Statistica Sinica Preprint No: SS-2023-0014	
Title	Mode-Based Classifier: A Robust and Flexible Discriminant Analysis for High-Dimensional Data
Manuscript ID	SS-2023-0014
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0014
Complete List of Authors	Wei Xiong, Wolfgang Karl Härdle, Jianrong Wang, Keming Yu and Maozai Tian
Corresponding Authors	Wei Xiong
E-mails	xwhehe.26@163.com
Notice: Accepted version subject to English editing.	

Statistica Sinica

MODE-BASED CLASSIFIER: A ROBUST AND FLEXIBLE DISCRIMINANT ANALYSIS FOR HIGH-DIMENSIONAL DATA

Wei Xiong, Wolfgang Karl Härdle, Jianrong Wang, Keming Yu and Maozai Tian

University of International Business and Economics, Humboldt-Universitat zu Berlin, Brunel University and Renmin University of China

Abstract: High-dimensional classification is both challenging and of interest in numerous applications. Componentwise distance-based classifiers, which utilize partial information with known categories, such as mean, median and quantiles, provide a convenient way. However, when the input features are heavy-tailed or contain outliers, performance of the centroid classifier can be poor. Beyond that, it frequently occurs that a population consists of two or more subpopulations, the mean, median and quantiles in this scenario fail to capture such a structure that can be instead preserved by mode, which is an appealing measure of considerable significance but might be neglected. This paper thus introduces and investigates componentwise mode-based classifiers that can reveal important structures missed by existing distance-based classifiers. We explore several strategies for defining the family of mode-based classifiers, including the unimodal classifiers, the multimodal classifier and the quantile-mode classifier. The unimodal classifiers are proposed based on componentwise unimodal distance and kernel mode estimation, and the multimodal classifier is constructed by identifying all the local modes of a distribution according to a novel introduced algorithm. We establish the asymptotic properties of these methods and demonstrate through simulation studies and three real datasets that the mode-based classifiers compare favorably to the current state-of-art methods.

Key words and phrases: componentwise modal distance, multimodal classifier, multimodality, quantile-

mode, unimodal classifier.

1. Introduction

In this work, we focus on the problem of classification for high-dimensional data, where the task is to assign a new observation to one class out of a finite collection of alternatives. Classification arises frequently from bioinformatics, computer vision, natural language processing, and a broad range of other fields, is an indispensable part of artificial intelligence. Most of the conventional methods depend on sufficient and balanced samples to make classifiers more efficient. More recently, neural networks and ensemble methods like Boosting and Random Forests have been proposed to solve different problems originated from different contexts. High-dimensional data poses significant challenges to many existing classification techniques, whose performance might be poor but computationally heavier due to the "curse of dimensionality".

The problem caused by high dimensions can be solved moderately by using distance-based classifiers which utilize a small portion of information of the population distributions with known categories. Distance-based classification methods take the distance between the new observed value and the core of the corresponding distribution as an important basis for classification. They distinguish themselves from other classifiers in one important respect: distance-based classifiers that allow the addition of new classes to existing classes without any cost is an approach to deal with real-life large scale datasets which are open-ended and dynamic (Mensink et al., 2013). Standard distance-based classifiers are the centroid-based classifiers, nearest class mean classifier (Webb, 2002), nearest-neighbor methods (Cover and Hart, 1967), support vector machines (Cortes and Vapnik, 1995) and so on. The performance of these classifiers critically relies on the applied distance metrics and the components of data vector, the predictive power can be very poor if some of the components of data vector suffer from high variability, which becomes more significant as the number

of features diverges. To alleviate this problem, componentwise distance-based classifiers that are established based on a sum of componentwise distances can be good alternatives, they adapt well to high-dimensional data without the frequently employed feature screening steps, but research on them is relatively sparse. Hall et al. (2009) introduce a componentwise median-based classifier, which utilizes the L_1-distance in the training set and performs well in high dimensions. As an extension of the median-based classifier, Hennig and Viroli (2016) propose a family of componentwise quantile-based classifiers by defining a componentwise quantile distance, and develop the optimal quantile classifier, in which the optimal quantile is chosen in the training set by maximizing the correct classification probability. Although the quantile classifiers are effective for discriminating high-dimensional data, their performance is limited to assigning each predictor the same importance. Accordingly, Lai and McLeod (2020) appoint weights to the componentwise distances and produce ensemble quantile classifier. Another closely related method is the directional quantile classifier that is built upon directional quantiles to account for possible interdependence among variables (Farcomeni et al., 2022).

Besides the above-mentioned statistics, i.e., mean, median and quantiles, the mode is demonstrated to be the most natural metric for describing central tendency in positively skewed data, and is the only measure that can be used for nominal scale data. This motivates us to focus on the componentwise mode-based classifiers. Why would we ever use mode-based classifiers in favor conventional componentwise distance-based classifiers? The answer, at a high level, is that mode can reveal structure that is missed by other statistics, and may cover more distributional information relevant for classification. Figure 1 and Table 1 give a definitive illustration of this point: we can see that (a) when the input features are heavy-tailed or contain outliers, the arithmetic mean is not applicable for representing the central tendency, leading to the poor performance of the componentwise centroid classifier especially for high-dimensional heterogeneous data. (b) It sometimes happens for two

distributions that the mean and median are identical or nearly identical, while the shapes are quite different, in this sense, performance of the componentwise centroid and median classifiers are similar, but might be unsatisfactory, since the commonly used measures of central tendency reflect limited information on other important features of underlying distribution, such as wiggles. (c) Quantile-based classifier is always powerful in the above two cases in view of the fact that quantiles can better characterize a distribution. However practically, it frequently occurs that a population consists of two or more subpopulations, the mean, median, and quantile in this scenario fail to capture such a structure, resulting in the poor prediction powers. These difficulties make mode attractive, the componentwise mode-based classifier is an improvement in these regards. By comparing the train and test error rates in Table 1, all classifiers except for the proposed one suffer from overfitting is observed. Therefore, the componentwise mode-based classifiers become natural to consider in this context, they are potentially useful but much neglected tool that can be employed to complement and advance the existing componentwise distance-based classifiers. To the best of our knowledge, relatively little has been done for the construction of mode-based classifiers. It is the intention of this work to fill this gap.

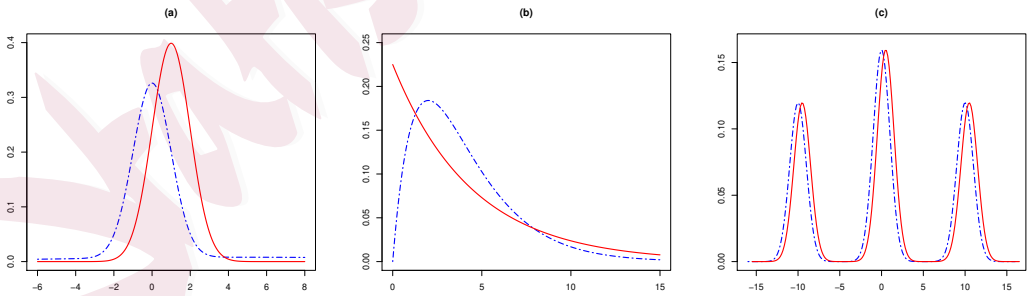

Figure 1: Examples of mode-based classifier versus centroid, median and quantile-based classifiers. Each p vectors are generated from two different populations. Panel (a): $X_j \sim 0.8N(0,1) + 0.2N(5,10^2)$, $Y_j \sim N(1,1)$; (b) $X_j \sim \chi^2(4)$, $Y_j \sim \exp(9/40)$; (c) $X_j \sim 0.3N(-10,1) + 0.4N(0,1) + 0.3N(10,1)$, $Y_j \sim X_j + 0.5$, for $j = 1, \ldots, p$.

With this motivation in mind, this work develops the family of the componentwise

Table 1: Classification error rates (%) over 100 replications. Proposed denotes the mode-based classifier.

		$p = 10$				$p = 100$			
Case	Method	Centroid	Median	Quantile	Proposed	Centroid	Median	Quantile	Proposed
(a)	Train error	39.01±3.12	14.81±1.70	13.29±1.33	11.59±1.41	20.71±1.41	0.13±0.14	0.00±0.00	0.00±0.00
	Test error	43.00±3.02	16.2±1.61	15.91±1.85	12.36±1.50	35.46±1.46	0.15±0.16	0.27±0.29	0.00±0.00
(b)	Train error	40.48±1.60	41.47±2.18	23.63±1.55	26.32±1.86	20.48±1.37	27.99±1.76	3.19±0.61	5.52±0.74
	Test error	44.74±1.57	45.37±1.52	26.60±1.94	30.77±1.81	29.25±1.76	43.74±1.89	6.22±1.59	9.19±1.48
(c)	Train error	43.51±1.67	38.55±2.07	36.01±1.63	24.53±1.71	26.68±1.46	16.92±1.55	11.92±1.09	2.99±0.68
	Test error	49.11±1.95	41.67±1.70	42.28±2.23	27.43±1.50	42.74±2.12	25.20±1.40	20.24±2.01	5.69±0.10

mode-based classifiers by defining appropriate distances for modes. The mode-based classifiers have several unusual properties, making the investigation of their properties more challenging. For instance, the theoretical mode of the sample is not usually equivalent to the mode of the population from which the data were drawn; the fact that the population mode is not necessarily unique to a given distribution, whereas the mean and median always represent a single value, making the mode-based classifier differ largely from the centroid and median classifiers. The loss function whose expectation is minimized at the mode is not uniquely defined, leading to the difficulty of choosing an appropriate distance metric. To address these challenges, this work provides a thorough exploration of the mode-based classifiers and investigates their properties. We not only focus much attention on unimodal classifiers, but develop multimodal classifiers. In particular, our contributions are as follows:

1. For unimodal populations, we introduce the family of unimodal classifiers based on unimodal distance in combination with kernel mode estimation. We propose a method for defining the optimal unimodal classifier, and prove that it enjoys consistency.

2. To significantly reduce the computational cost or improve the classification accuracy, alternative approaches for defining unimodal classifiers are also considered. These include the naive unimodal classifier and the hybrid quantile-mode classifier. The naive unimodal classifier is motivated by the relationship between mode, mean and median, and the quantile-

mode classifier is inspired by the nice properties of asymmetric Laplace distribution (ALD).

3. In some cases, there can be clear evidence of multimodality of a population. To address this, we first propose a novel mode detection algorithm to accurately identify all the local modes of a predictor and then formulate the multimodal classifiers, which outperform all the existing classifiers by a large margin in the presence of multimodality.

4. To allow mode-based classifiers for large-scale datasets, we introduce in addition a mode-diff (MD) filter, a feature screening technique, for speeding up classification.

5. Under the assumption that distributions for alternative populations differ by at most a location-shift, asymptotic theoretical properties of the mode-based classifier are derived. Though this assumption may be practically unrealistic, various numerical examples and real data analysis validate that the mode-based classifiers show competitive performance and always compare favorably to the current state-of-the-art classifiers in all scenarios including the distributions differing by shape and not just by location. We draw enlightening comparisons to current classifiers.

Generally speaking, the mode-based classifiers are robust and powerful for discriminating high-dimensional data especially with heavy-tailed, skewed or multimodal inputs. The rest of the paper is organized as follows. We define the family of mode-based classifiers in Section 2, where several variants are explored and an algorithm for detecting the multimodality is provided. In Section 3, we discuss the implementation issues associated with the mode-based classifiers. Section 4 respectively investigates the theoretical properties of unimodal classifiers for fixed p and $p \to \infty$. We present extensive simulation results in Section 5. Section 6 illustrates the usefulness of the proposed methods through three real datasets. Finally, we conclude this paper with some concluding remarks. All technical proofs and additional simulation results are presented in the Supplementary Materials.

