Noname manuscript No.
(will be inserted by the editor)

Nonparametric "Anti-Bayesian" Quantile-Based Pattern Classification

Fatemeh Mahmoudi ·
Mostafa Razmkhah ·
B. John Oommen

Received: date / Accepted: date

Abstract Parametric and nonparametric pattern recognition have been studied for almost a century based on a Bayesian paradigm, which is, in turn, founded on the principles of Bayes theorem. It is well-known that the accuracy of the Bayes classifier cannot be exceeded. Typically, this reduces to comparing the testing sample to mean or median of the respective distributions. Recently, Oommen and his co-authors have presented a pioneering and non-intuitive paradigm, namely, that of achieving the classification by comparing the testing sample with another descriptor, which could also be quite distant from the mean. This paradigm has been termed as being "Anti-Bayesian", and it essentially uses the quantiles of the distributions to achieve the pattern recognition. Such classifiers attain the optimal Bayesian accuracy for symmetric distributions even though they operate with a non-intuitive philosophy. While this paradigm has been applied in a number of domains (briefly explained in the body of this paper), its application for nonparametric domains has been limited. This paper explains, in detail, how such quantile-based classification can be extended to the nonparametric world, both using traditional and kernel-based strategies. The paper analyzes the methodology of such nonparametric schemes and their robustness. From a fundamental perspective, the paper utilizes the so-called "Large Sample" theory to derive strong asymptotic results that pertain to the equivalence between the parametric and nonparametric schemes for large samples. Apart from the new theoretical results, the paper also presents experimental results demonstrating their power. These re-

F. Mahmoudi and M. Razmkhah
Address of the first two authors: Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, P. O. Box 1159, Mashhad 91775, Iran. E-mail: razmkhah_m@um.ac.ir, raha.mhmy@yahoo.com.

B. John Oommen
Chancellor's Professor; *Life Fellow: IEEE* and *Fellow: IAPR*. Address: School of Computer Science, Carleton University, Ottawa, Canada: K1S 5B6. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway. E-mail: oommen@scs.carleton.ca.

sults pertain to artificial data sets, and also involves a real-life breast cancer data set obtained from the University Hospital Centre of Coimbra. The experimental results clearly confirm the power of the proposed "Anti-Bayesian" procedure, especially when approached from a nonparametric perspective.

Keywords "Anti-Bayesian" classification · Nonparametric quantile-based method · Mixture model · Sample quantile · Kernel density estimation · Robust classification

1 Introduction

Many scholars have investigated the problem of recognizing or classifying of patterns in data, as it is a fundamental problem boasting a long history. The process of Pattern Recognition (PR) (or synonymously, classification) includes two stages. In the first stage, the classifier is "trained" using a set of samples whose class identifications are known. In the second phase, known as the "testing" phase, one encounters an unknown sample which has to be assigned to one of the groups or classes [10]. The random variable representing the class is referred to by C, and the value of the class by c. The classification discussed in this paper is binary, and hence c can assume one of only two values, i.e., '+' (the positive class) or '-' (the negative class). A classifier is a function that assigns the label of a class to a test sample so as to optimize some criterion, for example, the classification accuracy. Bayesian classification is one of the traditional classification methods. Its competitive performance and optimality have made it a standard benchmark against other classifiers evaluated in terms of their optimality. The decision rule of a Bayesian approach classifies a test sample E using the equation:

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)},$$

where $p(\cdot)$ and $p(\cdot \mid \cdot)$ stand for probability and conditional probability of occurring events. In fact, E is assigned to class '+' iff

$$f_b(E) = \frac{p(c = +|E)}{p(c = -|E)} \geq 1,$$

where $f_b(E)$ is a function which is synonymously referred to as the "Bayesian classifier"[1] [9].

Recently, an innovative method, which is surprisingly unprecedented in its nature, was proposed by Oommen and Thomas [30,31]. The method, explained in more detail in the next section, is counter-intuitive in that it is based on testing the samples against non-central quantities of the various distributions.

[1] Over the last century, there are, indeed, tens of thousands of papers describing the art and science of Bayesian classification – for a myriad of distributions and applications. In this paper, we do not attempt a survey of the field.

It is, thus, aptly termed as being "Anti-Bayesian" (AB). The astonishing feature of this novel method is that in spite of its counter-intuitive strategy, it is as optimal as Bayesian classification for symmetric distributions such as the Normal distribution, and near-optimal for asymmetric distributions, such as the Gamma and Rayleigh distributions. When it was first discovered, it was expressed as being based on the order statistics of the distributions. However, subsequently, with a deeper insight, the authors demonstrated that the method was based on the distributions' quantiles rather than their order statistics [33]. Thereafter, it was referred to as "Classification by Moments of Quantile Statistics (CMQS)" [22].

Although CMQS and Bayesian classification attain an equivalent accuracy in symmetric distributions, they are different in terms of their respective procedures for classification. If the classes have equal *a priori* probabilities, in order to decide the class that the testing sample points should belong to, the process of Bayesian classification boils down to computing a distance (for example, the Mahalanobis distance) between the test instance and the means of the classes for symmetric distributions [31]. In a contrasting procedure, CMQS is based on the (Mahalanobis) distances of every test sample to the corresponding symmetric quantiles of the distributions. The amazing thing is that these quantile points can be quite distant from the mean points. To date, the formal properties of this method have been only proven in a parametric set-up, in which the classes are assumed to follow a known distribution, such as the Gaussian, Exponential etc. [32]. However, as explained presently, the *experiments* that have been done have validated some of the claims for nonparametric settings where there are no restrictions on the distributions of the data set.

Despite the fact that there are many advantages to parametric methods, there are still a number of situations in which nonparametric methods can act as a potential substitute technique to the family of parametric methods, especially when the distribution is unknown. Our position is the following: In parametric methods, one must assume the distributional form of the features' classes, which is what makes these schemes less desirable than the nonparametric ones, where one works with data whose distributions are badly skewed. In such situations, the distributional assumptions, for example, of normality, is doubtful, and consequently, parametric methods yield a poor accuracy or are not applicable at all [17]. On the other hand, when there are some outliers in the data set (i.e., when distribution is a mixture of two or more different distributions), one can safely assert that parametric statistics are not accurately robust in their nature [14,18]. The above-mentioned facts motivate us to develop nonparametric quantile-based classification strategies, which we believe, fundamentally expand the horizon of the published CMQS-based results.

The objectives of this paper are four-fold:

1. Firstly, we construct a nonparametric version of the CMQS method utilizing two different approaches. In the first, through a simple and easy-to-understand procedure, one obtains sample quantiles of the two classes, and an improved schema of CMQS is performed using *these* quantiles. In the

second approach, a unique distribution is fitted on the available data set using a Kernel Density Estimation (KDE) scheme.
2. Thereafter, the symmetric quantiles of the corresponding distributions are utilized to determine the classification boundary.
3. Apart from the latter aspects, we also demonstrate an intriguing phenomenon of nonparametric CMQS in dealing with outliers and skewed distributions. The contribution of this part of the paper is to focus on the high potency of nonparametric CMQS in classifying data sets containing outliers, and in explaining the logical rationale behind this performance.
4. Most importantly, from a foundational and theoretical perspective, this paper uses the so-called "Large Sample" theory as a premise to derive strong asymptotic results that pertain to the equivalence between the parametric and nonparametric paradigms for large samples. This, in one sense, closes the loop. In other words, although the basis of the paper is the parametric CMQS scheme used to motivate the nonparametric paradigm, the large sample analysis demonstrates that both the paradigms are identical when the number of samples is large. In other words, it provides the necessary theoretical foundations for further research in both the parametric and nonparametric worlds.

