

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

Directional Quantile Classifiers

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Farcomeni, A., Geraci, M., Viroli, C. (2022). Directional Quantile Classifiers. JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS, 31(3), 907-916 [10.1080/10618600.2021.2021209].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/947774> since: 2023-11-06

*Published:*

DOI: <http://doi.org/10.1080/10618600.2021.2021209>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

# Directional quantile classifiers

Alessio Farcomeni

Department of Economics and Finance, University of Rome “Tor Vergata”

Marco Geraci

MEMOTEF Department, Sapienza University of Rome

Department of Epidemiology and Biostatistics, University of South Carolina

and

Cinzia Viroli

Department of Statistical Sciences, University of Bologna

December 17, 2021

## Abstract

We introduce classifiers based on directional quantiles. We derive theoretical results for selecting optimal quantile levels given a direction, and, conversely, an optimal direction given a quantile level. We also show that the probability of correct classification of the proposed classifier converges to one if population distributions differ by at most a location shift and if the number of directions is allowed to diverge at the same rate of the problem’s dimension. We illustrate the satisfactory performance of our proposed classifiers in both small and high dimensional settings via a simulation study and a real data example. The code implementing the proposed methods is publicly available in the R package `Qtools`.

*Keywords:*  $L_1$  distance, Supervised Classification, Quantiles for multivariate data.

# 1 Introduction

The idea of using quantiles in classification is relatively recent and largely unexplored. The median classifier for high-dimensional problems proposed by Hall et al. (2009), which calculates the  $L_1$  distance of the coordinates of a multivariate data point from componentwise medians (rather than centroids), is particularly advantageous when data exhibit heavy-tailed or skewed distributions. Building on Hall et al.’s (2009) idea, Hennig and Viroli (2016a) proposed quantile classifiers which hinge on the sum of distances from componentwise quantiles at some generic level  $\theta \in (0, 1)$ . The ensemble quantile classifier by Lai and McLeod (2020) assigns weights to the componentwise distances by minimising a regularised loss function, where the regularisation parameter is determined by cross-validation.

In all the studies mentioned above, quantiles are calculated marginally for each input variable (componentwise). This implies that their calculation ignores the possible interdependence among variables. In this study, we consider directional quantiles for multivariate distributions (Kong and Mizera, 2012) to address such a limitation. Our choice is motivated by several reasons. First, as already mentioned, the dependence among variables is taken into account by computing linear combinations of input variables. Second, directional quantiles have a simple interpretation since the projections’ weights embody the relative importance of the variables involved in the classification problem. Finally, in the special case of  $p$  canonical directions (with  $p$  equal to the number of variables), the use of directional quantiles leads to the componentwise quantile classifier (Hennig and Viroli, 2016a), and thus inherits asymptotic optimal properties as shown in Appendix. The search of a directional quantile, it being based on a linear combination of variables, could be seen as a projection pursuit problem (Lee et al., 2005). However, directional quantiles require the search of the best projection associated with one, or more, optimal percentiles, thus making the problem particularly challenging. Directional quantiles have already found application in risk classification problems (Geraci et al., 2020) and proved to be a worthwhile alternative to risk classification based on componentwise quantile thresholds.

In general, the application of our methods does not require any assumption on the shape of the population distributions. We derive asymptotic theoretical properties of the proposed classifier, under the assumption that distributions for alternative populations differ by at most a location-

shift. While this assumption may be unrealistic in practice, empirical results support the merit of the proposed classifier also when the distributions differ by shape and not just by location.

The rest of the paper is organised as follows. In the next section, we introduce notation and basic definitions, followed by our proposal of directional quantile classifiers. Theoretical results are stated in Section 3. We report the results of a simulation study in Section 4 and of a real data analysis in Section 5. Concluding remarks are given in Section 6. All proofs of theoretical results are reported in Appendix A. A software implementation of our approach can be found in the package `Qtools` (Geraci, 2016), freely available on the Comprehensive R Archive Network (R Core Team, 2020).

## 2 Methods

### 2.1 Notation and definitions

Let  $\mathbf{X}^{(1)} = (X_1^{(1)}, X_2^{(1)}, \dots, X_p^{(1)})^\top$  and  $\mathbf{X}^{(2)} = (X_1^{(2)}, X_2^{(2)}, \dots, X_p^{(2)})^\top$  denote two  $p$ -variate random variables with absolutely continuous distributions  $F^{(1)}$  and  $F^{(2)}$  defined on the same space  $\mathcal{X} \subseteq \mathbb{R}^p$  for two populations  $\Pi^{(1)}$  and  $\Pi^{(2)}$ , respectively. The marginal distributions of the components of  $\mathbf{X}^{(k)}$  are denoted by  $F_j^{(k)}$ , for  $j = 1, 2, \dots$  and  $k = 1, 2$ . Further,  $I(\cdot)$  denotes the indicator function which is equal to 1 if its argument is true, and 0 otherwise.

Our goal is to assign a new observation  $\mathbf{y} = (y_1, y_2, \dots, y_p)^\top$  to either  $\Pi^{(1)}$  or  $\Pi^{(2)}$  according to how *close* the point is to one or the other. In quantile-based classification (Hennig and Viroli, 2016a), the distance is first calculated for each component of  $\mathbf{y}$  using the asymmetrically weighted loss function

$$\Phi^{(k)}(\theta; y_j) = \{\theta + (1 - 2\theta)I(y_j - Q_{X_j}^{(k)}(\theta) < 0)\}|y_j - Q_{X_j}^{(k)}(\theta)| \quad (1)$$

for  $j = 1, 2, \dots, p$  and  $k = 1, 2$ , where  $Q_{X_j}^{(k)}(\theta)$  is the componentwise quantile at level  $\theta \in (0, 1)$  for the  $k$ th population, which can be obtained by inversion of  $F_j^{(k)}$ . Subsequently,  $\mathbf{y}$  is assigned to  $\Pi^{(1)}$  if the discrepancy

$$d(\mathbf{y}, \theta) = \sum_{j=1}^p \{\Phi^{(2)}(\theta; y_j) - \Phi^{(1)}(\theta; y_j)\} \quad (2)$$

is positive, and to  $\Pi^{(2)}$  otherwise. The quantile classifier reduces to the componentwise median classifier of Hall et al. (2009) for  $\theta = 0.5$ . An extension of (2) to more than two populations is straightforward.

The classification rule based on (2) does not acknowledge the possible interdependence among the variables, since quantiles are obtained marginally for each variable. We address this limitation by using directional quantiles for multivariate data (Kong and Mizera, 2012). We now explain our idea informally and, in the next section, give a rigorous treatment.

Define  $\mathbf{u}$  to be a vector with unit norm in  $\mathbb{R}^p$ . Throughout this paper, our focus will be on the *projected* random variables  $\mathbf{u}^\top \mathbf{X}^{(k)} \equiv Z^{(k)}$ ,  $k = 1, 2$ , defined on  $\mathcal{Z} \subseteq \mathbb{R}$ . By assumption, the  $Z^{(k)}$ 's are continuous. We denote the corresponding distribution and density functions with  $G^{(k)}(\cdot; \mathbf{u})$  and  $g^{(k)}(\cdot; \mathbf{u})$ , respectively.

Our goal is to develop a classifier where the quantities in (1) are opportunely redefined on the corresponding *projections* along  $\mathbf{u}$  to capture the multivariate nature of the distributions, namely

$$\Phi^{(k)}(\theta; \mathbf{u}^\top \mathbf{y}) = \{\theta + (1 - 2\theta)I(\mathbf{u}^\top \mathbf{y} - Q_X^{(k)}(\theta; \mathbf{u}) < 0)\}|\mathbf{u}^\top \mathbf{y} - Q_X^{(k)}(\theta; \mathbf{u})| \quad (3)$$

for  $k = 1, 2$ , where  $Q_X^{(k)}(\theta; \mathbf{u}) \equiv Q_{\mathbf{u}^\top \mathbf{X}}^{(k)}(\theta)$  is the  $\theta$ th quantile of  $Z^{(k)}$ . The latter is obtained by inverting  $G^{(k)}$  and it can be recognised as the  $\theta$ th *directional quantile* of  $\mathbf{X}^{(k)}$  in the direction  $\mathbf{u}$  (Kong and Mizera, 2012).

By working with projections, we basically summarise a multivariate problem as a univariate one. Clearly, one difficulty to address is how many and which directions should be considered. To this end, we should note that not all the directions are equally useful for classification. To exemplify, consider Figure 1, which depicts bivariate normal samples from two independent populations centered at (1,1) and (3,3), respectively, and same variance. We want to assign the new observation  $\mathbf{y} = (1.3, 3.4)^\top$  to one of the two populations. The log-density at  $\mathbf{y}$  of two bivariate normal distributions with sample means and covariance matrices separately estimated from the two samples, is  $-8.8$  and  $-5.7$ , respectively. This suggests that  $\mathbf{y}$  has been generated more likely from  $F_2$  than from  $F_1$ .