2. Mode-based Classifiers

2.1 Componentwise distance-based classifiers

We consider constructing a mode-based distance discriminant rule for classifying new observations to one of the two classes. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be populations characterized by random variables X and Y defined on $\mathbb{R}^p$, $\mathcal{F}_1(\cdot)$ and $\mathcal{F}_2(\cdot)$ be probability distribution functions for each population. Define two sets $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ and $\mathcal{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ of p-variate data, given a new observation $\boldsymbol{z} = (z_1, \ldots, z_p)$, componentwise distance-based classifiers (Tibshirani et al., 2003; Jörnsten, 2004; Hall et al., 2009; Hennig and Viroli, 2016) assign $\boldsymbol{z}$ to the class that is closest, specifically, assign $\boldsymbol{z}$ to $\mathcal{C}_1$ if

$$\sum_{j=1}^{p} \{d(z_j, \theta_{\mathcal{Y}_j}) - d(z_j, \theta_{\mathcal{X}_j})\} > 0, \tag{2.1}$$

where $\boldsymbol{\theta}_{\mathcal{X}} = (\theta_{\mathcal{X}_1}, \ldots, \theta_{\mathcal{X}_p})$ and $\boldsymbol{\theta}_{\mathcal{Y}} = (\theta_{\mathcal{Y}_1}, \ldots, \theta_{\mathcal{Y}_p})$ are p-variate population statistics for $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively, $d(z_j, \theta_{\mathcal{X}_j})$ denotes a specific measure of distance from the jth component of $\boldsymbol{z}$ to the statistic of the set $\mathcal{X}_j = \{X_{1j}, \ldots, X_{mj}\}$. The L_2-distance in (2.1) is the most commonly used distance metric, based on which the centroid classifier (Tibshirani et al., 2003) is formulated by taking $\boldsymbol{\theta}_{\mathcal{X}} = (\bar{\mathcal{X}}_1, \ldots, \bar{\mathcal{X}}_p)$ and $\boldsymbol{\theta}_{\mathcal{Y}} = (\bar{\mathcal{Y}}_1, \ldots, \bar{\mathcal{Y}}_p)$, where $\bar{\mathcal{X}}_j$ denotes the jth component of the sample mean, i.e., $\bar{\mathcal{X}}_j = m^{-1} \sum_k X_{kj}$. If $\boldsymbol{\theta}_{\mathcal{X}}$ and $\boldsymbol{\theta}_{\mathcal{Y}}$ are taken to be the componentwise sample medians of $\mathcal{X}$ and $\mathcal{Y}$, respectively, $d(\cdot)$ is considered to be the L_1-distance in (2.1), the median-based classifier introduced by Hall et al. (2009) is obtained. In a similar fashion, given a quantile level $\tau \in (0, 1)$, if $\boldsymbol{\theta}_{\mathcal{X}}$ and $\boldsymbol{\theta}_{\mathcal{Y}}$ are taken to be the componentwise sample τ-quantiles of $\mathcal{X}$ and $\mathcal{Y}$, and $d(\cdot)$ is taken to be the quantile-loss function, that is, $d(z_j, \theta_{\mathcal{X}_j}) = \rho_\tau(z_j - q_{\mathcal{X}_j}(\tau))$, where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$, $I(\cdot)$ is an indicator function and $q_{\mathcal{X}_j}(\tau)$ is the empirical τ-quantile of $\mathcal{X}_j$, then the quantile-based classifier defined by Hennig and Viroli (2016) is achieved. The reason why L_2 metric is used for the centroid, L_1 metric for the median and the quantile-loss for the quantile is that the mean is the statistic that minimizes the sum of L_2 distances from points to their

centroid, the median minimizes the corresponding L_1 distance and the τ-quantile minimizes the quantile loss function at given τ. Inspired by the above facts, the mode-based classifier can be defined by proper selection of the distance metric, for which the decision rule is,

$$\sum_{j=1}^{p}\{d(z_j, \text{mode}(\mathcal{Y}_j)) - d(z_j, \text{mode}(\mathcal{X}_j))\} > 0.$$

2.2 Unimodal classifiers

We begin with the unimodal classifiers for two unimodal populations. By definition, the mode δ of a univariate random variable U is the only value at which the probability mass (density) function takes its maximum value. Literature focuses primarily on mode estimation under nonparametric kernel density estimation (KDE), see Parzen (1962); Eddy (1980); Birgé (1997); Meyer (2001) among others. Precisely, let $F_U(u)$ be the probability distribution function of U with density $f_U(u)$, and $L(U;\cdot)$ be the step-loss function (Manski, 1991), $L(U;\delta) = I\big(|U-\delta| \geq \sigma\big)$, where $\sigma > 0$ is a bandwidth. If $f_U(u)$ is symmetric about δ or if δ is the middle value of the interval of length 2σ that captures the most probability under $F_U(u)$, then $\hat{\delta} = \arg\min_\delta E\{L(U;\delta)\}$ is the mode of U. If $f_U(u)$ is highly skewed or it has multiple local maxima, the loss function of the form (2.2) is usually suggested,

$$L(U;\delta) = 1 - \gamma K\left(\frac{U-\delta}{\sigma}\right), \tag{2.2}$$

where $K(u)$ denotes a smooth kernel function and $\gamma = 1/K(0) > 0$ is a scaling constant. If $f_U(u)$ is strictly unimodal, the minimizer of (2.2) approaches mode(U) as σ goes to zero. Many smooth kernels are available here, such as Triangle kernel, Epanechnikov kernel, Gaussian kernel, and so on. This work mainly focuses on Gaussian Kernel, that is $K(u) = \varphi(u)$, where $\varphi(\cdot)$ is the standard normal density function. The reason why we take Gaussian kernel in this work is twofold: (i) the Gaussian kernel offers the advantage of generating a loss function that has both the mode and the mean as minimizers in limiting case (Kemp and Silva, 2012); (ii) the number of modes in a Gaussian kernel density estimate is nonincreasing

as σ increases (Minnotte, 1997). These two properties are not shared with other kernels.

For observations $u_1, \ldots, u_n$ independent and identically distributed (i.i.d.) with U, the empirical mode of U is the value δ that minimizes the sample analog of the expectation of (2.2), this is equivalent to maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_n}K\left(\frac{u_i-\delta}{\sigma_n}\right), \tag{2.3}$$

which is an estimate of the density of u_i at δ, σ_n is a bandwidth related to sample size n such that $\sigma_n \to 0$ as $n \to \infty$. We then utilize equations (2.2) and (2.3) to define the unimodal classifier. Denote by F_{rj} the marginal distributions of $\mathcal{F}_r$ and δ_{rj} the unique mode of F_{rj}, for $r = 1, 2$ and $j = 1, \ldots, p$. Let $\boldsymbol{\delta}_r = (\delta_{r1}, \ldots, \delta_{rp})$, and

$$\Psi_{rj}(z_j, \sigma_{rj}) = K\left(\frac{z_j-\delta_{rj}}{\sigma_{rj}}\right), \quad r = 1, 2, \ j = 1, \ldots, p,$$

where σ_{rj} is the bandwidth specified for estimating δ_{rj}. Define the *componentwise unimodal distance* as $d(z_j, \text{mode}(F_{rj})) := 1 - \gamma\Psi_{rj}(z_j, \sigma_{rj})$. Given two data sets $\mathcal{X}$ and $\mathcal{Y}$ each from populations $\mathcal{C}_1$ and $\mathcal{C}_2$, and a new observation $\boldsymbol{z} = (z_1, \ldots, z_p) \in \mathbb{R}^p$, the unimodal classifier assigns $\boldsymbol{z}$ to $\mathcal{C}_1$, if

$$\lambda(\boldsymbol{z}, \boldsymbol{\sigma}) = \sum_{j=1}^{p}\{\Psi_{1j}(z_j, \sigma_{1j}) - \Psi_{2j}(z_j, \sigma_{2j})\} > 0, \tag{2.4}$$

otherwise to $\mathcal{C}_2$, where $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ and $\boldsymbol{\sigma}_r = (\sigma_{r1}, \ldots, \sigma_{rp})$ for $r = 1, 2$.

Given classes $\mathcal{C}_1$ and $\mathcal{C}_2$ with prior probabilities π_1 and π_2, the probability of the correct classification for the unimodal classifier based on true modes $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$, is

$$\Gamma(\boldsymbol{\sigma}) = \pi_1\int I\{\lambda(\boldsymbol{z}, \boldsymbol{\sigma}) > 0\}d\mathcal{F}_1(\boldsymbol{z}) + \pi_2\int I\{\lambda(\boldsymbol{z}, \boldsymbol{\sigma}) \le 0\}d\mathcal{F}_2(\boldsymbol{z}). \tag{2.5}$$

Let $(\boldsymbol{z}_i, C_i), i = 1, \ldots, n$ be $\mathbb{R}^p \times \{1, 2\}$-valued i.i.d random variables, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^T$ satisfies that $\mathcal{F}_1 = P(\boldsymbol{z}_i | C_i = 1)$ and $\mathcal{F}_2 = P(\boldsymbol{z}_i | C_i = 2)$. Let $\hat{\delta}_{rj}$ be the empirical mode for the subsample defined by $C_i = r$, and $\hat{\sigma}_{rjn}$ be the optimal bandwidth for the KDE $\hat{f}_{rj}$ defined by $C_i = r$, $i = 1, \ldots, n$. The *ordinary unimodal classifier* is formulated by assigning

$\boldsymbol{z}$ to $C = 1$ if

$$\lambda_n(\boldsymbol{z}, \hat{\boldsymbol{\sigma}}_n) = \sum_{j=1}^{p} \{\Psi_{1jn}(z_j, \hat{\sigma}_{1jn}) - \Psi_{2jn}(z_j, \hat{\sigma}_{2jn})\} > 0, \tag{2.6}$$

where $\Psi_{rjn}(z, \sigma_n) = K\big\{(z - \delta_{rj})/\sigma_n\big\}$ for $r = 1, 2$, and the observed rate of correct classification is

$$\Gamma_n(\hat{\boldsymbol{\sigma}}_n) = \frac{1}{n}\left[\sum_{i:C_i=1} I\{\lambda_n(\boldsymbol{z}, \hat{\boldsymbol{\sigma}}_n) > 0\} + \sum_{i:C_i=2} I\{\lambda_n(\boldsymbol{z}, \hat{\boldsymbol{\sigma}}_n) \le 0\}\right]. \tag{2.7}$$

If $p = 1$ and further assume bandwidths $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2$ in (2.5), we obtain a simplified formula for calculating the probability of correct classification. When $\boldsymbol{\delta}_1 \le \boldsymbol{\delta}_2$,

$$\Gamma(\boldsymbol{\sigma}) = \pi_1 \mathcal{F}_1\big((\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2)/2\big) + \pi_2\big\{1 - \mathcal{F}_2((\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2)/2)\big\}, \tag{2.8}$$

otherwise,

$$\Gamma(\boldsymbol{\sigma}) = \pi_1\big\{1 - \mathcal{F}_1((\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2)/2)\big\} + \pi_2 \mathcal{F}_2\big((\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2)/2\big). \tag{2.9}$$

Equations (2.8) and (2.9) indicate that in the ideal case the correct population classification rate for unimodal classifier is irrelevant to the bandwidths. The proof of this formula is given in the supplementary materials.

Remark 1. Given two univariate unimodal populations ($p = 1$), without loss of generality, assume $\delta_2 \ge \delta_1$ and let $d = \delta_2 - \delta_1$ be the difference of two modes, a measure for the location difference in distributions. The two populations are becoming more separable with the increase of d. To investigate how the unimodal classifier works, we further assume $\delta_1 = 0$ if necessary after a shift in location and $\sigma_1 = \sigma_2$, then the misclassification rate calculated by (2.9) is $1 - \Gamma(\sigma) = 1 - \pi_1 \mathcal{F}_1(d/2) - \pi_2(1 - \mathcal{F}_2(d/2))$, a function of d. Figure 2 depicts four illustrative examples of how the theoretical classification rate varies across the location difference for two unimodal populations, assuming $\pi_1 = \pi_2 = 0.5$. Clearly, as shown in the second column of Figure 2, with the location shift growing larger, the misclassification rate diminishes to zero rapidly as expected. However, the assumption of $\sigma_1 = \sigma_2$ may be inappropriate for real data, since kernel methods are in principle sensitive to bandwidths.

To account for this crucial effects, we take simulated random samples from each pairs of populations, and denote by $c = \sigma_2/\sigma_1 \in (0, +\infty)$ a ratio between two bandwidths. For each $c \in [0.1, 10]$, we approximate the misclassification rate by a Monte-Carlo simulation average of 100 independent $1 - \hat{\Gamma}(c)$ obtained by the unimodal classifier determined by (2.4) with $m_1 = n_1 = 40,000$ training samples and $m_2 = n_2 = 10,000$ testing samples. Results are depicted in the third column of Figure 2, which conveys several informations: (a) the curve of the misclassification rate in each scenario exhibits strong convexity; (b) $c \leq 3$ always gives a smaller misclassification rate; (c) the location of the minimum point for each curve can be estimated accurately by misclassification rates within the training sample when n is large enough. Consequently, to better implement the unimodal classifier, it is essential to determine an optimal c value to maximize the correct classification rate.