The rest of the paper is structured as follows. First of all, in Section 2, we present a relatively brief overview of the state-of-the-art of the "Anti-Bayesian" paradigm for classification. Thereafter, in Section 3, we present some preliminaries to provide a groundwork for presenting the nonparametric procedures. Section 4 clarifies the newly-proposed methodology for nonparametric classification. Some theoretical analysis of the proposed scheme builds Section 5. The main results for the nonparametric paradigm are given in Section 6. The robustness of this nonparametric approach against outliers is considered in Section 7. Apart from the simulations done using artificial data sets (to explain the procedure proposed in the paper), the paper's formal results have been confirmed by testing them on a real-life data set[2], obtained from the University Hospital Centre of Coimbra. These results are found in Section 8. Section 9 concludes the paper.

2 The "Anti-Bayesian" paradigm

This section is intended to lay a foundation for the rest of the paper. It presents, relatively chronologically, how the "Anti-Bayesian" paradigm works, explains its foundations, and records the results that are available for the exponential family and for multi-dimensional features. We shall also mention how it has been used in border identification, prototype reduction, text classification and clustering.

The "Anti-Bayesian" paradigm works by performing all the comparisons for a testing sample with samples that could be distant from the distributions'

[2] We are very grateful to the anonymous Referee of the previous version of the paper, who requested this.

means or some other central statistics, such as medians or quantiles. This is the precise reason why CMQS is referred to as being "Anti-Bayesian" classification. By way of example, consider Figure 1, where the two classes ω_1 and ω_2, have density functions given in "solid" and "dashed" lines, respectively, in a *uni*-dimensional space.

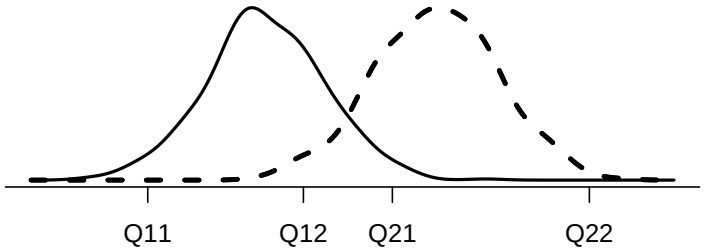

Fig. 1 This figure explain *how* the "Anti-Bayesian" paradigm works. The Quantile-based points Q_{11} and Q_{12} are computed for ω_1, and Q_{21} and Q_{22} are computed for ω_2. The classification is based on comparisons with regard to these Quantile-based points and not with regard to the means. Additional details are in the text.

The "Anti-Bayesian" paradigm first determines two points[3] Q_{11} and Q_{12} for the first class, ω_1, and two points Q_{21} and Q_{22} for the second class, ω_2. These points are typically distant from the mean, and could even lie towards the extreme boundaries of the domain of the feature space. Indeed, they are determined by certain well-defined quantiles which are also symmetric with respect to the median of each class; they are referred to as the "Quantile Statistics" (QS) points of the two classes. The classification is now achieved by comparing the testing sample, x^*, with Q_{12} for ω_1, and Q_{21} for ω_2. The reader should observe that although the classification is achieved in a counter-intuitive manner, the accuracy is exactly or close to the optimal Bayesian accuracy. Indeed, the testing is done as follows:

- If $x^* < Q_{12}$, then $x^* \in \omega_1$;
- If $x^* > Q_{21}$, then $x^* \in \omega_2$;
- If $Q_{12} < x^* < Q_{21}$, the decision is based on $D(x^*, Q_{12})$ and $D(x^*, Q_{21})$, where $D(a, b)$ stands for the distance between the points a and b.

The pioneering and fundamental results of the "Anti-Bayesian" paradigm published in [31]. These authors provided a theoretical framework for adequately responding to the question of why the border points are more informative for the task of classification. To justify these claims, the authors

[3] Initially, the authors of [31] stated that the classification was based on the Order Statistics of the distribution, and this was later rectified [33].

submitted a formal analysis and the results of various experiments which were performed for many distributions. The results were clearly conclusive. The results presented in [31], were then extended in [21] for members of the exponential family. They theoretically proved that the proposed approach could attain the optimal bound for symmetric distributions like the Doubly Exponential, Gaussian, and symmetric Beta. However, the proposed approach could attain a near-optimal bound for non-symmetric distributions such as the Rayleigh. Analogous results are derived by Thomas and Oommen [32] for the multidimensional features.
All of the above results operated with a parametric setting. The results assumed the distributional form for the class-conditional distributions. The departure from the parametric to a nonparametric model (which only assumed the existence of the training/testing data) were the results that related to Prototype Reduction Schemes (PRSs), Border Identification (BI), text classification and clustering. These are briefly stated below.

"Anti-Bayesian" prototype reduction schemes: The objective of PRSs is to reduce the number of training vectors, while simultaneously attempting to guarantee that the classifier built on the reduced design set performs as well, or nearly as well, as the classifier built on the original design set. Some initial results involving the development of PRSs using an "Anti-Bayesian" paradigm are found in [34].

"Anti-Bayesian" border identification algorithms: The BI algorithms, a subset of PRSs, aim to reduce the number of training vectors so that the reduced set (the border set) contains only those patterns that lie near the border of the classes, and yet have sufficient information to perform a meaningful classification. The only-reported results pertaining to "Anti-Bayesian" BI are found in [35].

"Anti-Bayesian" text classification and clustering: In all the prior recorded Text Classification (TC) papers reported in the literature, the schemes worked using the fundamental principle that once the statistical features are inferred from the syntactic/semantic indicators, the classifiers themselves are the well-established statistical ones. The pioneering application of "Anti-Bayesian" principles in TC and clustering were pioneered by two teams, and these results are contained in [22] and [12], respectively.

The reader will observe that such "Anti-Bayesian" quantile-based PR has been applied to the above domains, where the data is assumed to follow an exact distribution. It has also been used for some nonparametric domains, as explained above. In fact, since the exact distributions of real-world data sets are usually not known, in working with parametric methods, one often encounters problems that arise due to distribution-based assumptions, implying that its application for nonparametric models has been limited. This is the primary intent of this paper. This paper explains, in detail, how such quantile-based PR can be extended to the nonparametric world, using both traditional and kernel-based strategies. The paper analyzes the methodology of such nonpara-

metric schemes and their robustness, and also presents experimental results demonstrating their power.

3 Preliminaries of Nonparametric Classification

Suppose the random variables $X_1, \ldots, X_n$ are arranged in an ascending order. Let the k^{th} smallest one be denoted by $X_{(k)}$, $1 \leq k \leq n$. Then, $X_{(1)} < \cdots < X_{(n)}$ are called the order statistics of the random data set. One of the most important characteristics of the order statistics is that they can be used to summarize the data set. Indeed, there are situations in which the minimum, maximum or just the k^{th} data point is of great importance. These statistics have had a number of applications in inferential topics such as estimation, sufficiency, normality testing and statistical quality control. For more details about the properties and applications of order statistics, we refer the reader to [5,11].