Now compute  $\Phi^{(k)}(0.9; \mathbf{u}^\top \mathbf{y})$ ,  $k = 1, 2$ , as in (3) for four normalised directions. The results are reported in Table 1. Based on a principle of minimum distance, we assign  $\mathbf{y}$  to  $F_2$ , thus consistently with a maximum likelihood principle, for three, though not all four, directions.

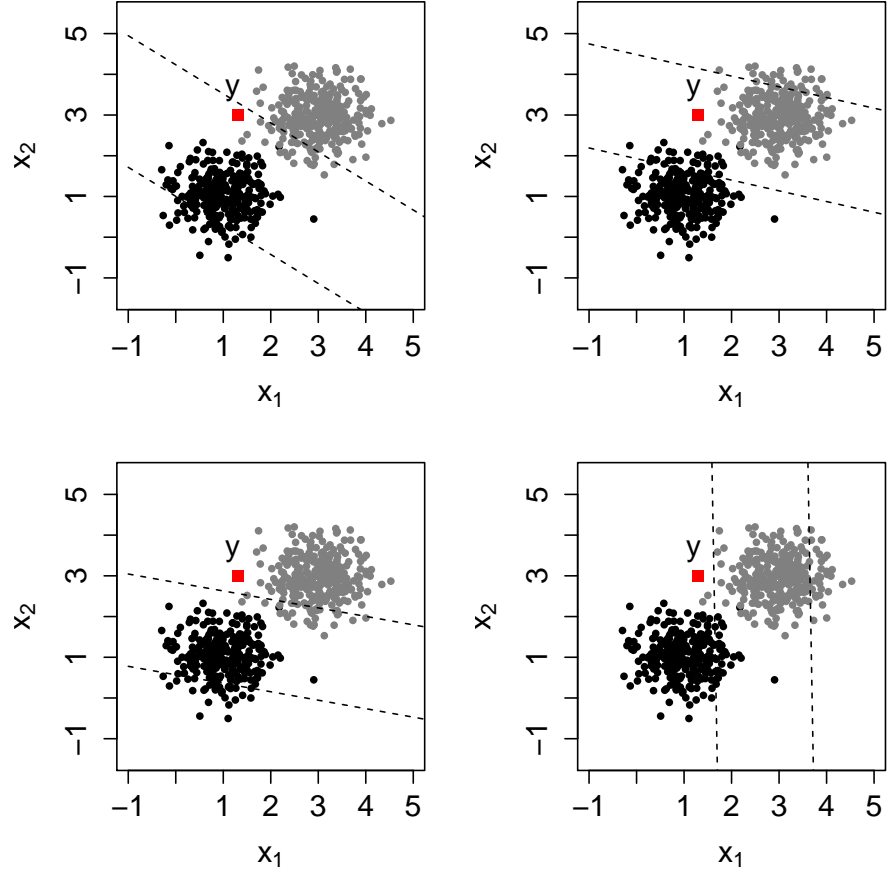


Figure 1: Simulated data depicting bivariate normal samples from two independent distributions (black and grey dots). The filled squares (labelled ‘y’) mark the point with coordinates (1.3, 3.4), while dashed lines mark directions.

$\mathbf{u}^\top$	$\Phi^{(1)}$	$\Phi^{(2)}$
$(-0.58, -0.81)$	0.27	0.01
$(0.25, 0.97)$	1.58	0.07
$(-0.20, -0.98)$	0.30	0.08
$(1.00, 0.02)$	0.03	0.23

Table 1: Distance  $\Phi^{(k)}(0.9; \mathbf{u}^\top \mathbf{y})$ ,  $k = 1, 2$ , calculated for simulated data using four different directions  $\mathbf{u}$ .

## 2.2 Directional quantile classifier

Let  $\vartheta = \{\theta_1, \theta_2, \dots, \theta_R\}$  be a set of  $R$  distinct quantile levels on  $(0, 1)$ . Also, define the set  $v_r = \{\mathbf{u}_{r1}, \mathbf{u}_{r2}, \dots, \mathbf{u}_{rS_r}\}$  containing  $S_r$  normalised directions associated with  $\theta_r$ ,  $r = 1, \dots, R$ , and let  $v = \{v_1, v_2, \dots, v_R\}$ . (Note that for convenience one may set  $S_r = S$  for  $r = 1, \dots, R$ .)

As mentioned in the previous section, we need to be wary of particular directions that may lead us to a classification error. Therefore, we introduce weights  $\omega_{rs}$  associated with each direction  $\mathbf{u}_{rs}$  to decrease (or increase) their relative importance. Let  $\boldsymbol{\omega} = (\omega_{11}, \dots, \omega_{1S_1}, \dots, \omega_{RS_R})^\top$  denote the vector of all such weights. We propose the discrepancy

$$d(\mathbf{y}, \vartheta, v, \boldsymbol{\omega}) = \sum_{r=1}^R \sum_{s=1}^{S_r} \omega_{rs} \{\Phi^{(2)}(\theta_r; \mathbf{u}_{rs}^\top \mathbf{y}) - \Phi^{(1)}(\theta_r; \mathbf{u}_{rs}^\top \mathbf{y})\}, \quad (4)$$

where  $\Phi^{(k)}$  is defined in (3). Then our *directional quantile classifier* (DQC) assigns the observation  $\mathbf{y}$  to  $\Pi^{(1)}$  if  $d(\mathbf{y}, \vartheta, v, \boldsymbol{\omega}) > 0$ , or to  $\Pi^{(2)}$  otherwise. Note that if  $R = 1$ ,  $S_r = p$ ,  $\omega_{rs} = 1$ , and  $v = \{e_1, e_2, \dots, e_p\}$  the standard basis in  $\mathbb{R}^p$ , then (4) reduces to (2).

A difficulty associated with the calculation of (4) is the selection of quantile levels, directions, and weights in the training data, say  $\mathbf{x}$ , that give the best performance on the test data, say  $\mathbf{y}$ . For some prior probabilities  $\pi_1$  and  $\pi_2$ , let

$$\begin{aligned} \psi(\mathbf{x}, \vartheta, v, \boldsymbol{\omega}) &= \pi_1 \int_{\mathcal{X}} I\{d(\mathbf{x}, \vartheta, v, \boldsymbol{\omega}) > 0\} dF^{(1)}(\mathbf{x}) \\ &\quad + \pi_2 \int_{\mathcal{X}} I\{d(\mathbf{x}, \vartheta, v, \boldsymbol{\omega}) \leq 0\} dF^{(2)}(\mathbf{x}) \end{aligned} \quad (5)$$

denote the population probability of correct classification by the DQC. Note that maximising (5) is equivalent to minimising the theoretical misclassification rate. For any given level  $\theta$  and direction  $\mathbf{u}$ , the optimal misclassification rate is obtained when

$$\pi_1 \int_{\mathcal{X}} \Phi^{(1)}(\theta; \mathbf{u}^\top \mathbf{x}) dF^{(1)}(\mathbf{x}) < \pi_1 \int_{\mathcal{X}} \Phi^{(2)}(\theta; \mathbf{u}^\top \mathbf{x}) dF^{(1)}(\mathbf{x})$$

and

$$\pi_2 \int_{\mathcal{X}} \Phi^{(2)}(\theta; \mathbf{u}^\top \mathbf{x}) dF^{(2)}(\mathbf{x}) < \pi_2 \int_{\mathcal{X}} \Phi^{(1)}(\theta; \mathbf{u}^\top \mathbf{x}) dF^{(2)}(\mathbf{x}),$$

which is equivalent to minimise

$$\begin{aligned} &\pi_1 \int_{\mathcal{X}} \{\Phi^{(1)}(\theta; \mathbf{u}^\top \mathbf{x}) - \Phi^{(2)}(\theta; \mathbf{u}^\top \mathbf{x})\} dF^{(1)}(\mathbf{x}) \\ &\quad + \pi_2 \int_{\mathcal{X}} \{\Phi^{(2)}(\theta; \mathbf{u}^\top \mathbf{x}) - \Phi^{(1)}(\theta; \mathbf{u}^\top \mathbf{x})\} dF^{(2)}(\mathbf{x}). \end{aligned} \quad (6)$$

In the general problem with  $K$  populations, the minimum misclassification rate is obtained when

$$\sum_{k=1}^K \pi_k \int_{\mathcal{X}} \Phi^{(k)}(\theta; \mathbf{u}^\top \mathbf{x}) dF^{(k)}(\mathbf{x}) < \sum_{k=1}^K \pi_k \int_{\mathcal{X}} \min_{k' \neq k} \Phi^{(k')}(\theta; \mathbf{u}^\top \mathbf{x}) dF^{(k)}(\mathbf{x}). \quad (7)$$