2.3 The empirically optimal unimodal classifier

In the implementation of the proposed unimodal classifier, two issues are critically important. One is the selection of bandwidths, determining which values of bandwidths provide the optimal degree of misclassification rate is hard. The other is the design of algorithm to achieve better classification performance while maintaining computational feasibility.

To address this problem and make it scalable to high-dimensional setting, we first define $\theta = \sigma_{2j}/\sigma_{1j}$ for $j = 1, \ldots, p$, and then introduce an approach to choose the bandwidths in the family of possible unimodal classifiers determined by an optimum θ value based on misclassification rates within the training samples. Take the univariate populations as an example. When $p = 1$, $\sigma_{21} = \theta\sigma_{11}$, the probability of the correct classification (2.5) becomes

$$\Gamma(\sigma_{11}, \theta) = \tag{2.10}$$
$$\pi_1 \int I\left\{K\Big(\frac{z-\delta_1}{\sigma_{11}}\Big) > K\Big(\frac{z-\delta_2}{\theta\sigma_{11}}\Big)\right\} d\mathcal{F}_1(z) + \pi_2 \int I\left\{K\Big(\frac{z-\delta_1}{\sigma_{11}}\Big) \leq K\Big(\frac{z-\delta_2}{\theta\sigma_{11}}\Big)\right\} d\mathcal{F}_2(z).$$

It remains the problem of how to find the two appropriate parameters to maximize the

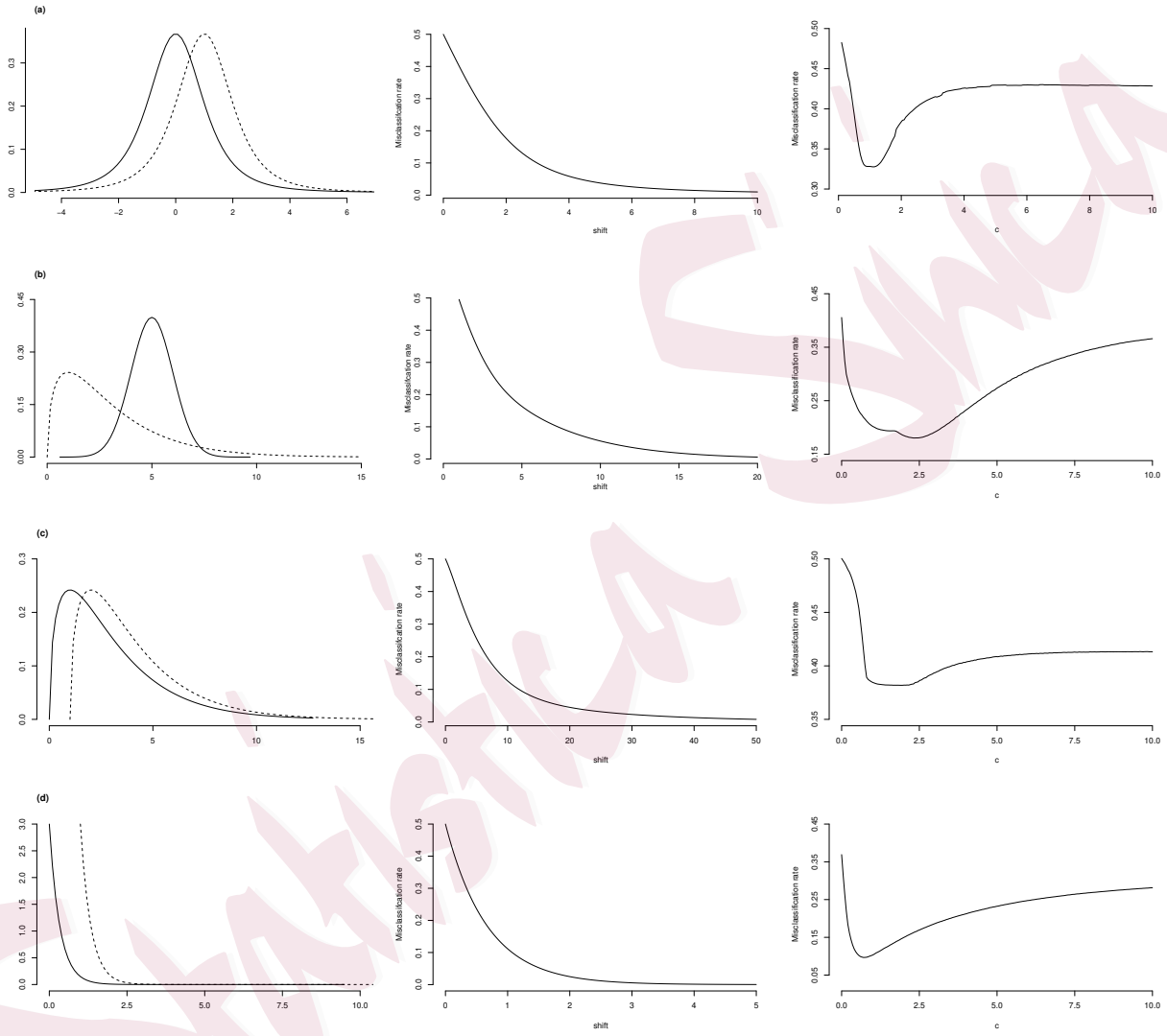

Figure 2: Misclassification rates in four different settings. The probability density functions of two populations are shown in each panel of the first column; each panel of the second column depicts the misclassification rate as the function of d assuming $\sigma_1 = \sigma_2$; The misclassification rate as a function of $c = \sigma_2/\sigma_1$ is shown in the last column. (a) location-shifted Gaussians; (b) a Gaussian distribution and a chi-squared distribution; (c) location-shifted chi-squared distributions; (d) location-shifted exponentials.

correct classification rate. We restrict ourselves to first select σ_{11} and then choose θ. Once the "best" $\hat{\sigma}_{11}$ is determined, the optimal θ can be subsequently identified by maximizing $\Gamma(\theta, \hat{\sigma}_{11})$. In Supplementary Materials, we present an illustrating example to show how to choose the "best" bandwidth σ_{11}.

Empirically, let $\pi_r = P(C_i = r)$ for $r = 1, 2$. Let $F_{r1}, \ldots, F_{rp}$ be the marginal unimodal distributions of $\mathcal{F}_r$ for $r = 1, 2$, and f_{rj} are the corresponding density functions. Denote by δ_{rj} the unique mode of F_{rj}. For arbitrarily small $0 < t < 1$, define $T = [t, 1/t]$. Let $\hat{\delta}_{rj}$ be the empirical mode for the subsample defined by $C_i = r$, and $\hat{\sigma}_{1jn}$ be the optimal bandwidth for the KDE $\hat{f}_{1j}$ defined by $C_i = 1$, $i = 1, \ldots, n$. The empirically *optimal unimodal classifier* is formulated by assigning $\boldsymbol{z}$ to $C = 1$ if

$$\lambda_n(\boldsymbol{z}, \hat{\theta}, \hat{\boldsymbol{\sigma}}_{1n}) = \sum_{j=1}^{p} \{\Psi_{1jn}(z_j, \hat{\sigma}_{1jn}) - \Psi_{2jn}(z_j, \hat{\theta}\hat{\sigma}_{1jn})\} > 0, \tag{2.11}$$

where $\Psi_{rjn}(z, \sigma_n) = K\big\{(z - \delta_{rj})/\sigma_n\big\}$ for $r = 1, 2$, and $\hat{\theta} = \arg\max_{\theta \in T} \Gamma_n(\theta, \hat{\boldsymbol{\sigma}}_{1n})$ is the estimated optimal θ from $(\boldsymbol{z}_1, C_1), \ldots, (\boldsymbol{z}_n, C_n)$, the observed rate of correct classification for θ in sample $(\boldsymbol{z}_i, C_i), i = 1, \ldots, n$ is

$$\Gamma_n(\theta, \hat{\boldsymbol{\sigma}}_{1n}) = \frac{1}{n}\left[\sum_{i:C_i=1} I\{\lambda_n(\boldsymbol{z}, \theta, \hat{\boldsymbol{\sigma}}_{1n}) > 0\} + \sum_{i:C_i=2} I\{\lambda_n(\boldsymbol{z}, \theta, \hat{\boldsymbol{\sigma}}_{1n}) \leq 0\}\right].$$

Practically, a grid search of θ is in principle possible. If the argmax is not unique like in the last panel in the third row of Figure 2, any maximizer can be chosen.

Remark 2. We search for the optimal value of θ in a compact interval T. To save computational cost, a small interval, such as $[1/3, 3]$, seems to be appropriate as discussed in Remark 1. $\Gamma_n(\theta)$ is thus assessed on a grid of equispaced values between 1/3 and 3.

Remark 3. Especially when the sample size is small, the issue of obtaining the precise componentwise mode of an "n sample, p-dimensional" data is of comparable difficulty to selecting the accurate bandwidth. To this end, we provide a fast to implement procedure

to seek the unique mode for a unimodal distribution but at the cost of some accuracy. This procedure we term as *naive unimodal classifier* is based on the formula

$$\text{mode} = 3 \times \text{median} - 2 \times \text{mean}. \tag{2.12}$$

The above relation holds directly for a symmetric distribution, and also holds approximatively for a wide range of asymmetric distributions Stuart (1994). The naive unimodal classifier assigns $\boldsymbol{z}$ to C_1, if $\sum_{j=1}^{p}\{K(z_j - \delta_{1j}) - K(z_j - \delta_{2j})\} > 0$, where the modes δ_{rj} are determined by (2.12) for $r = 1, 2$. Actually, we have tried various distance metrics and kernel functions, but the results vary slightly, therefore we still use Gaussian kernel for naive unimodal classifier. Even though this naive method might be problematic in practice, as it overlooks the intrinsic relationship between the mode and the datasets, at least the computational cost can be reduced considerably.

2.4 Multimodal classifiers

One more feature of mode-based classifiers that makes them differ largely from their competitors is that the mean, median and quantile always serve as a single value, whereas the density of a random variable U can have several modes (local maxima). To develop the multimodal classifiers, we first provide an algorithm to search for the multiple modes.

1. Multiple modes detector for multimodal distributions. The local modes can be interpreted as $\delta_{local} = \arg\max_{u\in\mathcal{I}} f_U(u)$, where $\mathcal{I}$ is a closed interval and the maximum is taken from the interior of the interval. Formally, the local mode set of U is defined as

$$\mathcal{M}(U) = \big\{u : f_U'(u) = 0, f_U''(u) < 0\big\},$$

which may consist of several points. Note that f_U is required to be twice differentiable.

It is assumed that $\mathcal{M}(U)$ has g distinct points. Given observations $U_1, \ldots, U_n \overset{i.i.d}{\sim} U$, the KDE of f_U using Gaussian kernel is $\hat{f}_U(\delta) = \frac{1}{n}\sum_{i=1}^{n}\varphi_\sigma(U_i - \delta)$, where $\varphi_\sigma = \sigma^{-1}\varphi(\cdot/\sigma)$.

Algorithm 1 MeanShift Algorithm: local modes detection

Input: data $\{U_i\}_{i=1}^n$, bandwidth $\sigma > 0$.

Initialize: $t = 0$, $\delta_0^{(1)} < \ldots < \delta_0^{(G)}$.

Output: the local mode set $\mathcal{M}(U) = \{\delta_m^{(1)}, \ldots, \delta_m^{(G)}\}$.

while not converge **do**

1. Compute $\delta_{t+1}^{(l)} = \mu(\delta_t^{(l)})$ for $l = 1, \ldots, G$.
2. Set $t := t + 1$.

end while

Hence, the local modes of U satisfy the following equation

$$\frac{d\hat{f}_U(\delta)}{d\delta} = \frac{1}{n\sigma^3}\sum_{i=1}^{n}\varphi\left(\frac{U_i - \delta}{\sigma}\right)(\delta - U_i) = 0.$$

It follows that the local mode estimator δ_m has an equivalent expression,

$$\delta_m = \sum_{i=1}^{n}\varphi\left(\frac{U_i - \delta_m}{\sigma}\right)U_i \Big/ \sum_{i=1}^{n}\varphi\left(\frac{U_i - \delta_m}{\sigma}\right). \tag{2.13}$$

However, equation (2.13) cannot be solved analytically, so we obtain the solution δ_m iteratively using a MeanShift tool. Set $\mu(\delta_m) := \sum_{i=1}^n \varphi_\sigma\big(U_i - \delta_m\big)U_i / \sum_{i=1}^n \varphi_\sigma\big(U_i - \delta_m\big)$, then $\mu(u) - u$, the so-called mean shift, takes value zero at mode δ_m by (2.13). Define $\delta_{t+1} = \mu(\delta_t)$, the sequence $\{\delta_t\}_{t=0,1,2,\ldots}$ is shown to converge to a nearby local mode δ_m, which is a fixed point of $\delta_{t+1} = \mu(\delta_t)$, for a given starting point δ_0 (Comaniciu and Meer, 2002). To detect all local modes for a multimodal distribution, we provide the MeanShift Algorithm 1.