Assume $X_1, \ldots, X_n$ are independent and identically distributed (iid) random variables having the cumulative distribution function (cdf) $F(\cdot)$, and the probability density function (pdf) $f(\cdot)$. Then, the pdf of $X_{(k)}$ is given by

$$f_{(k)}(x) = k\binom{n}{k} f(x) F^{k-1}(x) \bar{F}^{n-k}(x)$$

where $\bar{F}(\cdot) = 1 - F(\cdot)$ and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

In parametric methods, the distribution of the data set is considered to be known. More precisely, a complete knowledge of the underlying distribution for the real data set is a prerequisite to pursue the parametric procedure. However, there are well-established elegant nonparametric techniques to resolve the problem of estimating the cdf when one is disinclined to assume a distribution, or when there are situations in which parametric methods possess infirmities and thus yield weak accuracies. In such cases, one can utilize the methods which have the potential to estimate a distribution rather than considering a distribution for the data sets. Two common approaches for the estimation of the distribution function are the methods that use the empirical distribution function, and those that use kernel-based schemes, briefly described below.

Let $X_1, \ldots, X_n$ be iid random variables that come from an unknown cdf $F(\cdot)$. The empirical distribution function is then defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \tag{1}$$

where $I(A)$ is the indicator function on the event A. In other words, for a given x, it represents the ratio of the random variables among $X_1, \ldots, X_n$ which are less than or equal to x. Since the mathematical expectation of $F_n(x)$ is the true cdf $F(x)$, it is straightforward to show that this estimator is unbiased. Moreover, $nF_n(x)$ has binomial distribution with parameters n and $F(x)$; hence, it can be shown that $F_n(x)$ is a consistent estimator of $F(x)$.

On the other hand, Rosenblatt [24] pioneered and proposed kernel-based function estimation in order to estimate the distribution function. The kernel density estimator (KDE) of the univariate pdf $f(\cdot)$ is given by

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - X_i}{h}),$$

where h is an scale parameter referred to as the smoothing parameter or bandwidth, and $k(\cdot)$ is a kernel function satisfying the following conditions:

- $k(\cdot)$ is a continuous and symmetric function,
- $\int_{-\infty}^{\infty} k(u)du = 1$.

Further, the cdf may be estimated by integrating the kernel estimator of the corresponding pdf. More precisely, based on a sample of size n, the kernel estimator of the cdf $F(\cdot)$ can be represented as below:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{2}$$

where $K(u) = \int_{-\infty}^{u} k(t)dt$. Altman and Léger [4] studied the problem of bandwidth selection for the kernel-based estimation of distribution functions. When there is no *a priori* knowledge, the Gaussian kernel is the most widely-used kernel function, and this is defined as:

$$k(x) = (2\pi)^{-1/2} e^{-x^2/2}. \tag{3}$$

In this study, we use the kernel presented in Eq. (3) in all our experiments. The optimal choice for bandwidth that minimizes the mean integrated squared error is given by:

$$h \approx 1.06\hat{\sigma}n^{-1/5}, \tag{4}$$

where $\hat{\sigma}$ represents the sample standard deviation and n is the size of training data set. We refer the interested reader to [26] for applications of this method in classification.

This study deals with univariate kernel-based density estimation. We refer the reader to [26] for more information about multivariate kernel-based density estimation.

4 Proposed Methodology for Nonparametric Classification

Before we proceed to the nonparametric formulation, it is pertinent to highlight that the parametric CMQS works with the symmetric $(\frac{n-k+1}{n+1})^{th}$ and $(\frac{k}{n+1})^{th}$ quantiles, which attains the optimal and near-optimal classification accuracy in symmetric and asymmetric distributions, respectively. In fact, every test instance is compared to the symmetric quantiles of the classes. Since the form of the distribution is of primary importance, and since this is what yields the population's quantiles, this assumption is a cornerstone requirement.

Let the random variable X have an univariate cdf $F(\cdot)$, then, the population quantile of order p, $(0 < p < 1)$ is defined as

$$\xi_p = \inf\{x : F(x) \geq p\}, \tag{5}$$

which is also referred to as quantile function. Assuming X is a continuous random variable, the ξ_p satisfies $F(\xi_p) = p$. Additional details of such a method of approximation is found in [28], and we encourage the interested reader to this reference for further details.

The distributional assumption in parametric methods, imposes a stringent restriction. Indeed, one has to choose a distribution, which is then assumed to be *fitted* to the data set. In this set up, the quantiles of the assumed or fitted distribution are derived using Eq. (5). As mentioned earlier, if the distribution is badly skewed, the classification based on a nonparametric scheme works more efficiently than a parametric CMQS [17] based on a symmetric distribution, such as normal distribution. To allow this, we mention that several nonparametric estimators for quantile functions have been investigated for the case of unknown distributions, and in what follows, we present two of these.

4.1 Nonparametric Classification based on Empirical Quantiles

The Empirical Quantile (EQ), which is defined as the inverse of the empirical cdf, is one of the nonparametric solutions which has the potential of being exploited to estimate the quantiles without assuming any distributional form. Consider the case when $X_{(1)}, \ldots, X_{(n)}$ are the order statistics of a random sample of size n from the underlying population with unknown cdf $F(\cdot)$. Then, using Eq. (1) and Eq. (5), it can be shown that the EQ function is

$$\hat{\xi}_p = F_n^{-1}(p) = X_{([np])}, \tag{6}$$

where $[a]$ stands for the integer part of real number a. One can thus easily see that the order statistics play an important role in inferences related to the quantiles[4].

Having characterized the concept of EQs and formalized "Anti-Bayesian" classification, we are now in a position to present our proposed nonparametric classification method based on EQs, denoted by NCEQ. To explain the details of this procedure, we assume that the problem involves a binary classification involving two classes, ω_1 and ω_2. Additionally, let $(\hat{Q}_{11}, \hat{Q}_{12})$ and $(\hat{Q}_{21}, \hat{Q}_{22})$ be the symmetric EQs of orders $\left(\frac{k}{n+1}, \frac{n-k+1}{n+1}\right)$, for $k < n/2$, of classes ω_1 and ω_2, respectively. Similar to the criteria of classifying testing samples in CMQS, there are two (Euclidean or Mahalanobis) distances in NCEQ which must be compared in order to decide on which class the testing sample belongs to. There are two differences between the criteria by which NCEQ classifies

[4] With going into too many details, we refer the reader to [5], which is a key reference in this field.

the testing samples in comparison with the criterion of classification based on CMQS. Firstly, in NCEQ, the empirical quantiles are utilized instead of quantiles of distributions. Secondly, the distances are not constant values as they were in CMQS. In fact, they are dependent to the data sets which makes them *random variables*. Hence, prior to proceeding further, a criterion by which random variables can be compared must be defined. In probability theory and statistics, "*Stochastic order*" is defined to determine if a random variable is less than another. Let X and Y be two random variables such that

$$Pr(X > x) \leq Pr(Y > x), \qquad \forall x \in (-\infty, \infty). \tag{7}$$

Then X is said to be smaller than Y in the usual stochastic order, denoted by $X \leq_{st} Y$ [29]. By employing this concept to compare the distances of each testing sample x to the corresponding EQs, we can now define the rule of classifying x in NCEQ as follows

$$D(x, \hat{Q}_{12}) \leq_{st} D(x, \hat{Q}_{21}) \Rightarrow x \in \omega_1, \tag{8}$$

otherwise $x \in \omega_2$, where $D(a, b)$ stands for the distance between a and b. Note that the classification rule in (8) is true when $\hat{Q}_{12} \leq_{st} \hat{Q}_{21}$. Otherwise, the EQs $\hat{Q}_{11}$ and $\hat{Q}_{22}$ are used instead of $\hat{Q}_{12}$ and $\hat{Q}_{21}$, respectively. This rule is referred to as "dual NCEQ".