Let  $\Delta^{(k)}(\mathbf{x}, \theta, \mathbf{u}) = \Phi^{(k)}(\theta; \mathbf{u}^\top \mathbf{x}) - \min_{k' \neq k} \Phi^{(k')}(\theta; \mathbf{u}^\top \mathbf{x})$ . Given a sample of  $n$  observations  $\mathbf{x}_i$  and corresponding class labels  $\ell_i \in \{1, 2, \dots, K\}$ , we aim to solve

$$\min_{\vartheta, v, \omega} \sum_{k=1}^K \sum_{i: \ell_i=k} \sum_{r=1}^R \sum_{s=1}^{S_r} \omega_{rs} \Delta^{(k)}(\mathbf{x}_i, \theta_r, \mathbf{u}_{rs}). \quad (8)$$

Problem (8) may seem daunting, but luckily we can solve for  $\omega$  rather easily. Given  $\vartheta$  and  $v$ , problem (8) is linear with unit-norm constraints and can be minimised by using the Lagrange multiplier method. This problem has a closed-form solution given by  $\hat{\omega} = (\hat{\omega}_{11}, \dots, \hat{\omega}_{1S_1}, \dots, \hat{\omega}_{RS_R})^\top$  with generic  $r$ st element

$$\hat{\omega}_{rs} = \frac{\tilde{\Delta}_{rs}}{\sqrt{\sum_{r=1}^R \sum_{s=1}^{S_r} \tilde{\Delta}_{rs}^2}}, \quad (9)$$

where  $\tilde{\Delta}_{rs} = \sum_{k=1}^K \sum_{i: \ell_i=k} \Delta^{(k)}(\mathbf{x}_i, \theta_r, \mathbf{u}_{rs})$ .

We now turn to how to choose directions and quantile levels. A crude solution would consist in doing a multidimensional grid search on  $p+1$  dimensions. However, such a solution would become computationally prohibitive even at modest values of  $p$ . Thankfully, we are able to mitigate the computational cost of a naïve numerical solution with some theoretical results (Section 3); in particular, with Theorem 1, which guarantees that for each projection there exists (at least) a quantile that leads to the optimal Bayes misclassification probability, and Theorem 2, which, conversely, identifies the best direction for a given quantile level. Unfortunately, a theoretical result for the simultaneous optimisation with respect to  $\theta$  and  $\mathbf{u}$  does not exist. Nevertheless, we show that our DQC is asymptotically optimal (i.e. the misclassification rate goes to zero) when the number of directions increases with  $p$  and  $n$  (Theorem 3) under certain assumptions.

It shall be clear that, in principle, there are several alternative ways of implementing our proposed classification procedure. After investigating some alternatives (results not shown), we found that a strategy that gives satisfactory results in different data settings is the one that we pseudocoded in Algorithm 1. First, we define a grid of  $\theta$  values with length  $R$  spanning the



unit interval and, for each of these values, we randomly draw a set of  $S_r$  normalised directions from the orthant space defined by the optimal direction of Theorem 2 (that is, the  $p$ -dimensional Euclidean space of the vectors having the same component signs of the optimal direction).

We conclude this section by summarizing the key points concerning the implementation of our proposed classifier:

- (a) since the DQC requires choosing  $R$  distinct quantile levels and  $S_r$  distinct normalised directions,  $r = 1, \dots, R$ , the computational burden may easily become excessive, especially when  $n$  is large;
- (b) for low or moderate sample sizes, we propose using Algorithm 1, with a fixed grid of  $R \gg 1$  quantile levels (in our simulation study, we set  $R = 50$ ) and a uniform random sample of  $S_r > 20$  directions (in our simulation study, we set  $S_r = 100$ ). The theoretical basis for this strategy is supported by Theorem 3, at least when the distributions of competing populations differ by at most a location-shift, as it ensures consistency of the DQC when  $R$  and  $S_r$  grow large. On the other hand, Theorem 2 partially supports Algorithm 1 because we sample from the same orthant of the optimal direction.
- (c) to reduce the computational burden when the sample size is large, one may set  $R = 1$ ,  $S_r \gg 20$ , for  $r = 1, \dots, R$ , and find the optimal  $\theta$  according to Theorem 1 or set  $R \gg 1$ ,  $S_r = 1$ , for  $r = 1, \dots, R$ , and find the optimal  $\mathbf{u}$  according to Theorem 2;
- (d) since a theoretical result for the optimal choice of  $\theta$  and  $\mathbf{u}$  does not exist, we do not recommend setting  $R = 1$  and  $S_r = 1$ .

### 3 Theoretical results

In this section, we present theoretical results concerning our DQC. The proofs of lemmas and theorems are reported in the Appendix.

---

**Algorithm 1** Pseudocode of the algorithm for directional quantile classification of  $\mathbf{y}$ 

---

Fix  $\theta_1, \dots, \theta_R$  spanning the unit interval.

**for**  $r = 1, \dots, R$  **do**

    Compute  $\hat{u}_r$ , the optimal direction according to (12) in Theorem 2.

**for**  $s = 1, \dots, S_r$  **do**

**for**  $j = 1, \dots, p$  **do**

**if**  $\hat{u}_{rj} \geq 0$  **then**

                Sample  $u_{rsj} \sim U[0, 1]$

**else**

                Sample  $u_{rsj} \sim U[-1, 0]$

**end if**

**end for**

**end for**

    Normalize  $u_{rsj} \leftarrow u_{rsj} / \sqrt{\sum_{h=1}^p u_{rsh}^2}$

**end for**

Compute optimal weights  $\omega_{rs}$  for  $s = 1, \dots, S_r, r = 1, \dots, R$ , as in (9).

Compute discrepancy  $d(\mathbf{y}, \vartheta, v, \boldsymbol{\omega})$  as in (4)

**if**  $d(\mathbf{y}, \vartheta, v, \boldsymbol{\omega}) > 0$  **then**

    Assign  $\mathbf{y}$  to  $\Pi^{(1)}$

**else**

    Assign  $\mathbf{y}$  to  $\Pi^{(2)}$

**end if**

---

### 3.1 Optimal quantile level $\theta$

Although we suggest to consider a grid of quantiles as in Algorithm 1, there could be situations in which using a single quantile is preferable (e.g., because one wishes to identify such a quantile or for computational reasons). In this section we show that under general assumptions there exists an optimal quantile. We derive the theoretical rate of correct classification as a function of  $\theta$ , for given  $\mathbf{u}$ . We assume  $K = 2$  populations, although results can be generalised to  $K > 2$ .

**Lemma 1** For given  $\mathbf{u}$ , let  $Q_\alpha(\theta; \mathbf{u}) = \min\{Q_X^{(1)}(\theta; \mathbf{u}), Q_X^{(2)}(\theta; \mathbf{u})\}$  with corresponding inverse  $G_\alpha(\cdot; \mathbf{u})$ , density  $g_\alpha(\cdot; \mathbf{u})$ , and prior probability of correct classification  $\pi_\alpha$ , and let  $Q_\beta(\theta; \mathbf{u}) = \max\{Q_X^{(1)}(\theta; \mathbf{u}), Q_X^{(2)}(\theta; \mathbf{u})\}$  with corresponding inverse  $G_\beta(\cdot; \mathbf{u})$ , density  $g_\beta(\cdot; \mathbf{u})$ , and prior probability of correct classification  $\pi_\beta$ . The probability of correct classification of the directional quantile classifier is

$$\psi(\theta) = \pi_\alpha G_\alpha(\tilde{Q}(\theta; \mathbf{u}); \mathbf{u}) + \pi_\beta \{1 - G_\beta(\tilde{Q}(\theta; \mathbf{u}); \mathbf{u})\} \quad (10)$$

where  $\tilde{Q}(\theta; \mathbf{u}) = \theta Q_\alpha(\theta; \mathbf{u}) + (1 - \theta) Q_\beta(\theta; \mathbf{u})$ . Analogously, the theoretical misclassification rate is

$$1 - \psi(\theta) = \pi_\alpha \{1 - G_\alpha(\tilde{Q}(\theta; \mathbf{u}); \mathbf{u})\} + \pi_\beta G_\beta(\tilde{Q}(\theta; \mathbf{u}); \mathbf{u}). \quad (11)$$

**Theorem 1** Assume that the density functions  $g_\alpha(z; \mathbf{u})$  and  $g_\beta(z; \mathbf{u})$  exist for  $z$  and are nonzero on the same compact domain  $\mathcal{Z}$ . Further assume that there is a point  $z_0$  with  $\pi_\alpha g_\alpha(z_0; \mathbf{u}) = \pi_\beta g_\beta(z_0; \mathbf{u})$  so that  $\pi_\alpha g_\alpha(z; \mathbf{u}) > \pi_\beta g_\beta(z; \mathbf{u})$  for  $z$  on one side of  $z_0$  and  $\pi_\alpha g_\alpha(z; \mathbf{u}) < \pi_\beta g_\beta(z; \mathbf{u})$  for  $z$  on the other side of  $z_0$ . Then the quantile classifier using the quantile  $\tilde{Q}(\theta; \mathbf{u})$  that minimises the theoretical misclassification probability achieves the optimal Bayes misclassification probability in the projected space defined by  $\mathbf{u}$ .