In Algorithm 1, the starting points can be chosen as the G quantiles of $\{U_i\}_{i=1}^n$, and the values $\delta_m^{(l)}$ in $\mathcal{M}(U)$ are not necessarily distinct. The set $\widehat{\mathcal{M}}(u)$ is ordered, i.e. $\hat{\delta}_m^{(1)} \leq \ldots \leq \hat{\delta}_m^{(G)}$. The choice of the number G of starting points depends on the number of modes one expects, it can be decided roughly by counting the number of possible peaks of the KDE of U. However, kernel density estimates may result in different conclusions about the number of modes via different bandwidth selectors. As a consequence, we advocate

conducting a mode testing procedure before employing the MeanShift algorithm. The mode testing procedure is elaborated in Remark 4. To ensure that all the (local) modes of U can be detected, we suggest using a sufficiently large number of starting points, that is $G \geq g$. The choice of $G > g$ certainly implies that some branches will be found more than once, however, with an adequately high number of iterations (usually 30 is enough) all estimates belonging to the same branch are approximately equal.

2. Multimodal classifier. With the above preparation work, we next introduce the multimodal classifier that allows the number of local modes to vary across the marginal distributions F_{rj}. Denote by $\delta_{rj}^{(1)} < \ldots < \delta_{rj}^{(g_{rj})}$ the g_{rj} local modes of F_{rj} for $r = 1, 2$ and $j = 1, \ldots, p$. Let $\Psi_{rj}^{(l)}(z_j, \sigma_{rj}) = K\{(z_j - \delta_{rj}^{(l)})/\sigma_{rj}\}$ for $l = 1, \ldots, g_{rj}$. Define the multimodal distance between the jth component of $\boldsymbol{z}$ and the modes of F_{rj} by

$$d(z_j, \text{mode}(F_{rj})) := \min_{1 \leq l \leq g_{rj}} \left\{1 - \gamma \Psi_{rj}^{(l)}(z_j, \sigma_{rj})\right\}, \quad r = 1, 2, \quad j = 1, \ldots, p,$$

where σ_{rj} is the bandwidth for estimating density f_{rj}. Given two data sets $\mathcal{X}$ and $\mathcal{Y}$ from populations $\mathcal{C}_1$ and $\mathcal{C}_2$, and a new observation $\boldsymbol{z}$, $\boldsymbol{z}$ is assigned to $\mathcal{C}_1$, if

$$\lambda^*(\boldsymbol{z}, \boldsymbol{\sigma}) = \sum_{j=1}^{p} \left[\max_{1 \leq l \leq g_{1j}} \left\{\Psi_{1j}^{(l)}(z_j, \sigma_{1j})\right\} - \max_{1 \leq l \leq g_{2j}} \left\{\Psi_{2j}^{(l)}(z_j, \sigma_{2j})\right\} \right] > 0, \tag{2.14}$$

otherwise to $\mathcal{C}_2$. Given classes $\mathcal{C}_1$ and $\mathcal{C}_2$ with prior probabilities π_1 and π_2, the probability of the correct classification for the multimodal classifier based on the true modes is

$$\Gamma(\boldsymbol{\sigma}) = \pi_1 \int I\{\lambda^*(\boldsymbol{z}, \boldsymbol{\sigma}) > 0\} d\mathcal{F}_1(\boldsymbol{z}) + \pi_2 \int I\{\lambda^*(\boldsymbol{z}, \boldsymbol{\sigma}) \leq 0\} d\mathcal{F}_2(\boldsymbol{z}). \tag{2.15}$$

The observed rate of correct classification is

$$\Gamma_n^* = \frac{1}{n} \left[\sum_{i:\mathcal{C}_i=1} I\{\lambda_n^*(\boldsymbol{z}, \hat{\sigma}) > 0\} + \sum_{i:\mathcal{C}_i=2} I\{\lambda_n^*(\boldsymbol{z}, \hat{\sigma}) \leq 0\} \right].$$

Remark 4. (*mode testing procedure*) It is desirable to identify multimodality when it exists. As is shown, detecting the number of local maxima of the marginal distributions is considerably important to accurately implement the multimodal classifier. A variety of tests are

accessible to decide whether a data set is distributed bimodally or in a multimodal fashion, while different techniques have designed for different goals. Denote by j the true number of modes of a random variable X, given $k \in \mathbb{R}^+$, the testing problem is formulated as

$$H_0 : j \leq k \quad \text{versus} \quad H_1 : j > k.$$

Numerous procedures to this problem have been recommended, many of them can be classified into several collections: tests based on critical smoothing bandwidth (Silverman, 1981; Hall and York, 2001; Fisher and Marron, 2001); tests based upon the excess mass (Hartigan and Hartigan, 1985; Müller and Sawitzki, 1991; Cheng and Hall, 1998); tests based on the combination of smoothing and excess mass (Ameijeiras-Alonso et al., 2019). In addition to the above-mentioned formal testing procedures, a complementary step to identify the number of modes in a data distribution is the exploration of a nonparametric kernel density estimator, as stated in Section 2.4.

2.5 Alternative quantile-mode classifier

A disadvantage of the mode-based classifier is that it is sensitive to bandwidth selectors especially for the small sample size. This issue makes it accessible and appealing by employing the *quantile-mode classifier*, which is also a complement of the quantile-based classifiers proposed by Hennig and Viroli (2016). The quantile-mode classifier is motivated by the attractive feature that the mode is the most informative quantile for some distributions, for instance, asymmetric Laplace distribution (ALD). Specifically, the ALD takes the form as

$$f(y; \mu, \tau, \alpha) = \frac{1}{\alpha} \exp\left\{ -\frac{1}{\alpha} \rho_\tau (y - \mu) \right\},$$

where $\tau \in (0, 1)$ and α is a scale parameter. By maximizing the likelihood in ALD with respect to μ, the parameter μ is not only the τ-quantile but the mode of the ALD. Thereby, the selection of the most likely quantile of a distribution provides a convenient solution to the search for its mode. Any one of the classifiers decided by (2.11) and (2.14) can be made

even more powerful and robust by replacing the modes of the summands in the sums with the respective informative quantiles.

Denote by $q_U(\tau) = F_U^{-1}(\tau)$ the τ-quantile of U, which is the value q that minimizes $E\{\rho_\tau(U-q)\}$. Given $\tau \in (0,1)$, define the quantile-mode (qm) distance as

$$d(z_j, \mathrm{qm}(F_{rj})) := 1 - \gamma \Psi_{rj}(z_j, \tau), \quad r = 1, 2, \ j = 1, \ldots, p,$$

where $\Psi_{rj}(z_j, \tau) = K\big(z_j - q_{rj}(\tau)\big)$ and $q_{rj}(\tau) = F_{rj}^{-1}(\tau)$ for $r = 1, 2$ and $j = 1, \ldots, p$. We assign $\boldsymbol{z}$ to $\mathcal{C}_1$ if

$$\tilde{\lambda}(\boldsymbol{z}, \tau) = \sum_{j=1}^{p} \{\Psi_{1j}(z_j, \tau) - \Psi_{2j}(z_j, \tau)\} > 0, \tag{2.16}$$

the empirically optimal quantile-mode classifier is determined by assigning $\boldsymbol{z}$ to $\mathcal{C}_1$ if

$$\tilde{\lambda}_n(\boldsymbol{z}, \hat{\tau}) = \sum_{j=1}^{p} \{\Psi_{1jn}(z_j, \hat{\tau}) - \Psi_{2jn}(z_j, \hat{\tau})\} > 0,$$

where $\hat{\tau} = \arg\max_{\tau \in \tilde{T}} \Gamma_n(\tau)$, $\tilde{T} = [t, 1-t]$ for arbitrarily small $t > 0$ and the empirical correct classification rate for quantile-mode classifier is

$$\Gamma_n(\tau) = \frac{1}{n}\left[\sum_{i:\mathcal{C}_i=1} I\big\{\tilde{\lambda}_n(\boldsymbol{z}, \tau) > 0\big\} + \sum_{i:\mathcal{C}_i=2} I\big\{\tilde{\lambda}_n(\boldsymbol{z}, \tau) \le 0\big\}\right].$$

3. Some Implementation Issues

We discuss in this section some implementation issues on the mode-based classifiers, including the unimodal classifiers, the multimodal classifier and the quantile-mode classifier.

1. (Multi-class extension) Distance-based classifiers can readily allow themselves for R-class ($R > 2$) extensions. The rule of mode-based classifiers assigns any point $\boldsymbol{z} \in \mathbb{R}^p$ to the population that has the largest modal distance. Although the discussion of this work is primarily restricted to binary classification problems, multi-class extensions are straightforward. A multi-class real data example is analyzed in Section 6.

2. (Correction of the modal distance) Different distance metrics (losses) result in different classifiers. Mode-based classifiers assign an observation to a class based upon the

componentwise distance of equal weights. Take the unimodal classifier (2.4) as an illustration, the two quantities $\sum_{j=1}^{p}\Psi_{1j}(z_j,\sigma_{1j})$ and $\sum_{j=1}^{p}\Psi_{2j}(z_j,\sigma_{2j})$ in (2.4) contribute equally to the classification rule. If the shapes of marginal distributions for two populations are largely different, say F_{11} is symmetrically distributed and F_{21} is highly skewed, the equal weighting scheme seems to become inappropriate. A better option is to assign some weights to the distance metrics, i.e., $\sum_{j=1}^{p}\omega_{rj}\Psi_{rj}(z_j,\sigma_{rj})$, where $\omega_{rj}>0$ is the weight for $r=1,2$. The mode-based classifiers can be modified by reweighting the compenentwise distance metrics, for convenience, we choose $\omega_{rj}=1/\sigma_{rj}$. Then, an analog of the unimodal classifier is constructed in a modified modal distance fashion. We classify $\boldsymbol{z}$ to $\mathcal{C}_1$ if

$$\sum_{j=1}^{p}\left(\frac{1}{\sigma_{1j}}\Psi_{1j}(z_j,\sigma_{1j})-\frac{1}{\sigma_{2j}}\Psi_{2j}(z_j,\sigma_{2j})\right)>0. \tag{3.17}$$

Similarly, the optimal unimodal classifier with distance correction is obtained by replacing σ_{2j} with $\theta\sigma_{1j}$ in (3.17), the reweighted multimodal classifier is attained by replacing $\max_{1\le l\le g_{rj}}\left\{\Psi_{rj}^{(l)}(z_j,\sigma_{rj})\right\}$ with $\sigma_{rj}^{-1}\max_{1\le l\le g_{rj}}\left\{\Psi_{rj}^{(l)}(z_j,\sigma_{rj})\right\}$ for $r=1,2$. These modified classifiers generally show better performance especially in the analysis of real data that is more sensitive to outliers and in the settings of different types of distributions for predictors. Interested readers can try any other weights or even seek to derive the optimal weights.

3. (Data scaling) One problem inherent to distance-based classifiers is that the poor performance of such classifiers can be caused by the differences in the scales of the populations, since scale differences can mask location differences (Chan and Hall, 2009). For approaches that employ distance measures, data scaling is particularly crucial, it is regarded as an essential step prior to the implementation of the mode-based classifiers.

4. (Dependence structure of predictors) In practice, finding a classifier with both good prediction accuracy and low computational complexity is challenging especially confronted with high-dimensional data. Performance of classifiers depends greatly upon the volume of the input variables and the dependence structure within the data. The mode-based classifiers may still work empirically well in some situations with dependence, see simulation example

1 and its explanations.