It is pertinent to mention that this methodology, the NCEQ strategy, is akin to the nonparametric schemes invoked for applying the "Anti-Bayesian" paradigm for obtaining prototypes [34], Border Identification [35], in text classification [22] and in clustering [12]. In all these cases[5], the respective authors have empirically computed the quantiles sought for (for example, those that pertain to the $\frac{1}{3}$ and $\frac{2}{3}$ quantile locations of the respective distributions), and thereafter achieved the "Anti-Bayesian" classification. The difference here is that we have formally applied the EQ-based results from Eq. (6) to get these locations, and then used the corresponding distance comparisons to formulate the class assignments.

4.2 Nonparametric Classification using Kernel-based Quantile Estimation

The alternate approach is to estimate the quantile function by inverting the kernel estimate of the cdf in (2). The solution that uses this approach is said to involve kernel-based quantile (KQ) estimation. The kernel estimation enables us to fit an unknown distribution to any data set, using which it is feasible to find the quantiles of the distributions without requiring us to assume a distributional form. Of course, this requires the second order properties of the kernel estimators of both the distribution function and the quantiles, as explained in detail in [7].

[5] To be fair to the authors of [12], [22], [34] and [35], one must grant them the credit that they were able to achieve their nonparametric results by using the "Anti-Bayesian" paradigm in *multidimensions*, as opposed to unidimensions, as we have done here!

Let us denote the nonparametric classification using KQ estimation by NCKQ. We will presently demonstrate that this method possesses a superior performance than the parametric CMQS when it concerns classification for skewed distributions. Towards this end, we assume a Gaussian kernel with the bandwidth specified in Eq. (4).

Unlike the NCEQ strategy described above, there is no "prior art" when it concerns the NCKQ. It is completely different from the methods used in [12], [22], [34] and [35]. Firstly, the approximations to the distributions of both classes are done in a novel way. Secondly, the quantiles themselves are computed using the corresponding computations for these kernel-based approximations. Finally, the classifications are achieved by invoking the distance computations on these quantiles. In fact, a similar rule of classifying testing samples to Eq. (8) is performed with a difference which is utilizing the quantiles of KDE instead of EQs. All of these novel elements constitute some of the fundamental contributions of this paper.

5 Equivalence of NCEQ and CMQS for Large Samples

In this section, we utilize the so-called "Large Sample" theory to derive certain asymptotic results that pertain to the equivalence between NCEQ and CMQS for large samples. This theory revolves around the asymptotic properties of sample estimators, which is widely used in statistical problems. Toward this end, first of all, we recall a main theorem regarding the asymptotic distribution of EQs from [5], using which we point out the convergence phenomena of the EQs to the corresponding population quantiles[6].

Theorem 1 Let $X_1, \ldots, X_n$ be iid random variables with a cdf $F(\cdot)$ that is absolutely continuous and whose corresponding pdf is $f(\cdot)$. Also, let $k = [np] + 1$. Then, the k^{th} order statistic converges to $N(0,1)$ in distribution, such that

$$\sqrt{n} f\left(F^{-1}(p)\right) \frac{X_{(k)} - F^{-1}(p)}{\sqrt{p(1-p)}} \longrightarrow N(0,1), \tag{9}$$

where $N(0,1)$ denotes the standard normal distribution.

From Eq. (9), one can deduce that $X_{(k)}$ converges *almost surely* to $F^{-1}(p)$, which means that:

$$P\big(\lim_{n\to\infty} \left|X_{(k)} - F^{-1}(p)\right| > \epsilon\big) = 0.$$

Now, by invoking an insight into the following theorem proposed by Alfred [3], we are able to show the equivalence between both the classification schemes NCEQ and CMQS for large samples.

[6] The proof of the theorem is omitted, since it is found in the literature. Also, we refer the interested reader to [6] for more information about the various types of convergence.

Theorem 2 Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be univariate normally distributed random variables. Then,

$$X \leq_{st} Y \text{ if and only if } \mu_1 \leq \mu_2 \text{ and } \sigma_1^2 = \sigma_2^2.$$

This leads us to the main result that we have alluded to. Before we proceed, let us emphasize that as the authors of [31,21] have discussed in detail, there are two essential conditions in a binary classification that we work with that have to be imposed on the distributions of the two classes. The first one is that the distributions of both classes must be identical. These authors have further categorized these distributions, which have equal shape and scale parameters, as being identical distributions. The second basic condition is that the second class is located at the right of the first class. Based on these assumptions, we present the main result of this section in the following theorem.

Theorem 3 In a binary classification problem, assuming the scale parameters of respective classes are the same, the NCEQ is equivalent to CMQS for sufficiently large samples.

Proof. Suppose that classes ω_1 and ω_2 follow class conditional distributions F_1 and F_2, respectively. By examining Eq. (9), one can simply observe that for large sample sizes, the QSs $\hat{Q}_{12}$ and $\hat{Q}_{21}$ have approximate normal distributions as

$$N\left(F_1^{-1}\big(\frac{n-k+1}{n+1}\big), \left(\frac{(n+1)\sqrt{n} f_1\big(F_1\big(\frac{n-k+1}{n+1}\big)\big)}{\sqrt{k(n-k+1)}}\right)^{-2}\right) \tag{10}$$

and

$$N\left(F_2^{-1}\big(\frac{k}{n+1}\big), \left(\frac{(n+1)\sqrt{n} f_2\big(F_2\big(\frac{n-k+1}{n+1}\big)\big)}{\sqrt{k(n-k+1)}}\right)^{-2}\right), \tag{11}$$

respectively. As previously mentioned, in the classification problems that we work with, the scale parameters of all classes are the same, and the distributions of respective classes are assumed to be symmetric. This leads us to the following equation

$$f_1\big(F_1\big(\frac{n-k+1}{n+1}\big)\big) = f_2\big(F_2\big(\frac{n-k+1}{n+1}\big)\big)$$

which, using Eqs. (10) and (11), is equivalent to the assertion that the variances of $\hat{Q}_{12}$ and $\hat{Q}_{21}$ are the same. Moreover, using Eqs. (10) and (11), we see that the asymptotic expected values of $\hat{Q}_{12}$ and $\hat{Q}_{21}$ are $Q_{12} = F_1^{-1}\big(\frac{n-k+1}{n+1}\big)$ and $Q_{21} = F_2^{-1}\big(\frac{k}{n+1}\big)$, respectively. Hence, assuming $Q_{12} < Q_{21}$, it is deduced that the asymptotic mean of $\hat{Q}_{12}$ in (10) is less than that of $\hat{Q}_{21}$ in (11). By invoking the above in Theorem 2, we can deduce that $\hat{Q}_{12} \leq_{st} \hat{Q}_{21}$. In order to prove our claim, we now need to demonstrate that

$$D(x, \hat{Q}_{12}) \leq_{st} D(x, \hat{Q}_{21}) \Longleftrightarrow D(x, Q_{12}) < D(x, Q_{21}). \tag{12}$$