The consistency of the classifier may be illustrated with an example. Consider a two class decision problem where one population is a location-shift version of the other. Figure 2 shows two distributions which have both the same right skewness. The quantiles  $Q_\alpha(\theta)$  and  $Q_\beta(\theta)$  are marked by dashed lines. The median classifier (Hall et al., 2009) in the upper panel leads to a non-optimal misclassification probability equal to 0.30. However, the misclassification probability is reduced to 0.28 by setting  $\theta = 0.202$ .

### 3.2 Optimal direction $\mathbf{u}$

In the following theorem we show how to derive the optimal direction that minimises the misclassification rate at a given  $\theta$ .

**Theorem 2** Let  $\mathbf{W} = (W_1, W_2, \dots, W_p)^\top$  be a  $p$ -variate random variable such that  $Q_{W_j}(\theta) = 0$ , for  $j = 1, \dots, p$ , and let  $\boldsymbol{\mu}^{(k)} = (\mu_1^{(k)}, \mu_2^{(k)}, \dots, \mu_p^{(k)})^\top$  be a vector of constants,  $k = 1, 2$ . We

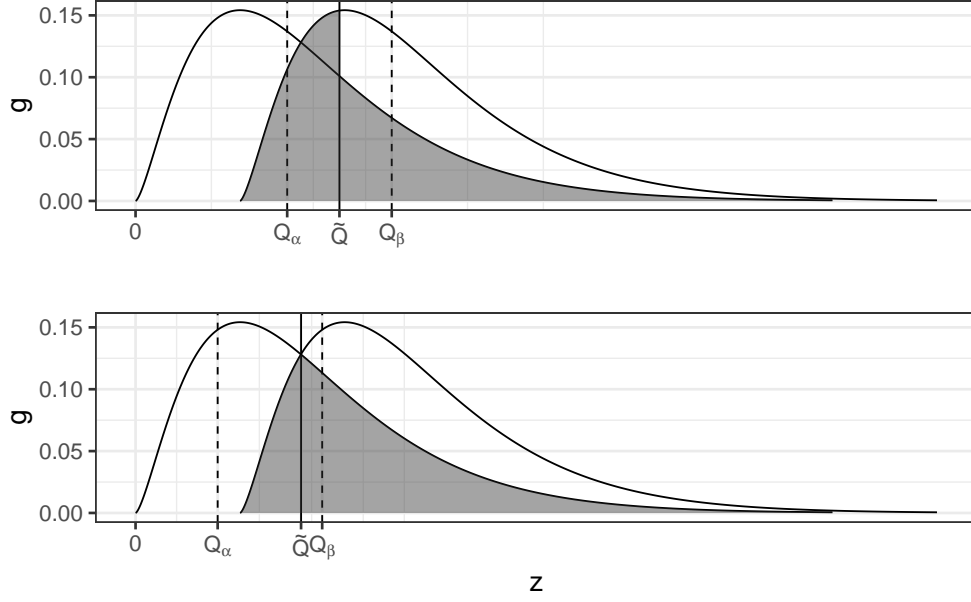


Figure 2: Misclassification probability (shaded grey area) with two location-shifted skewed distributions according to median classifier (upper panel) and the optimal quantile classifier (lower panel).

assume that  $\mathbf{X}^{(k)} = \mathbf{W} + \boldsymbol{\mu}^{(k)}$  and its probability distribution function is  $F^{(k)}$ , for  $k = 1, 2$ . Moreover, assume that  $Q_X^{(2)}(\theta; \mathbf{u}) > Q_X^{(1)}(\theta; \mathbf{u})$ , where  $Q_X^{(k)}(\theta; \mathbf{u})$  is the  $\theta$ -quantile of  $Z^{(k)} \equiv \mathbf{u}^\top \mathbf{X}^{(k)}$ . (Notice that there is no loss of generality with this assumption since the case  $Q_X^{(2)}(\theta; \mathbf{u}) \leq Q_X^{(1)}(\theta; \mathbf{u})$  can be reformulated as  $Q_X^{(2)}(\theta; -\mathbf{u}) > Q_X^{(1)}(\theta; -\mathbf{u})$ .) Under these assumptions, the normalised direction  $\mathbf{u}$  that minimises the misclassification error (6) is

$$\hat{\mathbf{u}} = \frac{Q_X^{(2)}(\theta) - Q_X^{(1)}(\theta)}{\|Q_X^{(2)}(\theta) - Q_X^{(1)}(\theta)\|}, \quad (12)$$

where  $Q_X^{(k)}(\theta) \equiv \boldsymbol{\mu}^{(k)}$ .

The generalization of Theorem 2 to  $K > 2$  populations involves  $K(K - 1)/2$  optimal directions for each of all the possible pairwise comparisons.

### 3.3 Asymptotic misclassification rate

In this section, we show that under certain assumptions, the correct classification probability converges to unity when the number of dimensions grows to infinity along with the sample size and the number of projections. The proof is built following a strategy similar to that used in Hall et al. (2009, Theorem 2), although our premises start from milder assumptions. In particular the projections are not required to obey the “ $\psi$ -mixing condition” (Bradley, 2005), which is rather strict in practice. Our theorem is developed for any  $\theta_r \in (0, 1)$ , unit weights  $\omega_{rs} = 1$ , and  $R = 1$ . Thus, the asymptotic result holds for sub-components of the summation in (8), which are then weighted and summed to minimise the misclassification rate. Hence, the overall criterion inherits the optimal properties of its additive components.

As we did with the theorems in the previous sections, we present this theorem for  $K = 2$  classes. Its extension to  $K > 2$  classes requires contrasting each class against the remaining  $K - 1$  classes, consistently with (7).

**Theorem 3** *Consider a given quantile  $\theta$  and a set of directions  $v = \{\mathbf{u}_1, \dots, \mathbf{u}_S\}$  sampled from a unit  $p$ -sphere and let  $n = \max(n_1, n_2)$ , with  $n_1$  and  $n_2$  denoting the sample sizes of the two groups in the training set. Assume*

- (i) *For a constant  $A_1 > 0$ ,  $S \geq A_1 n$ .*
- (ii) *The  $p$  variables  $X_1^{(k)}, X_2^{(k)}, \dots, X_p^{(k)}$  have each the same distribution as  $W_1 + \mu_1^{(k)}, W_2 + \mu_2^{(k)}, \dots, W_p + \mu_p^{(k)}$ , respectively. Moreover,  $Q_{W_j}(\theta) = 0 \forall j$  and  $\sup_{j \geq 1} \text{Var}(W_j) = A_2 < +\infty$ .*
- (iii) *The first moments of the projections are uniformly bounded in a strong sense. This implies that  $\forall c > 0$  and  $\forall \mathbf{u}, \exists \mathbf{v}$  with  $|\mathbf{u}^\top \mathbf{v}| > c$  such that*

$$\inf_{s \geq 1} \inf_{|\mathbf{u}_s^\top \mathbf{v}| > c} \theta \mathbb{E} |\mathbf{u}_s^\top \mathbf{W} + \mathbf{u}_s^\top \mathbf{v}| - (1 - \theta) \mathbb{E} |\mathbf{u}_s^\top \mathbf{W}| > 0.$$

- (iv) *For some  $\epsilon > 0$ , the proportion of values  $s \in \{1, 2, \dots, S\}$  for which*

$$|\theta \mathbf{u}_s^\top \boldsymbol{\mu}^{(2)} - (1 - \theta) \mathbf{u}_s^\top \boldsymbol{\mu}^{(1)}| > \epsilon$$

*multiplied by  $n^{1/2}$ , say  $n^{1/2} \sharp \mathcal{K}_\epsilon$ , is of larger order than  $S$ , which means  $S (n^{1/2} \sharp \mathcal{K}_\epsilon)^{-1}$  goes to zero as  $n$  and  $S$  increase.*

*Under the previous assumptions, the directional quantile classifier  $\mathcal{C}$  based on*

$$d(\mathbf{y}, \theta, v, \omega) = \sum_{s=1}^S \{\Phi^{(2)}(\theta; \mathbf{u}_s^\top \mathbf{y}) - \Phi^{(1)}(\theta; \mathbf{u}_s^\top \mathbf{y})\},$$