5. (Feature selection) Though componentwise distance-based classifiers can adapt well to high-dimensional data without feature screening steps, feature screening is naturally appealing to practitioners for classifying high-dimensional data, especially under the scenarios where the information of classification is poorly characterized in the original predictors. To identify important predictors in the context of mode, we introduce a novel marginal screening index based on statistical modes and refer to it as mode-diff (MD) filter. The MD filter is established based on the main idea of Fan and Lv (2008) who use marginal statistics to filter out many noise features and keep all important ones. Let $X_1, \ldots, X_p$ be predictors that are characterized by R populations $\mathcal{C}_1, \ldots, \mathcal{C}_R$. The MD filter Δ_k is defined by

$$\Delta_k = \min_{1 \le i < j \le R} \left\{ |\text{mode}(X_k|\mathcal{C}_i) - \text{mode}(X_k|\mathcal{C}_j)| \right\}, \quad k = 1, \ldots, p.$$

If $R = 2$, Δ_k is reduced to $|\text{mode}(X_k|\mathcal{C}_1) - \text{mode}(X_k|\mathcal{C}_2)|$. The above MD filter is established based on the unimodality assumption. When X_k is multimodal, Δ_k can be represented as

$$\Delta_k = \min_{1 \le i < j \le R} \left\{ |\Omega_{(m_i)}(X_k|\mathcal{C}_i) - \Omega_{(m_j)}(X_k|\mathcal{C}_j)| \right\}, \quad k = 1, \ldots, p,$$

where $\Omega_{(j)}(\cdot) = j^{th}\text{local-mode}(\cdot)$. A larger Δ_k indicates the more discriminative power of X_k. We suggest determining the contribution of each predictor by applying the MD filter to the dataset and filter out many noise predictors. The MD filter is model-free and robust, it captures nonlinear dependence between Y and X_k. This screening step generally results in improved classification performance and more efficient computation.

6. (Tuning parameter selection) The empirically optimal unimodal classifier is based on the bandwidths of the first population $\boldsymbol{\sigma}_1$ and a θ that is optimal for all predictors. In general, methods for choosing the empirical bandwidth include cross-validation, plug-in and the bootstrap methods. This paper explores five different data-driven bandwidth selectors: the least squares cross validation selector (LSCV,Bowman (1984)); the Sheather-Jones plug-in selector (PI, Sheather and Jones (1991)); the smoothed cross validation selector (SCV,

Hall et al. (1992)); the normal scale bandwidth selector (NS, Chacón et al. (2011)), and the Silverman's rule of thumb selector (Silverman (1986)) $\sigma_{\mathrm{ROT}} = 1.06 \min\{s, \mathrm{IQR}/1.34\} n^{-1/5}$, where s is the sample standard deviation and IQR is the corresponding interquartile range. We advocate approximating the "best" σ_{11} by comparing these five different bandwidth selectors. We also advocate a smoothed bootstrap estimate of bandwidth $\boldsymbol{\sigma}_1$. For $X_1, \ldots, X_n$ be independent observations from a density f, we first construct an initial estimate of the density $\hat{f}_n(x; \sigma_0) = n^{-1} \sum_{i=1}^{n} \varphi_{\sigma_0}(x - X_i)$ and then resample from that. This can be accomplished by adding a random amount $\sigma_0 \epsilon$ to each resampled X_j^*, where ϵ is distributed with density $\varphi(\cdot)$. So $X_j^* \to X_j^* + \sigma_0 \epsilon$. The bootstrap choice of bandwidth $\hat{\sigma}_b$ is obtained by minimizing $\mathrm{BIMSE}(\sigma, \sigma_0)$ over σ_0, where $\mathrm{BIMSE}(\sigma, \sigma_0) = B^{-1} \sum_{j=1}^{B} \int (f_{nj}^*(x; \sigma) - \hat{f}_n(x; \sigma_0))^2 dx$, and $f_{nj}^*(x; \sigma) = n^{-1} \sum_{i=1}^{n} \varphi_\sigma(x - X_i^*)$ for $j = 1, \ldots, B$, B is the number of bootstrap samples. Additionally, in our proposed procedure, the best value of θ is determined by the training error. It is also desirable to choose the best θ using the cross-validation method.

4. Theoretical Properties

4.1 Consistency of the unimodal classifier for fixed p

This section considers theoretical properties of mode-based classifiers. For convenience, we only provide the theoretical results for unimodal classifiers determined by (2.4) and (2.11). The proof can be adapted in a similar manner for other modal classifiers. Let $\tilde{\theta} = \arg\max_{\theta \in T} \Gamma(\theta, \boldsymbol{\sigma}_{1n})$ be the optimal θ concerning the true optimal unimodal classifier. We make the following assumptions to facilitate the technical proofs, while these assumptions may not be the weakest ones.

(A1) Assume the kernel $K(\cdot)$ is continuous and of bounded variation, that is, there exist positive constants c_0 and c_1 such that $\sup_{u \in \mathbb{R}} |K(u)| = c_0 < \infty$ and $\sup_{u \in \mathbb{R}} |K'(u)| = c_1 < \infty$.

(A2) For $r = 1, 2$ and $j = 1, \ldots, p$, f_{rj} is continuous in the neighborhood of the mode

δ_{rj}, and $\sup_{\{t:|t-\delta_{rj}|>s_j\}} f_{rj}(t) < f_{rj}(\delta_{rj})$, for every $s_j > 0$.

(A3) $\sigma_{rj,n}$ is a fixed sequence of numbers such that $\sigma_{rj,n} \to 0$ and $n\sigma_{rj,n}/\log(n) \to \infty$, for $r = 1, 2$ and $j = 1, \ldots, p$.

(A4) For all $\sigma_{rj,n} > 0$, $P\{\sum_{j=1}^{p}[\Psi_{1j}(\boldsymbol{z}, \sigma_{1j,n}) - \Psi_{2j}(\boldsymbol{z}, \sigma_{2j,n})] = 0\} = 0$.

(A4') For all $\theta \in T$ and $\sigma_{1j,n} > 0$, $P\{\sum_{j=1}^{p}[\Psi_{1j}(\boldsymbol{z}, \sigma_{1j,n}) - \Psi_{2j}(\boldsymbol{z}, \theta\sigma_{1j,n})] = 0\} = 0$.

Assumption (A1) is satisfied by majority of kernels used in practice. Assumption (A2) guarantees that each predictor has a uniquely defined mode. Assumption (A3) or stronger ones for bandwidths are conventionally imposed in nonparametric kernel density estimators to derive consistency. Assumptions (A4) and (A4') are formulated for the unimodal classifier and the optimal unimodal classifier, respectively. These assumptions that assume the two populations are well separated are most advantageous for classification.

Theorem 1. *(Consistency of ordinary unimodal classifier when p is fixed) Under assumptions (A1)-(A4), for any* $\epsilon > 0$,

$$\lim_{n\to\infty} P\left\{\left|\Gamma(\boldsymbol{\sigma}) - \Gamma_n(\hat{\boldsymbol{\sigma}}_n)\right| > \epsilon\right\} = 0,$$

where $\Gamma(\boldsymbol{\sigma})$ *and* $\Gamma_n(\hat{\boldsymbol{\sigma}}_n)$ *are defined in (2.5) and (2.7), respectively.*

Theorem 2. *(Consistency of optimal unimodal classifier when p is fixed) Under assumptions (A1)-(A3) and (A4'), for any* $\epsilon > 0$,

$$\lim_{n\to\infty} P\left\{\left|\Gamma(\tilde{\theta}, \boldsymbol{\sigma}_{1n}) - \Gamma_n(\hat{\theta}_n, \hat{\boldsymbol{\sigma}}_{1n})\right| > \epsilon\right\} = 0.$$

Theorem 1 indicates in large samples, the true correct classification probability can be obtained by using the recommended bandwidth selection procedures, and Theorem 2 further implies by using the empirically optimal $\hat{\theta}_n$ in unimodal classifiers, the true correct classification probability is achieved. Both Theorems demand more stringent assumptions on the kernel functions and bandwidths in comparison to the quantile classifiers, which reflects the fact that the mode-based classifier is more difficult to handle theoretically than

the quantile classifier, since determining the mode accurately is inherently challenging. In addition to the consistency property, we also investigate the rate of convergence of our unimodal classifiers. Results and proofs are provided in the Supplementary Material.

4.2 A theoretical result for $p \to \infty$

Theorem 2 handles the case of fixed finite p. Below we will present theoretical results for unimodal classifiers when p approaches to infinity. In contrast to results for mode-based classifiers, Hall et al. (2009) have proven that the misclassification probability for the median classifier, under certain conditions, diminishes to zero as $n, p \to \infty$. Hennig and Viroli (2016) demonstrate that the performance of the quantile classifier is nearly the same as that of the median classifier regarding $p \to \infty$, under much stronger conditions than those of Hall et al. (2009), while for finite n, the quantile classifier can be superior to the median classifier. For our mode-based classifiers, we establish that the misclassification probability for the unimodal classifiers converges to zero as $n, p \to \infty$. The results require not only the assumptions of Hall et al. (2009) with some modifications but the restrictions on kernel functions and bandwidths.

Let $T = [t, 1/t]$ for arbitrarily small $0 < t < 1$. Let $U = (U_1, U_2, \ldots)$ denote an infinite sequence of random variables each with a uniquely defined mode, equal to 0 if necessary after a shift in location, and the corresponding density functions are $f_U(u) = (f_1(u), f_2(u), \ldots)$. Let $(\nu_{X_1}, \nu_{X_2}, \ldots)$ and $(\nu_{Y_1}, \nu_{Y_2}, \ldots)$ be infinite sequences of constants. Assume that for each p, the p-vectors $X_1, \ldots, X_m$ are identically distributed as $(\nu_{X_1} + U_1, \ldots, \nu_{X_p} + U_p)$, $Y_1, \ldots, Y_n$ are identically distributed as $(\nu_{Y_1} + U_1, \ldots, \nu_{Y_p} + U_p)$. Denote by $\Psi_k(U_k, \sigma_k) = \gamma K(U_k/\sigma_k)$, where $\gamma = 1/K(0)$. Let C be a $[1, 2]$-valued random variable, and assume Z to be distributed as X_1 if $C = 1$ and as Y_1 if $C = 2$. Further assume that $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ and (Z, C) are totally independent. The following conditions are needed:

(A5) $\lim_{\lambda \to \infty} \sup_{k \geq 1} E\{|U_k| I(|U_k| > \lambda)\} = 0$.

(A6) For each $c > 0$, $\inf_{k\geq 1} \inf_{|x|\geq c} \left[E\{\Psi_k(U_k, \sigma_k)\} - E\{\Psi_k(U_k + x, \sigma_{kx})\}\right] > 0$.

(A6') For each $c > 0$, $\inf_{k\geq 1} \inf_{|x|\geq c} \inf_{\theta\in T} \left[E\{\Psi_k(U_k, \sigma_k)\} - E\{\Psi_k(U_k + x, \theta\sigma_k)\}\right] > 0$.

(A7) For each $\epsilon > 0$, $\inf_{k>1}\{\sup_{u:|u|>\epsilon} f_k(0) - f_k(u)\} > 0$.

(A8) With $\mathcal{B}$ denoting the class of Borel subsets of the real line,

$$\lim_{k\to\infty} \sup_{k_1,k_2:|k_1-k_2|\geq k} \sup_{B_1,B_2\in\mathcal{B}} |P(U_{k1} \in B_1, U_{k2} \in B_2) - P(U_{k1} \in B_1)P(U_{k2} \in B_2)| = 0.$$

(A9) $|\nu_{X_k} - \nu_{Y_k}|$ are uniformly bounded. For sufficiently small $\epsilon > 0$, the proportion of values $k \in \{1, \ldots, p\}$ for which $|\nu_{X_k} - \nu_{Y_k}| > \epsilon$ is bounded away from zero as p diverges.

Assumption (A5) ensures that the first moment of U_k is uniformly bounded, so that variables like Cauchy variable are not theoretically applicable here; Assumptions (A6), (A6') and (A7) are assumed to assure that variables U_k's have uniquely defined mode, equal to 0, and these assumptions are required to hold uniformly in k; Assumption (A6') is similar to (A6), but requires the condition (A6) uniformly holds for $\theta \in T$. Specifically, (A6) is designed for the unimodal classifier, and (A6') is assumed for the optimal unimodal classifier; Assumptions (A8) and (A9) are the same as in Hall et al. (2009), but have slightly different sense. (A8) implies that the approximate independence of variables, and (A9) requires that the componentwise differences of modes are bounded away from zero, and this proportion cannot be neglected. For each $p \geq 1$, let Z denote a random variable drawn from either the X or the Y population. Let $\mathcal{M}(Z)$ denote the unimodal classifier determined by (2.4), and $\mathcal{M}_{opt}(Z)$ denote the empirically optimal unimodal classifier determined by (2.11), we have the following Theorem 3. The proof is built following a strategy similar to that used in Hall et al. (2009) and Hennig and Viroli (2016), although our premises require more assumptions on kernel functions and bandwidths. The proof given by them has been adapted to take into account the additional parameters $\{\delta_{rj}, \sigma_{rj,n} : r = 1, 2, j = 1, \ldots, p\}$ and the change of distance metrics from quantile-loss to the unimodal distance.

