbf{X})]\|^2 = \langle E_{F_j}[u(\mathbf{x} - \mathbf{X}_1)], E_{F_j}[u(\mathbf{x} - \mathbf{X}_2)] \rangle = E_{F_j}\{\langle u(\mathbf{x} - \mathbf{X}_1), u(\mathbf{x} - \mathbf{X}_2) \rangle\}$, where $\mathbf{X}_1, \mathbf{X}_2 \sim F_j$ are independent random vectors for $1 \leq j \leq J$.

Since here we are dealing with expectations of random vectors with bounded norm, a simple application of Dominated Convergence Theorem implies that for almost every realization $\mathbf{x}$ of $\mathbf{X} \sim F_j$ $(1 \leq i \leq J)$, as $d \to \infty$,

$$\text{SPD}(\mathbf{x}, F_j) \xrightarrow{a.s.} 1 - \sqrt{\frac{1}{2}} \text{ and } \text{SPD}(\mathbf{x}, F_i) \xrightarrow{a.s.} 1 - \sqrt{\frac{\sigma_j^2 + \nu_{ji}}{\sigma_j^2 + \sigma_i^2 + \nu_{ji}}} \text{ for } i \neq j. \tag{13}$$

Therefore, for $\mathbf{X} \sim F_j$, we get $z(\mathbf{X}) = (\text{SPD}(\mathbf{X}, F_1), \ldots, \text{SPD}(\mathbf{X}, F_J))^T \xrightarrow{a.s.} \mathbf{c}_j$, as $d \to \infty$. □

Proof of Theorem 4 (b) : Recall that for $h > 1$, $\text{LSPD}_h(\mathbf{x}, F) = E_F[h^d K_h(\mathbf{t})] - \|E_F[h^d K_h(\mathbf{t}) u(\mathbf{t})]\|$, and since we have assumed $\mathbf{X}$s to be standardized, here we have $h^d K_h(\mathbf{t}) = K((\mathbf{x} - \mathbf{X})/h) = g(\|\mathbf{x} - \mathbf{X}\|/h)$. Let $\mathbf{X} \sim F$ and $\mathbf{X}_i \sim F_i$ where $1 \leq i \leq J$. Then, using (8)

and (9) above, and the continuity of g, for almost every realization $\mathbf{x}$ of $\mathbf{X} \sim F_j$, one gets

$$g\left(\frac{\|\mathbf{x}-\mathbf{X}_i\|}{\sqrt{d}}\frac{\sqrt{d}}{h}\right) \stackrel{a.s.}{\to} g(0) \text{ or } g(e_{ji}A),$$

depending on whether $\sqrt{d}/h \to 0$ or A, for almost every realization $\mathbf{x}$ of $\mathbf{X} \sim F_j$. The proof now follows from a simple application of Dominated Convergence Theorem. □

Proof of Theorem 4 (c) : Since $g(s) \to 0$ as $s \to \infty$, using the same argument as used in the proof of Theorem 3(b), for $\mathbf{X}_i \sim F_i$ and almost every realization $\mathbf{x}$ of $\mathbf{X} \sim F_j$, we have

$$g\left(\frac{\|\mathbf{x}-\mathbf{X}_i\|}{\sqrt{d}}\frac{\sqrt{d}}{h}\right) \stackrel{a.s.}{\to} 0 \text{ as } \sqrt{d}/h \to \infty.$$

The proof now follows from a simple application of Dominated Convergence Theorem. □

Lemma 2 : Recall $\mathbf{c}_j$ and $\mathbf{c}'_j$ for $1 \le j \le J$ defined in Theorem 3 (a) and (b), respectively. For any $1 \le j \ne i \le J$, $\mathbf{c}_j = \mathbf{c}_i$ if and only if $\sigma_j = \sigma_i$ and $\nu_{ji} = \nu_{ij} = 0$. Similarly, $\mathbf{c}'_j = \mathbf{c}'_i$ if and only if $\sigma_j = \sigma_i$ and $\nu_{ji} = \nu_{ij} = 0$.

Proof of Lemma 2 : The 'if' part is easy to check in both cases. So, it is enough to prove the 'only if' part and that too for the case of $J = 2$. Note that if $\mathbf{c}_1 = (c_{11}, c_{12})^T$ and $\mathbf{c}_2 = (c_{21}, c_{22})^T$ are equal, we have

$$\frac{\sigma_1^2+\nu_{12}}{\sigma_1^2+\sigma_2^2+\nu_{12}} = 1/2 \text{ and } \frac{\sigma_2^2+\nu_{12}}{\sigma_1^2+\sigma_2^2+\nu_{12}} = 1/2.$$

These two equations hold simultaneously only if $\sigma_1^2 = \sigma_2^2$ and $\nu_{12}(= \nu_{21}) = 0$.