*makes the correct choice asymptotically. More specifically, as  $p \rightarrow \infty$ , the classifier  $\mathcal{C}$  makes the correct decision with probability*

$$P^{(1)}\{\mathcal{C}(\mathbf{Y}) = 1\} + P^{(2)}\{\mathcal{C}(\mathbf{Y}) = 2\}$$

*converging to 1 if both  $n_1$  and  $n_2$  diverge with  $p$ , where  $P^{(k)}$ ,  $k = 1, 2$  denotes the probability computed under the assumption that  $Y$  is drawn from population  $k$ .*

The proof of Theorem 3 is shown in Appendix. Here we only comment on the assumptions, which are similar to those given in Hall et al. (2009, Theorem 2). In particular, condition (i) requires the number of directions  $S$  and the training sample size be of the same order. Condition (ii) implies that classes differ up to a location-shift  $\mu_j^{(k)}$  from a  $\theta$ -quantile centered distribution, or, in other words, that the discriminative information is contained in the marginal quantiles of the  $p$  variables. We also assume finite variances of the  $W_j$ 's, thus avoiding the more restrictive mixing condition assumed in Hall et al. (2009, Theorem 2). Condition (iii) concerns uniform continuity and boundedness along every direction  $s$ . Assumption (iv) is related to the proportion of nonzero signals that can decrease to 0 as the number of directions  $S$  and the sample size  $n$  increase, without affecting the consistency of the classifier.

## 4 Simulation study

We assessed the performance of the proposed classifier in a simulation study under five scenarios with two populations. In the first scenario, observations were generated independently from a multivariate distribution with normal marginals. In the second scenario, observations were generated independently from a multivariate distribution with Student's  $t_3$  marginals. In the third scenario, observations were generated as in the second scenario, but each variable was subsequently transformed according to  $x \mapsto \log(|x|)$  to induce asymmetry. In the fourth scenario observations were generated as in the second scenario, but each variable was subsequently transformed according to  $x \mapsto \log(|x|)$  or to  $x \mapsto -\log(|x|)$  depending on whether observations

belonged to one or the other population, respectively. Finally, in the fifth scenario observations were generated independently from a multivariate distribution with exponential marginals. For each scenario, we considered both uncorrelated and correlated variables. This gave ten data generating processes.

Data were generated for each combination of overall sample size  $n \in \{50, 100, 500\}$  (with  $n/2$  observations in each class) and dimension  $p \in \{10, 50, 100, 500\}$ . All in all, this resulted in  $10 \times 3 \times 4 = 120$  simulation cases. The two populations differed by a location shift equal to 0.4, except for the fourth scenario where the location shift was naturally determined by the opposite skewness. The variance-covariance matrix used to generate correlated variables was defined by using  $\Sigma = \mathbf{A}^\top \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p) \mathbf{A}$  with values of  $\mathbf{A}$  and  $\sigma$ 's determined so that pairwise correlations were on the interval  $(-0.63, 0.80)$ . Observations in the training and test datasets were generated in the same way. Data generation under each setting was replicated 100 times.

We compared the directional quantile classifier (DQC) in terms of misclassification rate on the test data with that of the centroid classifier (Centroid) (Tibshirani et al., 2002), median classifier (Median) (Hall et al., 2009), componentwise quantile classifier (CQC) (Hennig and Viroli, 2016a), ensemble quantile classifier (EQC) (Lai and McLeod, 2020), Fisher's linear discriminant analysis (LDA),  $k$ -nearest neighbour (KNN) (Cover and Hart, 1967), penalised logistic regression (PLR) (Park and Hastie, 2008), support vector machines (SVM) (Cortes and Vapnik, 1995; Wang et al., 2008), and naïve Bayes classifier (Bayes) (Hand and Yu, 2001). Tuning parameters for PLR, KNN, and SVM were selected using cross-validation. For the CQC, the Galton correction was used to reduce skewness and optimal quantile was selected by minimising the error rate on the training set (Hennig and Viroli, 2016a).

We used the package `Qtools` (Geraci, 2016, 2020) for the directional quantile classifier; the package `quantileDA` (Hennig and Viroli, 2016b) for the centroid, median and componentwise quantile classifiers; the package `eqc` (Lai and McLeod, 2019) for the ensemble quantile classifier; the package `MASS` (Venables and Ripley, 2002) for linear discriminant analysis; the package `class` (Venables and Ripley, 2002) for  $k$ -nearest neighbour; the package `e1071` (Meyer et al., 2019) for support vector machines and Bayes classifiers; and the package `stepAIC` (Park and

Hastie, 2018) for penalised logistic regression. All analyses were carried out in R version 4.0.0 (R Core Team, 2020).

For estimation, we applied Algorithm 1 using a sequence of  $R = 50$  quantile levels  $\theta$  between 0.01 and 0.99, and  $S_r = 100$  directions per quantile. The misclassification rates averaged over 100 replications for all simulation cases are reported in Table 2 for the Gaussian marginal case, and the remaining scenarios can be found in the Web Supplement. The results indicate that the performance of our proposed classifier improves as  $n$  and  $p$  increase, in agreement with the theoretical results. Our classifier generally outperforms our competitors, with some exceptions with low dimension and/or low sample size, in which it is anyway still among the best approaches. Moreover, as  $n$  and  $p$  increase the standard errors of the misclassification rates tend to zero, thus indicating the stability of the classification across the several experiments.

## 5 Clinical trial on Crohn’s disease

We analysed data from a matched case-control study in first-degree relatives (FDRs) of Crohn’s disease (CD) patients originally published by Sorrentino et al. (2014). The goal of the study was to identify asymptomatic FDRs with early CD signs using several intestinal inflammatory markers. The latter included hemoglobin, erythrocyte sedimentation rate, C-reactive protein, fecal calprotectin, and average mature ileum score. In our analysis, we grouped subjects into 2 classes, one with signs of inflammation ( $n_1 = 9$  subjects with early or frank CD) and one with normal values of markers ( $n_2 = 26$  subjects with no signs of inflammation, including healthy controls). In a separate analysis, we augmented the dataset with 45 artificial markers generated from independent standard normal distributions to investigate the impact of uninformative noise on the performance of the DQC. For estimation, we applied Algorithm 1 using a sequence of  $R = 25$  quantile levels  $\theta$  between 0.01 and 0.99, and  $S_r = 10000$  directions per quantile. We compared the estimated misclassification rate of our DQC to the rates estimated for all the classifiers included in our simulation study (Section 4). For the purpose of this study, the misclassification rate was estimated by the proportion of subjects that are misclassified when each of them is left out of analysis (leave-one-out validation). The classification error estimates are reported in Table 3. The proposed DQC outperforms its competitors in both the original ( $p = 5$ ) and noisy ( $p = 50$ )



versions of the dataset.

## 6 Conclusions

We proposed directional quantile classifiers, whose predictive ability is consistently good in both simulation and real data studies, on small and large dimensional classification problems. In particular, the empirical results show that our approach either outperforms its competitors or, when this is not the case, its performance is still in the ballpark of that of the best classifiers. Such a reliable behaviour across different scenarios is not shared by the other selected classifiers. Moreover, the directional quantile classifiers enjoy optimal theoretical properties under certain assumptions. Our theoretical results indicate that one can sample directions from the optimal orthant defined by Theorem 2, thus reducing the computational burden, but not at a significant expense of the classifier’s performance. Should the computational burden be prohibitive, one can exploit Theorem 1 to use a single quantile ( $R = 1$ ). Our strategy allows us to balance the importance of quantile levels and directions used for classification by means of weights, which can be optimised using a convenient closed-form expression. The use of quantiles makes our approach particularly advantageous with heavy-tailed and asymmetric distributions, and when the populations differ by shape, given the natural ability of quantiles to flexibly model distributions. Also, our approach deals with correlation by means of projections, thus improving on the performance of quantile classifiers when correlation is ignored.

We often stressed that there are alternative ways for the practical implementation of the proposed DQC. In general, we recommend using multiple quantiles and directions in the optimal orthant. However, this particular strategy does not lend itself to a simple interpretation of the *mechanics* of the DQC, which can be seen as acting on a coarsened and projected versions of the entire multivariate distributions. In contrast, if a single quantile and direction are used, then their practical role in the classification process and subsequent interpretation become more evident. We cannot exclude that a single-quantile strategy might be theoretically a better choice (especially for moderate to large sample sizes in view of Assumption 1 which requires the number of directions to be proportional to the sample size), although during our preliminary investigations using simulated and real data we did not observe meaningful differences. If anything, using mul-

multiple directions has its own limitation since the number of directions needed to span a  $p$ -sphere with a regular grid becomes prohibitive already at modest values of  $p$ .