Theorem 3. *Assume that conditions (A1)-(A3) and (A5)-(A9) hold, both m and n diverge as $p \to \infty$. Then with probability converging to 1 as p increases, the unimodal classifier $\mathcal{M}_1$*

makes the correct decision, that is

$$P\{\mathcal{M}(Z) = 2|C = 1\} + P\{\mathcal{M}(Z) = 1|C = 2\} = 0.$$

Assume that conditions (A1)-(A3), (A5), (A6') and (A7)-(A9) hold, both m and n diverge as $p \to \infty$. Then with probability converging to 1 as p increases, the optimal unimodal classifier $\mathcal{M}_2$ makes the correct decision, that is

$$P\{\mathcal{M}_{opt}(Z) = 2|\mathcal{C} = 1\} + P\{\mathcal{M}_{opt}(Z) = 1|\mathcal{C} = 2\} = 0.$$

5. Numerical Experiments

This section carries out simulations to evaluate the performance of the proposed mode-based classifiers. We first consider the case of unimodal populations to investigate the performance of the three introduced componentwise unimodal classifiers, that is, the optimal unimodal classifier, the quantile-mode classifier and the naive unimodal classifier. We next design an example to examine the behavior of mulitmodal classifiers via multimodal populations, and the third example gives particular attention to the mixture of unimodal and multimodal populations. The MD filter is demonstrated in the last example on an artificial dataset. We further include 11 other successful classifiers in the literature for comparison, specifically, the centroid classifier, the median classifier, the quantile classifier, the ensemble quantile classifier (EQC), the Fisher's linear discriminant analysis (LDA), the naive Bayes classifier (nBayes), the nearest neighbor classifier (1-NN), the support vector machine (SVM), the adaptive Boosting (AdaBoost), the Gradient Boosted Decision Trees (GBDT), and the eXtreme Gradient Boosting (XGBoost). In all examples, we simulate two data sets $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ of p-variate data coming from two different populations characterized by $\mathcal{F}_X$ and $\mathcal{F}_Y$, where $X_1 = (X_{11}, \ldots, X_{1p})$ and $Y_1 = (Y_{11}, \ldots, Y_{1p})$. We set $n = 50, 100, 200$ and $p = 50, 200, 1000$ to mimic high-dimensional scenario and consider the balanced and imbalanced design, respectively. In the case of balanced design, we set $m = n$

for each dataset; while for imbalanced setting, we set $m = 2n$. Within each replication, each dataset is randomly and equally partitioned into a training set and a testing set, the training set contains $m/2$ samples simulated from population $\mathcal{F}_X$ and $n/2$ samples simulated from population $\mathcal{F}_Y$. The prior probability π_r can either be specified by user or estimated directly from the training data. We estimate π_1 and π_2 by $\hat{\pi}_1 = \frac{m}{m+n}$ and $\hat{\pi}_2 = \frac{n}{m+n}$ in our simulation studies. The experiment is replicated 100 times, we report the mean misclassification error rate and the standard deviation (SD) of the mean as well as the median of misclassification error rate with its associated robust estimate of standard deviation (RSD=IQR/1.34) obtained from the classification results for the testing tests. We also report the mean computational time (in seconds) required by each classifier to perform 100 repeated classifications. The following four examples are considered in this section:

Example 1. (Performance of unimodal classifiers) The two data sets $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ of p-variate data are simulated from the following cases.

Case 1: $U_j \sim t_2$ independently for $j = 1, \ldots, p$; $X_j \sim U_j$ and $Y_j \sim U_j + 0.5$;

Case 2: $U_j \sim \exp(3)$ independently for $j = 1, \ldots, p$; $X_j \sim U_j$ and $Y_j \sim U_j + 0.2$;

Case 3: $U_j \sim F(a_{1j}, b_{1j})$, $V_j \sim F(a_{2j}, b_{2j})$, a F-distribution with parameters $a_{1j}, a_{2j}, b_{1j}, b_{2j}$ randomly sampled in the range from 1 to 10; $X_j \sim U_j$ and $Y_j \sim V_j$ for $j = 1, \ldots, p$;

Case 4: $U_j \sim N(0, 1)$ independently for $j = 1, \ldots, p$, and p is splitted into four balanced blocks to which we apply various transformations: (i) $X_j \sim U_j$ and $Y_j \sim U_j + 0.5$; (ii) $X_j \sim U_j^2$ and $Y_j \sim U_j^2 + 0.5$; (iii) $X_j \sim \exp(U_j)$ and $Y_j \sim \exp(U_j) + 0.5$; (iv) $X_j \sim |U_j|^{0.5}$ and $Y_j \sim |U_j|^{0.5} + 0.5$;

Case 5: the same as Case 4 except that a dependence structure among variables is considered, that is, $(U_1, \ldots, U_p)^T \sim N_p(\mathbf{0}, \Sigma_{p\times p})$ with $\Sigma = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = \rho^{|i-j|}$. We consider $\rho = 0.2, 0.5$ and 0.8 in this case;

Case 6: $U_j \sim \chi^2_3$ independently for $j = 1, \ldots, p$. We divide p into five balanced blocks to which we apply equispaced location-shifts: (i) $X_j \sim U_j$ and $Y_j \sim U_j + 0.25$; (ii) $X_j \sim U_j$

and $Y_j \sim U_j + 0.5$; (iii) $X_j \sim U_j$ and $Y_j \sim U_j + 0.75$; (iv) $X_j \sim U_j$ and $Y_j \sim U_j + 1.0$; (v) $X_j \sim U_j$ and $Y_j \sim U_j + 1.25$;

Case 7: $U_j \sim \chi^2_3$ and $V_j \sim N(3, 1)$ independently for $j = 1, \ldots, p$. $X_j \sim U_j$ and $Y_j \sim V_j$.

Case 1 is designed to investigate the performance of the introduced unimodal classifiers in two location-shifted populations with symmetric and heavy-tailed distributions. Case 2 devotes more attention to the highly skewed data within two location-shifted populations. Case 3 examines different distributional shapes and levels of skewness for different populations; Case 4 deals with different transformations of standard normal distributions across p variables, and the dependence structure is further taken into account in Case 5; Case 6 addresses the problem of location-shifted populations with distinct location shifts across p variables. Case 7 finally analyzes two different populations with one of the populations being symmetric and the other being highly skewed.

The optimal unimodal classifier determined by (2.11) is implemented with R code. In each setting, we respectively adopt the normal scale bandwidth selector (NS,Chacón et al. (2011)) and the rule of thumb bandwidth selector (ROT, Silverman (1986)) to determine the proper σ_{1jn}, we evaluate the unimodal classifiers on a grid of equispaced θ values in $T = [1/3, 3]$. The optimal θ is chosen in each training set. For comparison purpose, the quantile-mode classifier determined by (2.16) and the naive unimodal classifier discussed in Remark 3 are taken into consideration. Results for misclassification rates of each classifier are displayed in the panel (b) of Figure 3, for a total of 162 different settings. Detailed results are provided in the Supplementary Materials.

Example 2. (Performance of multimodal classifier) We study the following cases:

Case 1: $U_j \sim \text{Beta}(\alpha, \beta)$, a beta distribution with parameters α and β, independently for $j = 1, \ldots, p$. We set $\alpha = \beta = 0.5$; $X_j \sim U_j$ and $Y_j \sim U_j + 0.2$;

Case 2: $U_j \sim w_j N(10, 1) + (1 - w_j)N(-10, 1)$, a mixture normal distribution, independently for $j = 1, \ldots, p$, w_j is randomly sampled from a uniform distribution on interval

$[0.4, 0.7]$; $X_j \sim U_j$ and $Y_j \sim U_j + 1$;

Case 3: $U_j \sim 0.3N(10,1) + 0.3N(-10,1) + 0.4t_3$ independently for $j = 1, \ldots, p$, $X_j \sim U_j$ and $Y_j \sim U_j + 0.5$;

Case 4: $U_j \sim \sum_{k=1}^{3} 0.25N(\mu_k, 1) + 0.25t_3$ independently for $j = 1, \ldots, p$, we set $\mu_1 = 10$, $\mu_2 = -10$ and $\mu_3 = -5$; $X_j \sim U_j$ and $Y_j \sim U_j + 1$;

Case 5: $U_j \sim \sum_{k=1}^{4} 0.2N(\mu_k, 1) + 0.2t_3$ independently for $j = 1, \ldots, p$, we set $\mu_1 = 10$, $\mu_2 = -10$, $\mu_3 = 5$, $\mu_4 = -5$; $X_j \sim U_j$ and $Y_j \sim U_j + 1$;

Case 1 and Case 2 deal with location-shifted populations with bimodal distributions, while the other cases consider the multimodal distributions with three or more modes of each marginal distribution. We apply the introduced multimodal classifier in this example by first determining the local modes using Algorithm 1 and then calculating the misclassification rate by classification rule (2.14). Hereafter we employ the ROT bandwidth selector to fit a multimodal classifier, since by panel (b) of Figure 3 classifiers with ROT selectors outperform those with NS selectors considerably. Results for misclassification rates of each classifier are displayed in the panel (c) of Figure 3, for a total of 90 different settings. Detailed results are provided in the supplementary materials.

Example 3. (Performance of mode-based classifiers) This example examines the performance of mode-based classifiers in the mixture of unimodal and multimodal distributions for two populations. p is divided into five balanced parts to which we apply different transformations of standard normal distribution: (i) $U_j \sim N(0,1)$, $X_j \sim U_j$ and $Y_j \sim U_j + 0.5$; (ii) $U_j \sim 0.5N(10,1) + 0.5t_2$, $X_j \sim U_j$ and $Y_j \sim U_j + 0.5$; (iii) $U_j \sim N(0,1)$, $X_j \sim \exp(U_j)$ and $Y_j \sim \exp(U_j) + 0.5$; (iv) $U_j \sim 1/3N(10,1) + 1/3N(-10,1) + (1/3)t_2$, and $Y_j \sim U_j + 0.5$; (v) $U_j \sim N(0,1)$, $X_j \sim U_j^2$ and $Y_j \sim U_j^2 + 0.5$. The (i), (iii) and (v) blocks consider unimodal populations, whereas (ii) and (iv) blocks are correspond to bimodal and trimodal populations, respectively. We employ the multimodal classifier determined by (2.14), results for misclassification rates of each classifier are displayed in the panel (d) of Figure 3, for a

total of 18 different settings. Detailed results are provided in the Supplementary Materials. In addition, results for misclassification rates of all the classifiers across the three examples are summarized in panel (a) of Figure 3, for a total of 270 different settings.

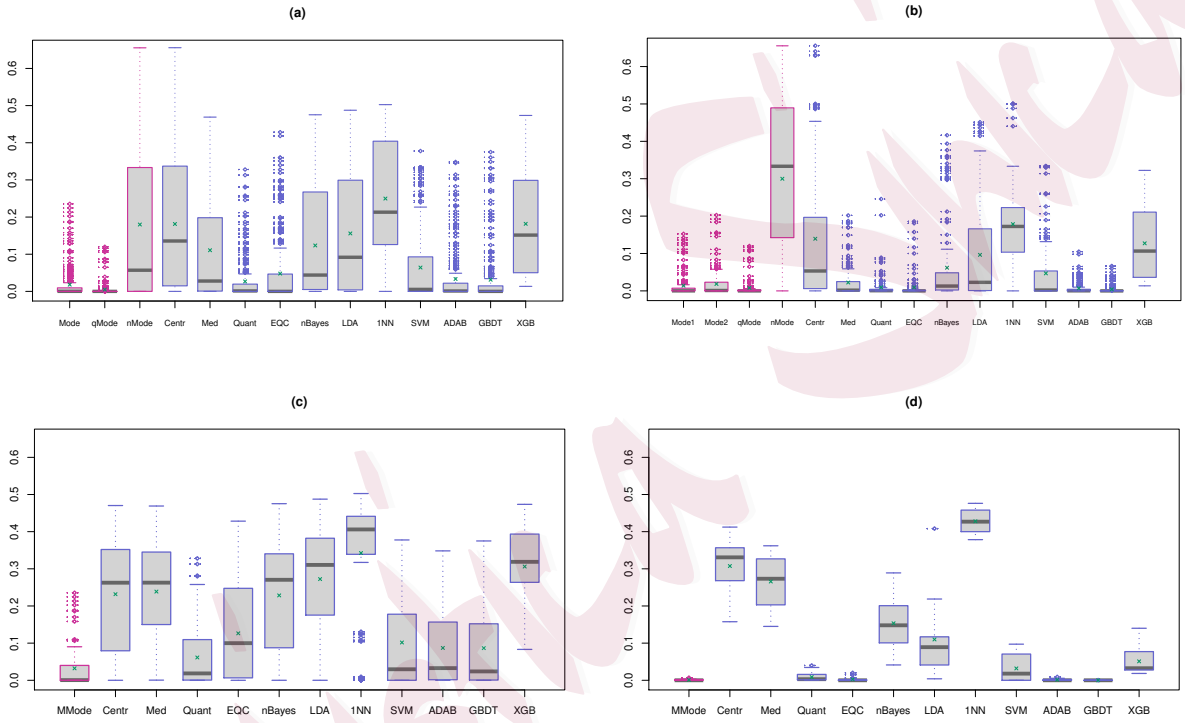

Figure 3: Performance of the classifiers for Examples 1-3. The labels denote the different classifiers, Mode1 is simplified for unimodal classifier with ROT bandwidth selector, Mode2 for unimodal classifier with NS selector, qMode for quantile-mode classifier, nMode for naive unimodal classifier, MMode for multimodal classifier, Centr for Centroid classifier, Med for Median classifier, ADAB for AdaBoost and XGB for XGBoost. Each panel shows the distribution of the misclassification rates for (a) all settings across three examples; (b) all settings of Example 1; (c) all settings of Example 2; (d) all settings of Example 3, with the cross indicating the mean.