Based on almost surly convergence of $\hat{Q}_{12}$ and $\hat{Q}_{21}$ respectively to $F_1^{-1}\big(\frac{n-k+1}{n+1}\big)$ and $F_2^{-1}\big(\frac{k}{n+1}\big)$, observed from (10) and (11), it is deduced that for sufficiently large data sets, the LHS of Eq. (12) leads to

$$\begin{aligned} D(x,\hat{Q}_{12}) \leq_{st} D(x,\hat{Q}_{21}) &\Longleftrightarrow x - F_1^{-1}\big(\frac{n-k+1}{n+1}\big) < F_2^{-1}\big(\frac{k}{n+1}\big) - x \\ &\Longleftrightarrow x < \frac{F_2^{-1}\big(\frac{k}{n+1}\big) + F_1^{-1}\big(\frac{n-k+1}{n+1}\big)}{2} \\ &\Longleftrightarrow x < \frac{Q_{12}+Q_{21}}{2}, \end{aligned}$$

which is precisely the criterion that CMQS utilizes for classifying the sample point x. Otherwise, when $Q_{12} > Q_{21}$, the equivalence of the mentioned PRs can be similarly proved based on using the rules of "dual CMQS" and "dual NCEQ". Hence the proof is complete.

6 Comparative Nonparametric Results

In this section, we shall demonstrate the strong performance of the NCEQ and NCKQ methods when they are compared to the Bayesian and Anti-Bayesian (or CMQS) parametric methods. Their competitive results are illustrated by executing rigorous tests for both symmetric (including normal, logistic and Laplace) and asymmetric (including gamma, log-normal, log-logistic and rayleigh) distributions. For these experiments, we assume that there are two classes ω_1 and ω_2 with pdfs $f(x)$ and $f(x-\theta)$, respectively, where the constant θ represents the location parameter. Setting $p(x|\omega_1) = f(x)$ and $p(x|\omega_2) = f(x-\theta)$, the Bayesian discriminant function can be determined using the following rule

$$p(x|\omega_1)p(\omega_1) \geq p(x|\omega_2)p(\omega_2), \tag{13}$$

where $p(\omega_1)$ and $p(\omega_2)$ are the *a priori* distributions of classes ω_1 and ω_2, respectively. As presented in [30], we also use here the equal priors for both classes. Althogh, recently, Meegen et al. [19] have used unequal priors in linear discriminant analysis. Also, Nguyen-Trang and Vo-Van [20] suggested an algorithm to identify the prior probabilities for classification problem by Bayesian method. Furthermore, to determine the parametric CMQS method, let us use the $(\frac{2}{3})^{rd}$ quantile of the class ω_1 and $(\frac{1}{3})^{rd}$ quantile of of the class ω_2, where based on (5), they are found as

$$Q_{12} = \xi_{\frac{2}{3}} \quad \text{and} \quad Q_{21} = \theta + \xi_{\frac{1}{3}},$$

respectively, where ξ_p stands for the pth quantile of the pdf $f(\cdot)$. As previously mentioned, when $Q_{12} > Q_{21}$, according to the dual CMQS, the quantiles

$$Q_{11} = \xi_{\frac{1}{3}} \quad \text{and} \quad Q_{22} = \theta + \xi_{\frac{2}{3}}$$

are used instead of Q_{12} and Q_{21}, respectively.

Table 1 Summarized results of some distributions including Bayesian discriminant functions and quantiles.

Distribution	Notation	*pdf*	Bayesian classifier	$\xi_{\frac{1}{3}}$	$\xi_{\frac{2}{3}}$
Normal	$N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$x=\mu+\frac{\theta}{2}$	$\mu-\sigma\Phi^{-1}(\frac{2}{3})$	$\mu+\sigma\Phi^{-1}(\frac{2}{3})$
Logistic	$Logis(\mu,\sigma)$	$\frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma\left(1+e^{-\frac{x-\mu}{\sigma}}\right)^2}$	$x=\mu+\frac{\theta}{2}$	$\mu-\sigma\log(2)$	$\mu+\sigma\log(2)$
Laplace	$Lap(\mu,\sigma)$	$\frac{1}{2\sigma}e^{-\frac{\lvert x-\mu\rvert}{\sigma}}$	$x=\mu+\frac{\theta}{2}$	$\mu+\sigma\log(\frac{2}{3})$	$\mu-\sigma\log(\frac{2}{3})$
Gamma	$\Gamma(\alpha,\beta)$	$\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$	$x=\theta\{1-e^{-\frac{\theta}{\beta(\alpha-1)}}\}^{-1}$	quantile for $\Gamma(2,1)$: $-\xi_{\frac{1}{3}}+\log(\xi_{\frac{1}{3}}+1)=\log(\frac{2}{3})$	quantile for $\Gamma(2,1)$: $-\xi_{\frac{2}{3}}+\log(\xi_{\frac{2}{3}}+1)=\log(\frac{1}{3})$
Log-normal	$LN(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$	$\log(\frac{x-\theta}{x})=$ $\frac{-(\log(x-\theta)-\mu)^2}{2\sigma^2}+\frac{-(\log x-\mu)^2}{2\sigma^2}$	$\exp\{\mu-\frac{\sigma}{\sqrt{2\pi}}\}$	$\exp\{\mu+\frac{\sigma}{\sqrt{2\pi}}\}$
Log-logistic	$LL(\alpha,\beta)$	$\frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{-\beta})^2}$	$(\beta-1)\log\left(\frac{x-\theta}{x}\right)=4\log\left(\frac{\alpha^{\beta}+(x-\theta)^{\beta}}{\alpha^{\beta}+x^{\beta}}\right)$	$\alpha 2^{1/\beta}$	$\alpha 2^{-1/\beta}$
Rayleigh	$Ray(\sigma)$	$\frac{x}{\sigma^2}e^{-x^2/2\sigma^2}$	$\log\frac{x}{x-\theta}=\frac{-\theta^2+2\theta x}{2\sigma^2}$	$\sigma\sqrt{2\log(\frac{3}{2})}$	$\sigma\sqrt{2\log(3)}$

All distributions studied in this section, as well as their pdfs, the Bayesian discriminant functions and quantiles used in the CMQS mechanism are summarized in Table 1. In this table, the quantile function of the normal distribution is denoted by $\Phi^{-1}(\cdot)$. Some of the reported details in this table can be found in [15,21].