## A Proofs of the theoretical results

### A.1 Proofs of Lemma 1 and Theorem 1

The proofs of Lemma 1 and Theorem 1 follow the arguments given in Hennig and Viroli (2016a, Supplementary Material). Here, we briefly sketch the main idea. The condition that a point  $z_0$  exists with  $\pi_\alpha g_\alpha(z_0; \mathbf{u}) = \pi_\beta g_\beta(z_0; \mathbf{u})$  so that  $\pi_\alpha g_\alpha(z; \mathbf{u}) > \pi_\beta g_\beta(z; \mathbf{u})$  for  $z$  on one side of  $z_0$  and  $\pi_\alpha g_\alpha(z; \mathbf{u}) < \pi_\beta g_\beta(z; \mathbf{u})$  for  $z$  on the other side, ensures that the densities cross in a point where the Gini transvariation area is minimized. The optimal value  $\theta$  that minimises the theoretical misclassification probability can be obtained by setting the first derivative of (11) to zero, from which

$$\pi_\alpha g_\alpha\{\tilde{Q}(\theta; \mathbf{u})\} = \pi_\beta g_\beta\{\tilde{Q}(\theta; \mathbf{u})\}.$$

By assumption, there exists  $\theta \in (0, 1)$  such that  $\tilde{Q}(\theta; \mathbf{u}) = z_0$ . Hence, the identity above is satisfied because  $Q_\alpha(\theta; \mathbf{u})$  and  $Q_\beta(\theta; \mathbf{u})$  are continuous functions of  $\theta$  that converge to the lower and upper bound of  $\mathcal{Z}$  for  $\theta$  approaching either 0 or 1, respectively. Furthermore, under the assumptions of Theorem 1, the optimal Bayesian classifier has a single decision boundary at  $\tilde{Q}(\theta; \mathbf{u})$ .

### A.2 Proof of Theorem 2

We start by the following general result that will be used in the proof.

**Lemma 2** *Let  $z$  be a realisation of either  $Z^{(1)}$  or  $Z^{(2)}$ , then*

$$\Phi^{(2)}(\theta; z) - \Phi^{(1)}(\theta; z) \leq Q_Z^{(2)}(\theta) - Q_Z^{(1)}(\theta),$$

where  $\Phi^{(k)}(\theta; z) = \theta \max(\eta^{(k)}, 0) + (1 - \theta) \max(-\eta^{(k)}, 0)$  and  $\eta^{(k)} = z - Q_Z^{(k)}(\theta)$ ,  $k = 1, 2$ .

In order to prove this, assume  $Q_Z^{(1)}(\theta) \leq Q_Z^{(2)}(\theta)$  without loss of generality. Let  $\Delta(\theta; z) = \Phi^{(2)}(\theta; z) - \Phi^{(1)}(\theta; z)$  and consider three possible, distinct cases:  $z \leq Q_Z^{(1)}(\theta)$ ,  $Q_Z^{(1)}(\theta) < z <$

$Q_Z^{(2)}(\theta)$ , and  $Q_Z^{(2)}(\theta) \leq z$ . If  $z \leq Q_Z^{(1)}(\theta)$ , then

$$\begin{aligned}\Delta(\theta; z) &= (1 - \theta)\{Q_Z^{(2)}(\theta) - z\} - (1 - \theta)\{Q_Z^{(1)}(\theta) - z\} \\ &= (1 - \theta)\{Q_Z^{(2)}(\theta) - Q_Z^{(1)}(\theta)\} \leq Q_Z^{(2)}(\theta) - Q_Z^{(1)}(\theta)\end{aligned}$$

by definition. If  $Q_Z^{(1)}(\theta) < z < Q_Z^{(2)}(\theta)$ , then

$$\begin{aligned}\Delta(\theta; z) &= (1 - \theta)\{Q_Z^{(2)}(\theta) - z\} - \theta\{z - Q_Z^{(1)}(\theta)\} \\ &= \theta\{Q_Z^{(1)}(\theta) - Q_Z^{(2)}(\theta)\} + Q_Z^{(2)}(\theta) - z \\ &\leq \theta Q_Z^{(1)}(\theta) - Q_Z^{(2)}(\theta) \leq Q_Z^{(2)}(\theta) - Q_Z^{(1)}(\theta).\end{aligned}$$

Finally, if  $Q_Z^{(2)}(\theta) \leq z$ , then

$$\begin{aligned}\Delta(\theta; z) &= \theta\{z - Q_Z^{(2)}(\theta)\} - \theta\{z - Q_Z^{(1)}(\theta)\} \\ &\leq Q_Z^{(2)}(\theta) - Q_Z^{(1)}(\theta).\end{aligned}$$

This completes the proof of Lemma 2. By Lemma 2, the differences  $\Phi^{(1)}(\theta; \mathbf{u}^\top \mathbf{x}) - \Phi^{(2)}(\theta; \mathbf{u}^\top \mathbf{x})$  and  $\Phi^{(2)}(\theta; \mathbf{u}^\top \mathbf{x}) - \Phi^{(1)}(\theta; \mathbf{u}^\top \mathbf{x})$  are upper bounded by  $Q_Z^{(2)}(\theta; \mathbf{u}) - Q_Z^{(1)}(\theta; \mathbf{u})$  since  $Q_Z^{(2)}(\theta; \mathbf{u}) > Q_Z^{(1)}(\theta; \mathbf{u})$ . Therefore the quantity in (6), which is to be minimised with respect to  $\mathbf{u}$  subject to  $\|\mathbf{u}\| = 1$ , is uniformly bounded above by

$$\begin{aligned}&\pi_1 \int_{\mathcal{X}} \{\Phi^{(1)}(\theta; \mathbf{u}^\top \mathbf{x}) - \Phi^{(2)}(\theta; \mathbf{u}^\top \mathbf{x})\} dF^{(1)}(\mathbf{x}) \\ &+ \pi_2 \int_{\mathcal{X}} \{\Phi^{(2)}(\theta; \mathbf{u}^\top \mathbf{x}) - \Phi^{(1)}(\theta; \mathbf{u}^\top \mathbf{x})\} dF^{(2)}(\mathbf{x}) + \lambda(\mathbf{u}^\top \mathbf{u} - 1) \\ &\leq Q_Z^{(2)}(\theta; \mathbf{u}) - Q_Z^{(1)}(\theta; \mathbf{u}) + \lambda(\mathbf{u}^\top \mathbf{u} - 1) \\ &= (Q_W(\theta; \mathbf{u}) + \mathbf{u}^\top \boldsymbol{\mu}^{(2)}) - (Q_W(\theta; \mathbf{u}) + \mathbf{u}^\top \boldsymbol{\mu}^{(1)}) + \lambda(\mathbf{u}^\top \mathbf{u} - 1) \\ &= \mathbf{u}^\top (\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}) + \lambda(\mathbf{u}^\top \mathbf{u} - 1).\end{aligned}$$

To find  $\mathbf{u}$ , we minimise the Lagrangian function  $\mathbf{u}^\top (\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}) + \lambda(\mathbf{u}^\top \mathbf{u} - 1)$  which has solution

$$\frac{\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}}{\|\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\|}.$$

Finally, equation (12) is obtained by observing that  $\boldsymbol{\mu}^{(k)}$  is the quantile of  $\mathbf{X}^{(k)}$  at  $\theta$ , since  $Q_{\mathbf{X}}(\theta) = 0$  by assumption.