It can be concluded from Figure 3 and the tables from Supplementary Materials that mode-based classifiers on the whole perform the best among all the classifiers, with lower misclassification rates and smaller standard errors, especially for the simulated heavy-tailed data. Specifically, we have the following observations. First, for all the classifiers, the mis-

classification rates decrease with the increase of sample size, while for fixed sample size, these methods seem to work better as the dimensionality increases in almost all settings. Second, for unimodal populations with symmetric and identical distributional shapes variables, the median and quantile classifiers show comparable performance to the unimodal classifiers and the naive unimodal classifiers. The chosen optimal values of θ for optimal unimodal classifiers are on average close to 1. However, the centroid, 1-NN, and XGBoost classifiers perform relatively worse at the presence of heavy-tailed features, such as t_2 location-shifted populations, by contrast, the unimodal classifiers work reasonably well and the quantile-mode classifiers are always slightly better in such scenarios. Third, in the cases of unimodal populations with asymmetric and identical distributional shapes variables, the unimodal classifiers and the quantile classifier have similar behaviors and outperform all the other competitors by large margin. As expected, the naive unimodal classifier established based on the relation between mean, median and mode in these cases are as bad as random guesses. Note that for case 3 of example 1, since the parameters a_{rj} and b_{rj} for $r = 1, 2$ are randomly sampled, it is possible that U_j has the identical distribution with V_j for some j, indicating that some predictors may lack discriminating power and do not contribute to the classification task. While, in such scenarios, the ensemble learning methods, such as GBDT and Adaboost, are capable of automatically selecting relevant variables during the fitting process, outperforming individual classifiers like the unimodal classifier. In sum, in terms of Example 1, unimodal classifiers with ROT bandwidth outperform those with NS bandwidths, the overall results of the AdaBoost and GBDT are comparative to those of unimodal classifiers, but they are rarely significantly better. Fourth, for multimodal populations, the multimodal classifier significantly outperforms the other competitors in all scenarios, conversely, the centroid, median, LDA, 1-NN and XGBoost behave relatively worse. Fifth, in the case of mixture of unimodal and multimodal distributions, the mode-based classifier again has a superior performance, with only the AdaBoost and GBDT achieving better

results sometimes. Not unexpectedly, the centroid and the median classifiers have unsatisfactory performance in this setting, and the 1-NN almost fails since it greatly depends on the similarity of features. Generally speaking, classifiers that compete well with the mode-based classifiers in one or two cases obviously drop behind in some others. Performances of the distance-based classifiers do not change much in the balanced and imbalanced design, and they are relatively insensible of the strength of dependence among components, since the information that differentiates each category could be collected from the original variables.

As suggested by one referee, we have also compared the computing time of our proposed classifiers with other competitors. As expected, the naive unimodal classifier requires the least amount of computing time, while the computational costs for both the optimal unimodal classifier and the multimodal classifier are comparable to that of the quantile classifier. The ensemble methods, such as GBDT and Adaboost, are relatively time-consuming. Due to limited space, we include the computational time comparison in Figure S2 and Tables S1-S24 in the Supplementary Materials.

Example 4. (Feature screening) Consider the following two scenarios: (i) $X_j \sim N(3,1), Y_j \sim \chi^2_3$; (ii) $U_j \sim 0.4N(0,1)+0.6N(3,0.25)$, $X_j \sim U_j$ and $Y_j \sim U_j+0.5$, for $j = 1, \ldots, 100$. Apart from the 100 predictors in the original dataset, 4900 independent noise variables following standard normal distribution and exponential distribution $\exp(1)-1$ are added respectively.

We include the mode-diff filter (MDF), two-sample t-tests (TT), Kolmogorov filter (KF, Mai and Zou (2013)), information gain based sure independence screening (IGS, Ni and Fang (2016)) and Pearson chi-square based sure independence screening (PCS, Huang et al. (2014)) for comparison. We first investigate whether the screening approaches can separate the important variables from the noise. For each screening approach, the top 100 predictors are kept, and in the two-sample t-tests, the 100 most important predictors are preserved in the sense of highest significance of two-sample t-tests. The proportion of the original 100 predictors captured by various screening methods are reported in Table 2. It is observed that

Table 2: Comparison of screening methods on the artificial dataset. The numbers (%) are averaged within 100 replications. Standard errors are in parentheses.

	Case (i)					Case (ii)				
noise	MDF	TT	KF	IGS	PCS	MDF	TT	KF	IGS	PCS
Screening performance										
$N(0,1)$	97.0(0.2)	1.3(0.1)	99.0(0.1)	65.1(0.4)	67.9(0.3)	100(0)	0(0)	93.5(0.2)	83.7(0.3)	83.2(0.3)
$\exp(1)-1$	99.9(0.0)	1.4(0.1)	98.9(0.1)	66.2(0.4)	67.1(0.4)	100(0)	0(0)	94.2(0.2)	84.7(0.3)	84.2(0.3)
Misclassification rate										
$N(0,1)$	0.1(0.1)	36.1(0.5)	0(0)	2.0(0.2)	2.0(0.2)	0.2(0.0)	59.2(0.2)	0.3(0.0)	1.8(0.0)	2.0(0)
$\exp(1)-1$	0(0)	38.0(0.4)	0(0)	0.1(0)	0(0)	0.2(0.0)	49.8(0)	0.3(0.1)	7.3(0.1)	12.5(0.1)

the mode-diff filter and the Kolmogorov filter have clearly better performance in keeping the true predictors. In particular, the mode-diff filter has a nearly perfect screening result especially for multimodal distributions. We further evaluate how variable screening can improve the classification accuracy. Again, we begin with the expanded dataset with the additional 4900 pure noise variables. The prediction is made by performing a mode-based classifier after screening. The misclassification rates over 100 replications are listed in Table 2. The mode-based method is observed to outperform all the other methods.

6. Real Data Examples

In this section, we demonstrate the mode-based classifiers on three real datasets, the prostate cancer dataset (Singh et al., 2002), the multiple myeloma dataset (Tian et al., 2003) and the DNA methylation dataset (Christensen et al., 2009). The Prostate Cancer dataset is collected and analyzed by a team of 15 scientists from a dozen institutions and is further researched by Efron (2008, 2010); Hall et al. (2009). It comprises $p = 6,033$ genes for $m = 50$ healthy males and $n = 52$ prostate cancer patients. This dataset is available in the R package "sda". The Multiple myeloma dataset comprises $p = 122,625$ genes for $m = 36$ patients with multiple myeloma in whom focal lesions of bone could not be detected and $n = 137$ patients in whom such lesions are detected. This dataset is available at https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS531. The DNA methy-

lation dataset contains 217 human tissues, with $n_1 = 19$ from placenta, $n_2 = 85$ from blood and $n_3 = 113$ from other solid tissue. Each tissue measures the expression values of $p = 1,413$ autosomal CpG loci associated with 773 genes. This dataset is available at https://github.com/ramhiser/datamicroarray. Obviously, all the three datasets are "small n, large p". The first two datasets are applied to fit binary classification models, while the third is used for assessing the performance of classifiers in multi-class instances.

To apply mode-based classifiers to these datasets and compare with other competitors, we first preprocess the data in the following steps: We remove the features that are least correlated with the target classification problem and then standardize each predictor to have zero mean and unit variance. For the first two datasets, we apply the proposed MD filter to the training set to select $p = 50, 200, 500$, and 1000 predictors by ranking MD values. For the third dataset, we keep $p = 20, 21, \ldots, 60$ to have a thorough understanding of the effects of dimensionality on each classifier's performance with limited sample size. We also include the centroid, the median, the quantile, EQC, nBayes, LDA, 1-NN, SVM, AdaBoost, GBDT and XGBoost for contrast analysis. We take $\pi_r = 1/R$ for $r = 1, \ldots, R$ for simplicity.

In the prostate cancer data, by a detailed exploration, most of the identified features are observed to have unimodal distributions and be positively skewed distributed. To get insight into the data and make sure of the number of modal groups for each predictor, we conduct mode testing for $p = 6,033$ genes via three above-mentioned testing procedures: SI (Silverman, 1981), FM (Fisher and Marron, 2001) and ACR (Ameijeiras-Alonso et al., 2019). We first test the null hypothesis of single mode versus the alternative of two or more modes, and then test the bimodality against more than two modes. The tests are performed at the nominal level of 0.05. If the null hypothesis is rejected, we continue the testing process, and test the null hypothesis of j modes versus the alternative of $j + 1$ or more modes. Results for testing for multimodality of each group are summarized in Figure S2 and Table S25 of the Supplementary Material, respectively. Figure S2 displays box plots

of the number of modes of each predictor for two groups determined by three tests, while the corresponding descriptive statistics are provided in Table S25. It can be seen that results of the SI test and ACR test are similar, but largely different from that of FM test. Among these three tests, SI test is relatively conservative, almost all the predictors are confirmed to be unimodally distributed by SI test. The median and mean of the number of modes for each predictor by ACR test is 1.0 and 1.5, respectively, indicating that most of the predictors are unimodally distributed, while a few predictors have more than two modes. In contrast to ACR and SI tests, FM test tends to discover larger number of modes, of which spurious modes caused by outlying data clusters might be also detected. About 27% of the predictors are tested to be unimodal, and the number of modes on average is around 18 by FM test. After implementing these testing tools, we apply unimodal classifiers based on the testing results of SI test and multimodal classifiers based on the results of FM and ACR tests to the preprocessed dataset. For unimodal classifiers, the optimal unimodal classifier with ROT bandwidth is implemented. The quantile-mode classifier and the naive unimodal classifier are also included. We use 10-fold cross-validation to assess the performance of all of the classifiers. Within each fold, the optimal θ of the optimal unimodal classifier is selected in the training set. Misclassification error rates for the first dataset are listed in Table 3, from which we can observe that the mode-based classifiers perform better than or comparably to the other 11 classifiers especially for larger dimensions, while the multimodal classifier based on FM test leads to poor results as compared to that based on SI and ACR test.