To compare the precision of the different classification methods for a binary classification problem, we generated two data sets, each of size 800, from classes ω_1 and ω_2. The accuracy of each algorithm was obtained by testing it 10 times, each invoking a 10-fold cross-validation mechanism. The results obtained are tabulated in Table 2 for both cases of symmetric and asymmetric distributions. In this table, we present the results for various values of θ which serve to displace the classes sufficiently for the various different distributional shapes. By examining the results of these experiments, one will observe that the accuracies of the proposed schemes are almost as high and efficient as their parametric versions, even though the distributions of the data are considered to be unknown.

Table 2 Numerical values of the Precision for different classification methods.

		Classification Method			
Distribution	θ	*Bayesian*	*CMQS*	*NCEQ*	*NCKQ*
$N(0,3)$	6	83.62	83.62	83.50	83.50
$Logis(0,1)$	2	87.60	87.60	87.50	87.50
$Lap(0,3)$	10	90.00	90.12	90.00	90.03
$\Gamma(2,1)$	4.5	95.40	95.06	94.67	94.67
$LN(0,1.5)$	7	86.57	88.87	87.50	80.00
$LL(0,1)$	3	79.37	72.56	74.45	74.40
$Ray(2)$	3	88.01	87.06	86.98	86.99

From Table 2, one easily observes that the proposed nonparametric classification methods perform almost as well as Bayesian classifier, while in some cases, the parametric CMQS method is unable to attain the optimality or near-optimality of the Bayesian classifier[7]. In some cases, such as for the Normal distribution, the parametric CMQS is more accurate than the NCEQ and NCKQ, although there is a negligible difference between the accuracies. In addition, the proposed nonparametric methods, which utilize EQs or KDE, do not depend on the knowledge of the distribution which makes it easier for the user to compute the discriminant function without paying attention to the *structure* of the distribution.

[7] It is pertinent to mention that the accuracy of *any* classifier can and will never exceed that of a Bayesian classifier. The amazing thing is that we have been able to attain to an accuracy quite close to the optimal, even though we have worked in a counter-intuitive manner, and also made no assumption about the underlying distribution!

7 The Robustness of Nonparametric Methods against Outliers

In real-world scenarios, sometimes data sets appear to contain "outliers". Outliers have various definitions depending on the structure of the data. One of the most generalized definitions is based on Hawkins' perspective [13], when he states that "an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Barnett and Lewis [8] have defined the outlier in a set of data to be "an observation or subset of observations which appears to be inconsistent with the remainder of the set of data".

To handle the scenario in which the data contains outliers, we formalize the above perspectives and model them by considering a mixture probability model. In general, we assume that $X_1, \ldots, X_n$ are iid observations from a population with the pdf

$$f(x) = \sum_{i=1}^{m} p_i f_i(x),$$

where m is the number of components in the mixture, $f_i(\cdot)$ is the pdf of the i^{th} component, and the $\{p_1, \ldots, p_m\}$ are the so-called "mixing" weights, such that $p_i \geq 0$ and $\sum_{i=0}^{n} p_i = 1$. While outlier detection has been investigated by many authors, Aitkin and Wilson [2] identified outliers in single sample using mixture models[8].

In what follows, we consider a two-component mixture model in which one of the components, with a large mixing weight represents the majority of the observations, while the other, with a small mixing weight, represents the minority of the observations. Observe that the second component may have different parameters from the first. Specifically, we assume that the data comes from a population with the pdf:

$$f(x) = pf_1(x; \boldsymbol{\theta}_1) + (1-p)f_2(x; \boldsymbol{\theta}_2), \quad 0 \leq p \leq 1, \tag{14}$$

where $\boldsymbol{\theta}_i$ $(i = 1, 2)$ represents the parameter vector of the i^{th} component.

In parametric CMQS, the data is assumed to follow an exact distribution which is often chosen to be Normal, because of its generality. In fact, since the exact pdf for a given real-world data set is not known, in working with parametric methods such as CMQS, it is usual to assume a Normal distribution to the data. But, in practice, since the data may not exactly come from a Normal distribution, the practitioner encounters various problems due to such an assumption. For instance, since data containing outliers follows a mixture model, a data set which is suspected to contain outliers does not lend itself to be compatible with the common methods. Indeed, in such cases, it is impossible to fit an appropriate distribution since it is, in actuality, a mixture of different distributions. Further, on the other hand, there is no consideration to accommodate the process of classifying outlying data points within the family

[8] For more details about outliers in statistical analysis, we refer the reader to [8, 16, 25, 27].

of parametric CMQS schemes. This is why proposing methods that are robust against outliers (specially when one works with real-life data) is essential. In these scenarios, using a nonparametric method is a safer and more stable way to resolve the classification problem. This is because of the presence of noise and outliers in real-world data sets.

With this in mind, we quantify the efficiency and robustness of NCEQ and NCKQ, and compare them with the parametric CMQS method. Toward this end, we invoke a binary classification method assuming that the pdf of the classes ω_1 and ω_2 are assumed to be:

$$f_{\omega_1}(x) = pf_1(x;\boldsymbol{\theta}_1) + (1-p)f_2(x;\boldsymbol{\theta}_2), \quad 0 \leq p \leq 1, \tag{15}$$

and

$$f_{\omega_2}(x) = pf_1(x-\gamma_1;\boldsymbol{\theta}_1) + (1-p)f_2(x-\gamma_2;\boldsymbol{\theta}_2), \quad 0 \leq p \leq 1, \tag{16}$$

respectively, where γ_1 and γ_2 are the constants representing the displacement of ω_2 with respect to ω_1.

To test the schemes and to demonstrate the superior performance of NCEQ and NCKQ in comparison with the parametric versions, we first generated 800 random data points from each mixture pdfs given in Eqs. (15) and (16) with $p = 0.75$ and various choices of skew pdfs f_1 and f_2. We then classified them using the parametric CMQS and the presently-proposed nonparametric schemes. In fact, after generating the data, in all the settings, we assumed that the distributions were unknown, and the data sets were treated in identical manners as one would do when one encountered real-life data sets in which the exact distributions were unknown. Thus, the results of the parametric CMQS were obtained by assuming that the data followed a Normal distribution possessing the mean and variance of the sample points [32]. Every algorithm was executed 10 times using a 10-fold cross-validation scheme. The results of the experiments are presented in Table 4, which is merely a summary of the results we have obtained for numerous experiments, but where, in the interest of brevity, we have only cited a few typical examples. In Table 4, $Exp(\theta)$ and $Pareto(\alpha, \beta)$ respectively stand for the exponential and Pareto distributions with the pdfs

$$f(x) = \theta e^{-\theta x}, \quad x > 0, \quad \theta > 0$$

and

$$f(x) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}, \quad \beta \leq x < +\infty, \quad \alpha > 0, \quad \beta > 0,$$

respectively. Note that $Exp(\theta)$ is equivalent to $\Gamma(1, \theta)$ distribution.

From the results given in Table 4, we deduce that the proposed nonparametric PR schemes, NCEQ and NCKQ, are more efficient than the parametric method in classifying data sets which have outlying points. In other words, when there are outliers in the data, one can safely assume that, as is done in the literature [2], the data follows a mixture model. In such situations, nonparametric methods are preferable due to the fact that they are distribution-free. This superiority is demonstrated by the fact that in Table 4, the NCEQ and

Table 3 Precision of the CMQS, NCEQ and NCKQ approaches in various mixture models.