### A.3 Proof of Theorem 3

Let  $Q_X^{(k)}(\theta; \mathbf{u}_s)$  be the empirical quantile computed on the projected training data  $\mathbf{u}_s^\top \mathbf{X}^{(k)}$ . We write

$$\Phi^{(k)}(\theta; \mathbf{u}_s^\top \mathbf{Y}) = \gamma_s^{(k)}(\theta) |\mathbf{u}_s^\top \mathbf{Y} - Q_X^{(k)}(\theta; \mathbf{u}_s)|,$$

where  $\gamma_s^{(k)}(\theta) = \theta + (1 - 2\theta)I\{\mathbf{u}_s^\top \mathbf{Y} < Q_X^{(k)}(\theta; \mathbf{u}_s)\}$ . Let  $\boldsymbol{\mu}_y$  denote the vector of quantiles of  $\mathbf{Y}$ , and put  $\boldsymbol{\mu}_y^{(k)} = \boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}_y$  for  $k = 1, 2$  and write  $\mathbf{V} = \mathbf{Y} - \boldsymbol{\mu}_y$ . By the triangular inequality

$$\begin{aligned} & \gamma_s^{(2)}(\theta) |\mathbf{u}_s^\top \mathbf{Y} - Q_X^{(2)}(\theta; \mathbf{u}_s)| - \gamma_s^{(1)}(\theta) |\mathbf{u}_s^\top \mathbf{Y} - Q_X^{(1)}(\theta; \mathbf{u}_s)| \\ &= \gamma_s^{(2)}(\theta) |\mathbf{u}_s^\top \mathbf{V} - \mathbf{u}_s^\top \boldsymbol{\mu}_y^{(2)}| - \gamma_s^{(1)}(\theta) |\mathbf{u}_s^\top \mathbf{V} - \mathbf{u}_s^\top \boldsymbol{\mu}_y^{(1)}| \\ &+ \tau_2 |Q_X^{(2)}(\theta; \mathbf{u}_s) - \mathbf{u}_s^\top \boldsymbol{\mu}^{(2)}| + \tau_1 |Q_X^{(1)}(\theta; \mathbf{u}_s) - \mathbf{u}_s^\top \boldsymbol{\mu}^{(1)}|, \end{aligned}$$

where  $\tau_1$  and  $\tau_2$  satisfy  $|\tau_k| \leq 1$ ,  $k = 1, 2$ . Hence

$$\begin{aligned} T_1 &\equiv \sum_{s=1}^S \gamma_s^{(2)}(\theta) |\mathbf{u}_s^\top \mathbf{Y} - Q_X^{(2)}(\theta; \mathbf{u}_s)| - \gamma_s^{(1)}(\theta) |\mathbf{u}_s^\top \mathbf{Y} - Q_X^{(1)}(\theta; \mathbf{u}_s)| \\ &= T_2 + \tau_1 R_1 + \tau_2 R_2, \end{aligned}$$

where  $T_2 = \sum_{s=1}^S \gamma_s^{(2)}(\theta) |\mathbf{u}_s^\top \mathbf{V} - \mathbf{u}_s^\top \boldsymbol{\mu}_y^{(2)}| - \gamma_s^{(1)}(\theta) |\mathbf{u}_s^\top \mathbf{V} - \mathbf{u}_s^\top \boldsymbol{\mu}_y^{(1)}|$ ,  $R_1 = \sum_{s=1}^S |Q_X^{(1)}(\theta; \mathbf{u}_s) - \mathbf{u}_s^\top \boldsymbol{\mu}^{(1)}|$  and  $R_2 = \sum_{s=1}^S |Q_X^{(2)}(\theta; \mathbf{u}_s) - \mathbf{u}_s^\top \boldsymbol{\mu}^{(2)}|$ . Given the convergence of the empirical quantiles to the population quantiles,

$$\begin{aligned} P^{(1)}(T_1 > c_1 - 2c_2 S n^{-1/2}) &\geq P^{(1)}(T_2 > c_1) - P(R_1 > c_2 S n^{-1/2}) - P(R_2 > c_2 S n^{-1/2}) \\ &\geq P^{(1)}(T_2 > c_1) - 2 \sum_{s=1}^S e^{-2n_1 \delta_s^{(1)}} - 2 \sum_{s=1}^S e^{-2n_2 \delta_s^{(2)}} \end{aligned}$$

for any  $c_1, c_2 > 0$ , where

$$\delta_s^{(k)} = \left[ \min \left\{ F^{(k)} \left( \mathbf{u}_s^\top \boldsymbol{\mu}^{(k)} + \frac{c_2 S}{n^{1/2}} \right) - \theta, \theta - F^{(k)} \left( \mathbf{u}_s^\top \boldsymbol{\mu}^{(k)} - \frac{c_2 S}{n^{1/2}} \right) \right\} \right]^2.$$

Now define

$$d_s = \mathbb{E} \left\{ \gamma_s^{(2)}(\theta) |\mathbf{u}_s^\top (\mathbf{V} - \boldsymbol{\mu}_y^{(2)})| - \gamma_s^{(1)}(\theta) |\mathbf{u}_s^\top (\mathbf{V} - \boldsymbol{\mu}_y^{(1)})| \right\}.$$

Given  $\epsilon > 0$ , let  $\mathcal{K}_\epsilon$  denote the set of indices  $s \in \{1, 2, \dots, S\}$  such that

$$|\gamma_s^{(2)}(\theta) \mathbf{u}_s^\top \boldsymbol{\mu}_2 - \gamma_s^{(1)}(\theta) \mathbf{u}_s^\top \boldsymbol{\mu}_1| > \epsilon$$

$\forall \theta \in (0, 1)$ . Under the assumption that  $\mathbf{Y}$  has distribution  $F^{(1)}$ , we have

$$\begin{aligned} d_s &= \mathbb{E} \left\{ \gamma_s^{(2)}(\theta) |\mathbf{u}_s^\top (\mathbf{Y} - \boldsymbol{\mu}_2)| - \gamma_s^{(1)}(\theta) |\mathbf{u}_s^\top (\mathbf{Y} - \boldsymbol{\mu}_1)| \right\} \\ &= \gamma_s^{(2)}(\theta) \mathbb{E}_1 |\mathbf{u}_s^\top (\mathbf{Z} + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)| - \gamma_s^{(1)}(\theta) \mathbb{E}_1 |\mathbf{u}_s^\top \mathbf{Z}|, \end{aligned}$$

where  $\mathbb{E}_1$  is the expectation under  $P^{(1)}$ . Therefore, by assumption (iii) and provided  $c \geq \epsilon$ , we have

$$\sum_{s \in \mathcal{K}_\epsilon} d_s \geq a(c)(\#\mathcal{K}_c)$$

where  $a(c) > 0$ , with  $a(c) = \gamma_s^{(2)}(\theta) \mathbb{E}_1 |\mathbf{u}_s^\top (\mathbf{Z} + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)| - \gamma_s^{(1)}(\theta) \mathbb{E}_1 |\mathbf{u}_s^\top \mathbf{Z}|$  in view of (iii). As a consequence, for  $\mathbb{E}_1(T_2) = \sum_{s=1}^S d_s$  and  $\epsilon \rightarrow 0$ , and  $\forall c$ , we have

$$\mathbb{E}_1(T_2) \geq a(c)(\#\mathcal{K}_c), \quad (13)$$

where  $\#A$  denotes the cardinality of the set  $A$ . By the Chebychev inequality and provided that  $c_1 < \frac{1}{2} \mathbb{E}_1(T_2)$ , we have

$$\begin{aligned} P^{(1)}(T_2 > c_1) &\geq 1 - P^{(1)}(|T_2 - \mathbb{E}_1(T_2)| > c_1) \geq 1 - c_1^{-2} \mathbb{E}_1\{T_2 - \mathbb{E}_1(T_2)\}^2 \\ &\geq 1 - c_1^{-2} \text{var}_1(T_2) \geq 1 - A_2 c_1^{-2} S, \end{aligned} \quad (14)$$

where  $\text{var}_1$  denotes the variance under  $P^{(1)}$  and the second inequality follows from assumption (ii); more specifically

$$\begin{aligned} \text{var}_1(T_2) &= \text{var}_1 \left\{ \sum_{s=1}^S \left( \gamma_s^{(2)}(\theta) |\mathbf{u}_s^\top (\mathbf{V} - \boldsymbol{\mu}_y^{(2)})| - \gamma_s^{(1)}(\theta) |\mathbf{u}_s^\top (\mathbf{V} - \boldsymbol{\mu}_y^{(1)})| \right) \right\} \\ &\leq \text{var}_1 \left\{ \sum_{s=1}^S \left( \gamma_s^{(2)}(\theta) \mathbf{u}_s^\top (\mathbf{V} - \boldsymbol{\mu}_y^{(2)}) - \gamma_s^{(1)}(\theta) \mathbf{u}_s^\top (\mathbf{V} - \boldsymbol{\mu}_y^{(1)}) \right) \right\} \\ &= \text{var}_1 \left\{ \sum_{s=1}^S \left( \gamma_s^{(2)}(\theta) \mathbf{u}_s^\top (\mathbf{W} + \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) - \gamma_s^{(1)}(\theta) \mathbf{u}_s^\top \mathbf{W} \right) \right\} \\ &\leq \sum_{s=1}^S A_2 \mathbf{u}_s^\top \mathbf{u}_s + 2 \sum_{s=1}^{S-1} \sum_{s'=s+1}^S A_2 \mathbf{u}_s^\top \mathbf{u}_{s'}. \end{aligned}$$

Stam (1982) proved that a uniform random variable on the sphere,  $\mathbf{U} \in R^p$ , converges to a standard Gaussian as  $p \rightarrow \infty$ . Therefore, for  $S \rightarrow \infty$ , by the strong law of large numbers we have

$$\frac{2 \sum_{s=1}^{S-1} \sum_{s'=s+1}^S A_2 \mathbf{U}_s^\top \mathbf{U}_{s'}}{S(S-1)} \xrightarrow{a.s.} A_2 \mathbb{E}(\boldsymbol{\Xi}_1^\top \boldsymbol{\Xi}_2) = 0,$$

where  $\Xi_1$  and  $\Xi_2$  are two independent standard Gaussians. This explains why the covariances become negligible in the last part of (14) as  $p$  increases.