Now consider the case $\mathbf{c}'_1 = \mathbf{c}'_2$. Recall that $c'_{11} = g(A\sqrt{2}\sigma_1)c_{11}$, $c'_{22} = g(A\sqrt{2}\sigma_2)c_{22}$, $c'_{12} = g(A\sqrt{\sigma_1^2+\sigma_2^2+\nu_{12}})c_{12}$ and $= c'_{21} = g(A\sqrt{\sigma_2^2+\sigma_1^2+\nu_{21}})c_{21}$. If possible, assume that $\sigma_1 > \sigma_2$. This implies that $A\sqrt{\sigma_1^2+\sigma_2^2+\nu_{12}} > A\sqrt{2}\sigma_1$ and hence

$$g(A\sqrt{2}\sigma_1) > g(A\sqrt{\sigma_1^2+\sigma_2^2+\nu_{12}}) \quad \text{(since } g \text{ is monotonically decreasing).} \tag{14}$$

Also, if $\sigma_1 > \sigma_2$, we must have

$$1/2 < \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} < \frac{\sigma_1^2+\nu_{12}}{\sigma_1^2+\sigma_2^2+\nu_{12}} < 1 \Leftrightarrow 1-\sqrt{1/2} > 1-\sqrt{\frac{\sigma_1^2+\nu_{12}}{\sigma_1^2+\sigma_2^2+\nu_{12}}}. \tag{15}$$

Combining (13) and (14), we have $c'_{11} > c'_{21}$, and this implies $\mathbf{c}'_1 \neq \mathbf{c}'_2$. Similarly, if $\sigma_1 < \sigma_2$, we get $c'_{12} > c'_{22}$ and hence $\mathbf{c}'_1 \neq \mathbf{c}'_2$. Again, if $\sigma_1 = \sigma_2$ but $\nu_{12} = \nu_{21} > 0$, similar arguments lead to $\mathbf{c}'_1 \neq \mathbf{c}'_2$. This completes the proof of the lemma. □

References

[1] Agostinelli, C. and Romanazzi, M. (2010) Local depth. *J. Statist. Plan. Inf.*, **141**, 817-830.

[2] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Mack, D. and Leine, A. J. (1999) Broad pattern of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci., USA*, **96**, 6745-6750.

[3] Chaudhuri, P., Ghosh, A. K. and Oja, H. (2009) Classification based on hybridization of parametric and nonparametric classifiers. *IEEE Trans. Pattern Anal. Machine Intell.,* **31**, 1153-1164.

[4] Chen, Y., Dang, X., Peng, H. and Bart Jr, H. L. (2009) Outlier detection with the kernelized spatial depth function. *IEEE Trans. Pattern Anal. Machine Intell.,* **31**, 288-305.

[5] de Jong, R. M. (1995) Laws of large numbers for dependent heterogeneous processes. *Econ. Theory*, **11**, 347-358.

[6] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A. (2011) e1071 : Misc functions of the department of statistics (e1071), TU Wien. R package version 1.5-27. *http://CRAN.R-project.org/package=e1071*

[7] Dutta, S. and Ghosh, A. K. (2012) On robust classification using projection depth. *Ann. Inst. Statist. Math.*, **64**, 657-676.

[8] Dzeroski, S. and Zenko, B. (2004) Is combining classifiers better than selecting the best one ? *Mach. Learn.*, **54**, 255-273.

[9] Friedman, J. (1996) Another approach to polychotomous classification. Technical Report, Dept. of Statistics, Stanford University. Available at *http://old.cba.ua.edu/∼mhardin/poly.pdf*.

[10] Gao, Y. (2003) Data depth based on spatial rank. *Statist. Prob. Lett.*, **65**, 217-225.

[11] Ghosh, A. K. and Chaudhuri, P. (2005) On data depth and distribution free disciminant analysis using separating surfaces. *Bernoulli*, **11**, 1-27.

[12] Ghosh, A. K. and Chaudhuri, P. (2005) On maximum depth and related classifiers. *Scand. J. Statist.*, **32**, 328-350.

[13] Ghosh, A. K., Chaudhuri, P. and Murthy, C. A. (2005) On visualization and aggregation of nearest neighbor classifiers. *IEEE Trans. Pattern Anal. Machine Intell.*, **27**, 1592-1602.