No.	f_1	γ_1	f_2	γ_2	$CMQS$	$NCEQ$	$NCKQ$
1	$Exp(1)$	3	$LL(1,1)$	20	59.62	85.00	90.25
2	$Exp(2)$	3	$LL(1,1)$	20	58.94	90.19	92.81
3	$Exp(3)$	3	$LL(1,1)$	20	58.75	91.81	94.50
4	$Pareto(1,2)$	3	$LL(1,1)$	20	60.19	86.81	86.83
5	$Pareto(2,3)$	3	$LL(1,1)$	20	60.81	77.69	78.00
6	$Pareto(2,5)$	3	$LL(1,1)$	20	61.56	77.81	77.90
7	$LN(0,0.5)$	2	$LL(1,1)$	3	59.19	89.56	93.44
8	$LN(0,1.5)$	2	$LL(1,1)$	3	50.56	89.06	89.50
9	$LN(0,2)$	2	$LL(1,1)$	3	50.00	76.62	76.00
10	$Pareto(1,2)$	3	$\Gamma(2,2)$	20	66.81	93.00	95.75
11	$Pareto(2,3)$	3	$\Gamma(2,1)$	20	67.37	80.20	83.87
12	$Pareto(2,5)$	3	$\Gamma(2,1)$	20	68.44	78.00	84.75

NCKQ yield much more accurate results than the CMQS. For example, for the experiment No. 11, the NCEQ and NCKQ yield 80.20% and 83.87% accuracies, respectively. As opposed to this, the CMQS yields an accuracy of only 67.37%, which is clearly demonstrated that for such outlier-ridden data, the currently-proposed non-parametric schemes are superior. Indeed, the performance of these new methods is clear. Additionally, it is obvious that in most cases, nonparametric classification by invoking kernel-based quantiles yields to an even higher efficiency than the NCEQ. The reason behind this result is because of the inclusion of additional information about the distributions in NCKQ, while NCEQ works only with the sample points and does not have any information about the distributions imposed by the kernels.

8 Testing on Real-life Data

Although the results presented in the previous section support a *prima facie* case for our analytic results, to further illustrate the procedure proposed in this paper, we have tested them on a real-life data set involving the breast cancer data which was created by the authors of [23], at the Faculty of Medicine of the University of Coimbra[9]. The data set had only 116 instances, rendering it particularly interesting because the sample size was so small. Thus, if one resorted to a *parametric* training scheme, we believe that the corresponding covariance matrices would have been singular. A histogram-based nonparametric scheme would have also yielded most bins to be empty, which is why, we believe, that our nonparametric AB method is pertinent.

In this data set, there were ten predictors, all of which were quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. In what follows, we explain how we carried out the procedure only for the predictor Glucose (mg/dL), while the dependent variable was labeled as "Healthy" or "Cancerous".

[9] The data may be obtained from the UCI Repository of Machine Learning databases at archive.ics.uci.edu/ml.

To evaluate the accuracy and precision of the various methods, in all the experiments we randomly choose 93 instances (approximately 80 percent of the total data set) and considered them as the "Training" set, while the remaining 23 instances were used as the "Testing" set. Based on the Training set, the quantiles of order $\frac{1}{3}$ and $\frac{2}{3}$ were then obtained for both the classes by invoking the three methods mentioned in the previous sections. The quantiles based on the *parametric* CMQS method were determined by fitting a normal distribution on the data elements of each class. On the other hand, the EQs utilized in the nonparametric NCEQ method, were easily obtained from Eq. (6). Finally, the quantiles used in the NCKQ method were derived by inverting the kernel estimate of the cdfs of each class. The KDEs of both classes are presented in Figure 2, from which one observes that the pdf of the Class ω_2 can be considered to be a mixture probability model, as in Eq. (14).

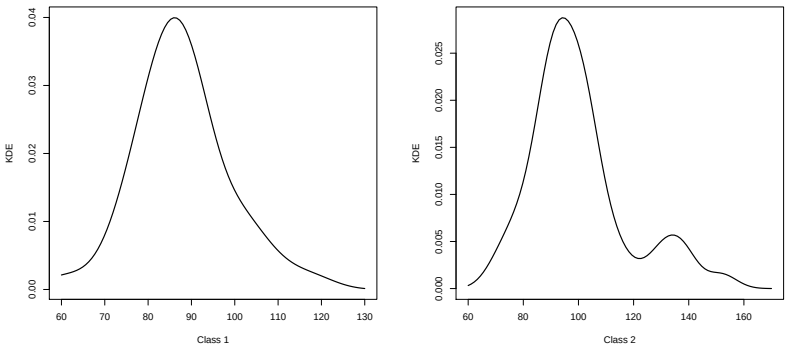

Fig. 2 Estimates of the kernel density functions for Class ω_1 and Class ω_2.

To appropriately affect the nonparametric computations, we fitted the following mixture model to the data in Class ω_2:

$$f(x) = 0.82\phi(x; 95, 10) + 0.18\phi(x; 158, 32), \tag{17}$$

where $\phi(x; \mu, \sigma)$ represents the normal distribution with mean μ and standard deviation σ. The value of the test statistic for the Kolmogorov-Smirnov test was obtained as 0.0512 with the corresponding p-value to equal 0.999. Based on these observations, the pdf described by Eq. (17) was suitable to fit the data in Class ω_2.

The quantiles obtained for the various methods are summarized in Table 4. This table also includes the precision of the methods obtained by classifying the instances of the Testing set into the "Healthy" or "Cancerous" classes. From these results, one can observe that the precision of the nonparametric methods (73.91%) is markedly higher than that of the parametric (69.56%).

This appears remarkable, but it can, well, be justified. Indeed, this is because, in the parametric setup, one would assume a Normal distribution for the data elements of each class, whereas, one sees that a mixture probability model is better suited for the data in Class ω_2. More precisely, using the assumption of a Normal distribution for the data in this class, would be erroneous, and lead to faulty indicators, inferences and conclusions. As opposed to this, the nonparametric methods do not rely on any such assumption, and yields more genuine results than the parametric ones.

Table 4 The quantiles and the corresponding precision values for the CMQS, NCEQ and NCKQ approaches.

	Quantiles of Class ω_1		Quantiles of Class ω_2		
Method	1/3	2/3	1/3	2/3	Precision
$NCKQ$	82.7505	91.5221	91.9627	104.6955	0.7391
$NCEQ$	84	90	92	103	0.7391
$CMQS$	83.1658	92.3691	93.2437	118.3963	0.6956

To demonstrate the key contribution of this paper, the reader must observe that this classification was done only on the basis of a *single* feature. It is also particularly important and pertinent to point out that the pdf of the second class (i.e., of ω_2) in the present data set, emerged as a result of a *mixture* model involving Normal distributions, while the data in the first class (i.e., of ω_1) was Normally distributed. Observe that, in practice, other distributions or mixture models could have been utilized to fit for the data, just as well.