For the multiple myeloma dataset, a glance at the KDE plots of each identified predictor reveals that most of the predictors are unimodally or bimodally distributed, as shown in Figure S1 of the Supplementary Materials, in which we provide the density plots for four randomly selected genes from each group. Thereby, we first naively assume that all the predictors have unimodal or bimodal distributions, and directly utilize the unimodal and bimodal classifiers without statistical tests. Next, to accurately decide on the number of

Table 3: Mean misclassification error rates (%) for the prostate cancer dataset, with standard errors (%) in parenthese

Classifiers	$p = 50$	$p = 200$	$p = 500$	$p = 1000$
Unimodal classifier (SI test)	14.91 (8.56)	1.00 (3.16)	0.00 (0.00)	0.00 (0.00)
Quantile-Mode classifier (SI test)	17.55 (10.68)	2.91 (4.69)	0.00 (0.00)	0.00 (0.00)
Naive unimodal classifier (SI test)	15.73 (10.75)	3.91 (5.05)	3.00 (6.75)	1.00 (3.16)
Multimodal classifier (FM test)	49.00 (19.24)	48.91 (20.25)	11.64 (5.84)	20.64 (7.21)
Multimodal classifier (ACR test)	10.91 (12.88)	4.00 (6.99)	0.00 (0.00)	0.00 (0.00)
Centroid classifier	7.00 (8.23)	3.91 (5.05)	2.91 (6.65)	2.00 (6.32)
Median classifier	11.82 (9.03)	3.82 (6.54)	1.91 (4.03)	1.00 (3.16)
Quantile classifier	12.55 (12.91)	9.91 (9.43)	15.73 (8.43)	14.73 (7.03)
Ensemble quantile classifier	3.00 (4.80)	2.82 (4.54)	0.00 (0.00)	0.91 (2.87)
Naive Bayes classifier	7.91 (7.87)	3.91 (5.05)	4.00 (5.16)	2.91 (4.69)
LDA	17.82 (9.16)	14.82 (15.04)	20.64 (11.87)	30.36 (18.40)
1-NN	19.45 (8.84)	28.27 (13.95)	48.91 (16.43)	46.73 (16.08)
SVM	47.18 (21.49)	4.00 (6.99)	0.91 (2.87)	2.00 (4.22)
AdaBoost	4.91 (5.18)	4.00 (6.99)	5.91 (5.09)	4.91 (7.01)
GBDT	4.00 (5.16)	1.00 (3.16)	3.82 (8.06)	2.00 (4.22)
XGBoost	14.64 (10.64)	24.45 (13.22)	20.73 (7.58)	24.27 (14.21)

modes of each predictor, we follow the same analysis path as in the first dataset and employ three mode testing procedures. Testing results are summarized in Table S26 of the Supplementary Materials, from which the evidence of $j = 1$ mode for most of the predictors is strongly suggested. We subsequently apply our mode-based classifiers to the case where the number of modes is ascertained by each testing tool. 10-fold cross-validation is still used to evaluate the performance of the classifiers. Table S27 reports misclassification rates of all the classifiers. We find the unimodal classifiers are substantially better than other methods, and a significant performance improvement over bimodal classifiers can be made by using testing procedures, especially when the number of predictors is large. This suggests that the unimodal classifier is more applicable.

For the DNA methylation dataset, all the predictors are tested to be unimodal. In Figure 4, we show the performance of the unimodal classifier, where $p = [20, 21, \ldots, 60]$ is considered. For each p, the misclassification rate is obtained by using 10-fold cross-validation. For comparison purpose, we also include the results of naive unimodal classifier and other 11 competitors. The results show that all the classifiers perform similarly well in this dataset,

yielding comparable error rates. Performance of nMode, EQC and nBayes are almost identical, but somewhat worse than the other classifiers, while the unimodal classifier and GBDT slightly outperform all the others. Notably, the naive unimodal classifier (nMode) shows inferior performance compared to the centroid classifier as expected, as nMode calculates the mode using the relationship between the average and the median. By contrast, the unimodal classifier (Mode) generally outperforms the centroid classifier, though it can exhibit greater variability in misclassification rates across dimensionality p. As the dimension gets larger, more outliers are likely to appear, and the heavy-tailed issue may further be amplified, this may lead to the significant increase of misclassification error of centroid classifier. In summary, the proposed mode-based classifiers achieve high degree of robustness against heavy-tailed predictors, while losing little or no efficiency under normality.

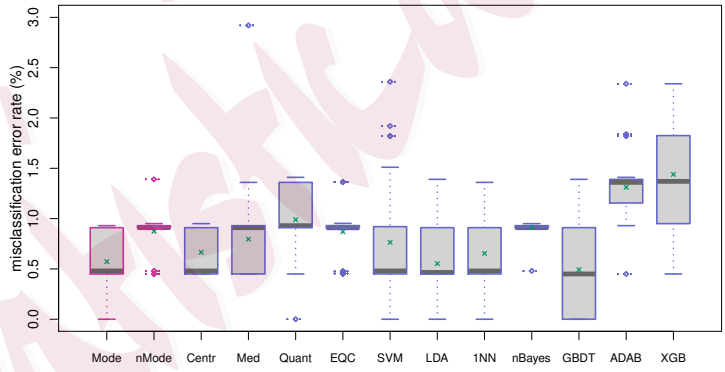

Figure 4: Box plots of misclassification error rates for DNA methylation dataset.

7. Concluding Remarks and Discussion

1. In this paper we have studied the componentwise mode-based classifiers for high-dimensional heterogeneous data. An algorithm that provide a means of detecting all the

local modes of a distribution is also presented. The introduced mode-based classifiers have many distinctive features: (i) they are robust and can handle a wider range of data distributions; (ii) they readily allow themselves to multi-class extensions; (iii) they perform remarkably well when multimodality exists; (iv) they are easy to implement and serve as complementary tools to the existing componentwise distance-based classifiers. The drawback of them, however, is that the computation is alway expensive, since in principle all pairwise distances between training instances and observations of each dimension are required to be computed and several smoothing parameters are needed to be tuned. To speed up classification, dimension reduction techniques can be used, and we introduce a mode-diff filter to effectively screen out inactive features for the target classification problem.

2. Theoretically, we only investigate the consistency of the proposed classifiers and establish the rate of convergence for unimodal classifiers when p is fixed. Another question is about the convergence rate and theoretical probability of misclassification for mode-based classifiers as both n and p go to infinity. Can the mode-based classifiers attain the optimal rate of convergence in high-dimensional scenario? What is the behavior of the theoretical misclassification error for different regimes of prior probabilities π_r and distributions of the predictors $\mathcal{F}_r$? These deserve some further studies.

3. The performance of the optimal unimodal classifier relies heavily on the choice of θ, which is required to be located in a closed interval mainly for mathematical derivation. To accurately identify the optimal θ, one could search in a wider range but at the cost of more computational complexity. In fact, determining the possible range of θ for real data is pretty hard due to the lack of prior information, thus how to select an appropriate θ with little computational cost needs further investigating.

4. To better implement mode-based classifiers, we should have a thorough understanding of the datasets as well as the number of modes of the marginal distributions for each predictor. Although a number of statistical tests are available to help to decide the number

of modes, the techniques employed in the field of multimodality testing have designed for different goals and they all shown their own defects. For example, the SI test is confirmed to be conservative, and the FM test tends to discover larger number of modes. So one may try other mode-testing procedures to alleviate this problem, but some new testing procedures that are suitable for the implementation of mode-based classifiers are highly demanded.

5. All results in this paper are based on misclassification probability with symmetric loss, an improvement over them might be produced by exploiting an asymmetric loss, for instance, choosing the decision boundary bounded away from zero, i.e., $\lambda_n(\boldsymbol{z}, \hat{\theta}, \hat{\boldsymbol{\sigma}}_{1n}) > c$ instead of $\lambda_n(\boldsymbol{z}, \hat{\theta}, \hat{\boldsymbol{\sigma}}_{1n}) > 0$.

6. Distance-based classifiers are typically powerful in distinguishing between populations that differ in location. However, scale differences can mask location differences, leading to the poor performance of classifiers. In consequence, the problem of making scale corrections for componentwise mode-based classifiers needs more discussion.

7. Results of three real datasets imply that the unimodal classifiers without statistical tests constantly perform on par with those using statistical tests, and show competitive performance even if the multimodality exists. For the sake of implementation, the more brief and feasible strategy in practice is to employ the optimal unimodal classifier directly.

Supplementary Materials

In the supplementary materials, we present additional results for simulation examples and real data analysis, and provide the technical results of Theorems 1-3.

Acknowledgements

The authors thank the referees and the associate editor very much for their many constructive and insightful comments which greatly improve the presentation. The research of W.

Xiong was supported in part by NSFC grants 12001101 and the Fundamental Research Funds for the Central Universities in UIBE CXTD14-05.

References

Ameijeiras-Alonso, J., R. Crujeiras, and A. Rodríguez-Casal (2019). Mode testing, critical bandwidth and excess mass. *Test 28*(3), 900–919.

Birgé, L. (1997). Estimation of unimodal densities without smoothness assumptions. *The Annals of Statistics 25*(3), 970–981.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*(2), 353–360.

Chacón, J., T. Duong, and M. Wand (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica 21*(2), 807–840.

Chan, Y. and P. Hall (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika 96*(2), 469–478.

Cheng, M. Y. and P. Hall (1998). Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(3), 579–589.

Christensen, B. C., E. A. Houseman, C. J. Marsit, S. Zheng, M. R. Wrensch, J. L. Wiemels, H. H. Nelson, M. R. Karagas, J. F. Padbury, and R. Bueno (2009). Aging and environmental exposures alter tissue-specific dna methylation dependent upon cpg island context. *PLoS genetics 5*(8), e10006024.

Comaniciu, D. and P. Meer (2002). Mean shift: A robust approach toward feature space

analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*(5), 603–619.

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning 20*(3), 273–297.

Cover, T. and P. Hart (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*(1), 21–27.

Eddy, W. F. (1980). Optimum kernel estimators of the mode. *The Annals of Statistics 8*(4), 870–882.

Efron, B. (2008). Microarrays, empirical bayes and the two-groups model. *Statistical Science 23*, 1–22.

Efron, B. (2010). The future of indirect evidence. *Statistical Science 25*, 145–157.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Farcomeni, A., M. Geraci, and C. Viroli (2022). Directional quantile classifiers. *Journal of Computational and Graphical Statistics 31*(3), 90–916.

Fisher, N. I. and J. S. Marron (2001). Mode testing via the excess mass estimate. *Biometrika 88*, 419–517.

Hall, P., J. Marron, and B. U. Park (1992). Smoothed cross-validation. *Probability Theory and Related Fields 92*(1), 1–20.

Hall, P., D. Titterington, and J. Xue (2009). Median-based classifiers for high-dimensional data. *Journal of the American Statistical Association 104*(488), 1597–1608.

Hall, P. and M. York (2001). On the calibration of silverman's test for multimodality. *Statistica Sinica 11*, 525–536.

Hartigan, J. A. and P. M. Hartigan (1985). The dip test of unimodality. *The Annals of Statistics 13*, 70–84.

Hennig, C. and C. Viroli (2016). Quantile-based classifiers. *Biometrika 103*(2), 435–446.

Huang, D., R. Li, and H. Wang (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business and Economics Statistics 32*, 237–244.

Jörnsten, R. (2004). Clustering and classification based on the l_1 data depth. *Journal of Multivariate Analysis 90*(1), 67–89.

Kemp, G. C. and J. S. Silva (2012). Regression towards the mode. *Journal of Econometrics 170*(1), 92–101.

Lai, Y. and I. McLeod (2020). Ensemble quantile classifier. *Computational Statistics & Data Analysis 144*, 106849.

Mai, Q. and H. Zou (2013). The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika 100*, 229–234.

Manski, C. F. (1991). Regression. *Journal of Economic Literature 29*(1), 34–50.

Mensink, T., J. Verbeek, F. Perronnin, and G. Csurka (2013). Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(11), 2624–2637.

Meyer, M. C. (2001). An alternative unimodal density estimator with a consistent estimate of the mode. *Statistica Sinica 11*, 1159–1174.

Minnotte, M. C. (1997). Nonparametric testing of the extence of modes. *The Annals of Statistics 25*(4), 1646–1660.

Müller, D. W. and G. Sawitzki (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association 86*(415), 738–746.

Ni, L. and F. Fang (2016). Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification. *Journal of Nonparametric Statistics 28*(3), 515–530.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics 33*(3), 1065–1076.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(3), 683–690.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological) 43*(1), 97–99.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell 1*(2), 203–209.

Stuart, A. (1994). Kendall's advanced theory of statistics. *Distribution theory 1*.

Tian, E., F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, and S. J. (2003). The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine 349*(26), 2483–2494.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science 18*(1), 104–117.

Webb, A. R. (2002). *Statistical pattern recognition*. John Wiley & Sons.

Wei Xiong, School of Statistics, University of International Business and Economics, Beijing, China

E-mail: xiongwei@uibe.edu.cn

Wolfgang Karl Härdle, Center for Applied Statistics and Economics, Humboldt-Universitat zu Berlin, Berlin, Germany

E-mail: haerdle@hu-berlin.de

Jianrong Wang, School of Statistics, University of International Business and Economics, Beijing, China

E-mail: athena420@126.com

Keming Yu, Mathematical Sciences, Brunel University, Uxbridge, UB8 3PH, London, United Kingdom

E-mail: Keming.Yu@brunel.ac.uk

Maozai Tian, School of Statistics, Renmin University of China, Beijing, China

E-mail: mztian@ruc.edu.cn