[14] Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2006) Classification using kernel density estimates : multi-scale analysis and visualization. *Technometrics*, **48**, 120-132.

[15] Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension low sample size data. *J. Royal Statist. Soc. Ser. B*, **67**, 427-444.

[16] Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models.* Chapman & Hall, New York.

[17] Hastie, T., Tibshirani, R. and Friedman, J. (2009) *Elements of Statistical Learning Theory.* Wiley, New York.

[18] Hoberg, R. and Mosler, K. (2006) Data analysis and classification with the zonoid depth. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **72**, 45-59.

[19] Hu, Y., Wang, Y., Wu, Y., Li, Q. and Hou, C. (2011) Generalized Mahalanobis depth in the reproducing kernel Hilbert space. *Statist. Papers*, **52**, 511-522.

[20] Jornsten, R. (2004) Clustering and classification based on the L_1 data depth. *J. Mult. Anal.*, **90**, 67-89.

[21] Jung, S. and Marron, J. S. (2009) PCA consistency in high dimension, low sample size context. *Ann. Statist.*, **37**, 4104-4130.

[22] Kittler, J., Hatef, M., Duin, R. and Matas, J. (1998) On combining classifiers. *IEEE Trans. Pattern Anal. Machine Intell.,* **20**, 226-239.

[23] Lange, T., Mosler, K., and Mozharovskyi, P. (2014) Fast nonparametric classification based on data depth. *Statist. Papers*, **55**, 49-69.

[24] Levina, E. and Bickel, P. J. (2005) Maximum likelihood estimation of intrinsic dimension. In *Adv. Neural Info. Process. Sys.*, MIT Press, Cambridge, MA. Vol. **17**, pp. 777-784.

[25] Li, J., Cuesta-Albertos, J. A. and Liu, R. (2012) Nonparametric classification procedures based on DD-plot. *J. Amer. Statist. Assoc.*, **107**, 737-753.

[26] Li, R.-Z., Fang, K.-T. and Zhu, L.-X. (1997) Some Q-Q probabillity plots to test spherical and elliptic symmetry. *J. Comput. Graph. Statist.*, **6**, 435-450.

[27] Liaw, A. and Wiener, M. (2002) Classification and regression by random forest. *R News*, **2**, 18-22.

[28] Lim, T.-S., Loh, W.-Y. and Shih, Y.-S. (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.*, **40**, 203-228.

[29] Liu, R., Parelius, J. and Singh, K. (1999) Multivariate analysis of data depth: descriptive statistics and inference. *Ann. Statist.*, **27**, 783-858.

[30] Lok, W. S. and Lee, S. M. S. (2011) A new statistical depth function with applications to multimodal data. *J. Nonparam. Statist.*, **23**, 617-631.

[31] Marron, J. S., Todd, M. J. and Ahn, J. (2007) Distance weighted discrimination. *J. Amer. Statist. Assoc.*, **102**, 1267-1271.

[32] Paindaveine, D. and Van Bever, G. (2013) From depth to local depth : a focus on centrality. *J. Amer. Statist. Assoc.*, **105**, 1105-1119.

[33] Paindaveine, D. and Van Bever, G. (2015) Nonparametrically consistent depth-based classifiers. *Bernoulli*, **21**, 62-82.

[34] Ramsey, J. O. and Silverman, B. W. (2005) *Functional Data Analysis.* Springer, New York.

[35] Ripley, B. (2011) tree: Classification and regression trees. R package version 1.0-29. *http://CRAN.R-project.org/package=tree*

[36] Royston, J. P. (1983) Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *J. Royal Statist. Soc. Ser. C*, **32**, 121-133.

[37] Serfling, R. (2002) A depth function and a scale curve based on spatial quantiles. *In Statistics and Data Analysis based on L_1-Norm and Related Methods* (Y. Dodge ed.), Birkhaeuser, 25-38.

[38] Vardi, Y. and Zhang, C. H. (2000) The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, **97**, 1423-1426.

[39] Xia, C., Lin, L. and Yang, G. (2008) An extended projection data depth and its applications to discrimination. *Comm. Statist. - Theory and Methods*, **37**, 2276-2290.

[40] Yee, T. W. and Wild, C. J. (1996) Vector generalized additive models. *J. Roy. Statist. Soc. Ser. B*, **58**, 481-493.

[41] Yurinskii, V. V. (1976) Exponential inequalities for sums of random vectors. *J. Mult. Anal.*, **6**, 473-499.

[42] Zuo, Y. and Serfling, R. (2000) General notions of statistical depth function. *Ann. Statist.*, **28**, 461-482.