9 Conclusions

For decades, parametric and nonparametric PR have been achieved using a Bayesian paradigm, which reduces to comparing the testing sample to central descriptors of the respective distributions. In this paper, we have pursued a recently-introduced pioneering and non-intuitive paradigm (the "Anti-Bayesian" paradigm) of achieving the PR by comparing the testing sample with quantile points, which could also be quite distant from the mean. It essentially uses the quantiles of the distributions to achieve the PR, and they attain the optimal Bayesian accuracy for symmetric distributions even though they operate with a non-intuitive philosophy. This paper explained, in detail, how such quantile-based PR can be extended to the nonparametric world, using both traditional and kernel-based strategies. From a fundamental perspective, the paper has also used the so-called "Large Sample" theory to derive strong asymptotic results that pertain to the equivalence between the parametric and nonparametric paradigms for large samples. Further, the paper analyzed the methodology of such nonparametric schemes and their robustness in the presence of outliers modeled using a mixture distribution. As far as we know, these are the first-reported results within such a nonparametric domain, and which merge the results from both the paradigms.

Apart from deriving the analytic results, the paper also included simulation results and the results by testing the methods on a real-life breast cancer data set gathered at the University Hospital Centre of Coimbra.

Using the schemes described in [32] amd [34], we believe that the results of this paper can be extended to multidimensional classification problems. However, this is currently open since it involves all the three nonparametric estimation strategies that we have proposed here.

Acknowledgements We are very grateful to the anonymous Referees of the previous version of the paper, who suggested various modifications and changes. Their suggestions have greatly enhanced the quality of this present version.

References

1. Ahsanullah, M. and Nevzorov, V. B., (2005). *Order statistics: Examples and Exercises*. Nova Publishers.
2. Aitkin, M. and Wilson, G. T., (1980). Mixture models, outliers, and the EM algorithm. *Technometrics*, 22 (3), 325–331.
3. Alfred, M., (2001). Stochastic ordering of multivariate normal distributions. *Annals of the Institute of Statistical Mathematics*, 53 (3), 567–575.
4. Altman, N. and Léger, C., (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46 (2), 195–214.
5. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N., (2008). *A First Course in Order Statistics*. SIAM, Philadelphia.
6. Athreya, K. B. and Lahiri, S. N., (2006). *Measure theory and probability theory*, Springer Science & Business Media.
7. Azzalini, A., (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68 (1), 326–328.
8. Barnett, V. and Lewis, T., (1978). *Outliers in Statistical Data*. John Wiley & Sons.
9. Binder, D. A., (1978). Bayesian cluster analysis. *Biometrika*, 65 (1), 31–38.
10. Bishop, C. M., (2006). *Pattern Recognition and Machine Learning*, (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.
11. David, H. A. and Nagaraja, H. N., (2004). *Order Statistics*. John Wiley & Sons.
12. Hammer, H., Yazidi, A. and Oommen, B. J., (2017). "Anti–Bayesian" Flat and Hierarchical Clustering Using Symmetric Quantiloids. *Information Sciences*, Vol. 418-419, 495-512.
13. Hawkins, D. M., (1980). *Identification of Outliers*. London: Chapman and Hall.
14. Hollander, M., Wolfe, D. A., and Chicken, E., (2013). *Nonparametric statistical methods*. John Wiley & Sons.
15. Hu, L., (2015). A note on order statistics-based parametric pattern classification. *Pattern Recognition*, 48 (1), 43–49.
16. Huber, P. J., (2011). Robust statistics. In: *International Encyclopedia of Statistical Science*, pp. 1248–1251.
17. Kothari, C. R., (2004). *Research Methodology: Methods and Techniques*. New Age International.
18. Leech, N. L. and Onwuegbuzie, A. J., (2002). A call for greater use of nonparametric statistics. In: *Mid-South Educational Research Association Annual Meeting*.
19. Meegen, C. V., Schnackenberg, S. and Ligges, U. (2019). Unequal Priors in Linear Discriminant Analysis. *Journal of classifications*, https://doi.org/10.1007/s00357-019-09336-2.
20. Nguyen-Trang, T. and Vo-Van, T. (2017). A new approach for determining the prior probabilities in the classification problem by Bayesian method. *Advances in Data Analysis and Classification*, vol 11, pp. 629–643.

21. Oommen, B. J. and Thomas, A., (2014). Optimal order statistics-based "Anti-Bayesian" parametric pattern classification for the exponential family. *Pattern Recognition*, 47, 40–55.
22. Oommen, B. J., Khoury, R. and Schmidt, A., (2015). Text classification using novel "Anti-Bayesian" techniques. In: *Nunez M., Nguyen N., Camacho D., Trawinski B. (eds) Computational Collective Intelligence. Lecture Notes in Computer Science*, 9329, 1–15.
23. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R. and Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, 18, 29.
24. Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27 (3), 832–837.
25. Rousseeuw, P. J. and Leroy, A. M., (2005). *Robust Regression and Outlier Detection*, John wiley & sons.
26. Santhanam, V., Morariu, V. I., Harwood, D., and Davis, L. S., (2016). A non-parametric approach to extending generic binary classifiers for multi-classification. *Pattern Recognition*, 58, 149–158.
27. Scott, D. W., (2004). Partial mixture estimation and outlier detection in data and regression. In: *Hubert M., Pison G., Struyf A., Van Aelst S. (eds) Theory and Applications of Recent Robust Methods. Statistics for Industry and Technology*, pp. 297–306.
28. Serfling, R. J., (2009). *Approximation theorems of mathematical statistics*, John Wiley & Sons.
29. Shaked, M. and Shanthikumar, J. G., (2007). *Stochastic orders*. Springer Science & Business Media.
30. Thomas, A. and Oommen, B. J., (2012). Optimal "Anti-Bayesian" parametric pattern classification for the exponential family using order statistics criteria. In: *Alvarez L., Mejail M., Gomez L., Jacobo J. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012. Lecture Notes in Computer Science*, vol 7441, 1–13.
31. Thomas, A. and Oommen, B. J., (2013)a. The fundamental theory of optimal "Anti-Bayesian" parametric pattern classification using order statistics criteria. *Pattern Recognition*, 46 (1), 376–388.
32. Thomas, A., Oommen, B. J., (2013)b. Order statistics-based parametric classification for multi-dimensional distributions. *Pattern Recognition*, 46 (12), 3472–3482.
33. Thomas, A., Oommen, B. J., (2014). Corrigendum to three papers that deal with "Anti-Bayesian" pattern recognition. *Pattern Recognition*, 47 (6), 2301–2302.
34. Thomas, A., and Oommen, B. J., (2013). Ultimate order statistics-based Prototype Reduction Schemes. In: *Cranefield S., Nayak A. (eds) AI 2013: Advances in Artificial Intelligence. AI 2013. Lecture Notes in Computer Science*, vol 8272, pp. 421–433.
35. Thomas, A. and Oommen, B. J., (2013). A novel Border Identification algorithm based on an "Anti-Bayesian" paradigm. In: *Wilson R., Hancock E., Bors A., Smith W. (eds) Computer Analysis of Images and Patterns. CAIP 2013. Lecture Notes in Computer Science*, vol 8047, pp. 196–203.