It remains to prove that  $c_1 < \frac{1}{2} E_1(T_2)$ . Consider  $c_1 = \frac{c_3 S}{n^{1/2}}$ , where  $c_3$  is a positive constant. By (13), the latter holds if  $c_3 S n^{-1/2} < \frac{1}{2} a(c) \mathcal{K}_c$ . But this is true because it implies that

$$S (n^{1/2} \# \mathcal{K}_c)^{-1} < \frac{1}{2} a(c) c_3^{-1},$$

where the term on the left goes to zero according to assumption (iv) while  $a(c) > 0$ , thus  $c_3^{-1} > 0$ .

For  $c_1 = \frac{c_3 S}{n^{1/2}}$ , we have

$$P^{(1)}(T_1 > c_3 S n^{-1/2} - 2c_2 S n^{-1/2}) \geq 1 - A_2 \frac{n}{c_3^2 S} - 2 \sum_{s=1}^S e^{-2n_1 \delta_s^{(1)}} - 2 \sum_{s=1}^S e^{-2n_2 \delta_s^{(2)}}.$$

We wish to choose  $c_3$  and  $c_2$  such as

$$P^{(1)}(T_1 > 0) \geq 1 - \epsilon.$$

Therefore, we fix  $\epsilon$  and choose  $c_3$  such that  $\frac{A_2}{c_3^2 A_1} \leq \epsilon$ , where  $A_1$  is defined in assumption (i). It follows that

$$\frac{A_2 S}{c_1^2} = A_2 \frac{n}{c_3^2 S} \leq \frac{A_2}{c_3^2 A_1} \leq \epsilon.$$

Then we choose  $c_2$  such that  $c_3 > 2c_2$  and observe that  $2 \sum_{s=1}^S e^{-2n_1 \delta_s^{(1)}} + 2 \sum_{s=1}^S e^{-2n_2 \delta_s^{(2)}} \rightarrow 0$  for  $n, S \rightarrow \infty$ . Since this is true for each  $\epsilon > 0$ , then  $P^{(1)}(T_1 > 0) \rightarrow 1$ , and similarly  $P^{(2)}(T_1 < 0) \rightarrow 1$ .

## References

- Bradley, R. C. (2005). Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys* 2, 107–144.
- Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning* 20, 273–297.
- Cover, T. M. and P. E. Hart (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 21–27.

- Geraci, M. (2016). Qtools: A collection of models and other tools for quantile inference. *R Journal* 8(2), 117–138.
- Geraci, M. (2020). *Qtools: Utilities for quantiles*. R package version 1.5.2.
- Geraci, M., N. S. Boghossian, A. Farcomeni, and J. D. Horbar (2020). Quantile contours and allometric modelling for risk classification of abnormal ratios with an application to asymmetric growth-restriction in preterm infants. *Statistical Methods in Medical Research* 29, 1769–1786.
- Hall, P., D. M. Titterton, and J.-H. Xue (2009). Median-based classifiers for high-dimensional data. *Journal of the American Statistical Society* 104, 1597–1608.
- Hand, D. and K. Yu (2001). Idiot’s Bayes - Not so stupid after all? *International Statistical Review* 69, 385–398.
- Hennig, C. and C. Viroli (2016a). Quantile-based classifiers. *Biometrika* 103(2), 435–446.
- Hennig, C. and C. Viroli (2016b). *quantileDA: Quantile classifier*. R package version 1.1.
- Kong, L. and I. Mizera (2012). Quantile tomography: Using quantiles with multivariate data. *Statistica Sinica* 22(4), 1589–1610.
- Lai, Y. and I. McLeod (2019). *eqc: Ensemble quantile classification*. R package version 1.2-2.
- Lai, Y. and I. McLeod (2020). Ensemble quantile classifier. *Computational Statistics & Data Analysis* 144, 106849.
- Lee, E.-K., D. Cook, S. Klinke, and T. Lumley (2005, Dec). Projection Pursuit for Exploratory Supervised Classification. *J. Comput. Graph. Stat.* 14(4), 831–846.
- Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch (2019). *e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien*. R package version 1.7-3.
- Park, M. Y. and T. Hastie (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics* 9, 30–50.

- Park, M. Y. and T. Hastie (2018). *stepAIC: L2 penalized logistic regression with stepwise variable selection*. R package version 0.93.
- R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Sorrentino, D., C. Avellini, M. Geraci, T. Dassopoulos, D. Zarifi, S. F. Vadalá di Prampero, and G. Benevento (2014). Tissue studies in screened first-degree relatives reveal a distinct Crohn’s disease phenotype. *Inflammatory Bowel Diseases* 20, 1049–1056.
- Stam, A. J. (1982). Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces. *Journal of Applied Probability* 19(1), 221–228.
- Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 99, 6567–6572.
- Venables, W. N. and B. D. Ripley (2002). *Modern applied statistics with S* (Fourth ed.). New York, NY: Springer.
- Wang, L., J. Zhu, and H. Zou (2008). Hybrid Huberized support vector machines for microarray classification and gene selection. *Bioinformatics* 24, 412–419.



Table 2: Misclassification rates averaged over 100 replications with standard errors in brackets for ten classifiers (DQC, directional quantile classifier; Centroid, centroid classifier; Median, median classifier; CQC, componentwise quantile classifier; EQC, ensemble quantile classifier; LDA, linear discriminant analysis; KNN, k-nearest neighbour; PLR, penalised logistic regression; SVM, support vector machines; Bayes, naive Bayes) in the first scenario where populations have normal marginals.

<i>Dimension <math>p</math></i>	<i>Uncorrelated</i>				<i>Correlated</i>			
	10	50	100	500	10	50	100	500
<i>Sample size <math>n = 50</math></i>								
DQC	0.294 (0.075)	0.111 (0.053)	0.047 (0.029)	0.000 (0.000)	0.304 (0.070)	0.113 (0.046)	0.045 (0.031)	0.000 (0.002)
Centroid	0.300 (0.066)	0.122 (0.048)	0.055 (0.031)	0.000 (0.000)	0.308 (0.065)	0.127 (0.051)	0.052 (0.033)	0.000 (0.002)
Median	0.334 (0.078)	0.173 (0.066)	0.096 (0.042)	0.004 (0.009)	0.343 (0.066)	0.184 (0.051)	0.095 (0.042)	0.002 (0.007)
CQC	0.349 (0.066)	0.211 (0.071)	0.205 (0.083)	0.075 (0.040)	0.350 (0.080)	0.227 (0.069)	0.197 (0.085)	0.072 (0.035)
EQC	0.344 (0.076)	0.184 (0.061)	0.097 (0.043)	0.002 (0.007)	0.347 (0.074)	0.197 (0.052)	0.105 (0.043)	0.002 (0.007)
LDA	0.317 (0.078)	0.390 (0.090)	0.212 (0.069)	0.148 (0.048)	0.318 (0.068)	0.391 (0.090)	0.206 (0.067)	0.143 (0.064)
KNN	0.358 (0.073)	0.182 (0.058)	0.104 (0.047)	0.010 (0.015)	0.351 (0.065)	0.178 (0.057)	0.108 (0.041)	0.004 (0.008)
PLR	0.310 (0.078)	0.146 (0.059)	0.064 (0.034)	0.000 (0.000)	0.313 (0.066)	0.155 (0.052)	0.066 (0.035)	0.000 (0.003)
SVM	0.319 (0.081)	0.124 (0.055)	0.055 (0.029)	0.000 (0.000)	0.330 (0.070)	0.133 (0.051)	0.053 (0.033)	0.000 (0.002)
Bayes	0.359 (0.080)	0.212 (0.059)	0.127 (0.049)	0.007 (0.013)	0.365 (0.077)	0.223 (0.060)	0.132 (0.051)	0.008 (0.012)
<i>Sample size <math>n = 100</math></i>								
DQC	0.272 (0.043)	0.092 (0.029)	0.027 (0.015)	0.000 (0.000)	0.282 (0.046)	0.089 (0.030)	0.031 (0.017)	0.000 (0.000)
Centroid	0.275 (0.043)	0.105 (0.032)	0.036 (0.016)	0.000 (0.000)	0.291 (0.046)	0.100 (0.030)	0.039 (0.021)	0.000 (0.000)
Median	0.314 (0.042)	0.154 (0.038)	0.073 (0.027)	0.000 (0.002)	0.326 (0.046)	0.151 (0.040)	0.076 (0.031)	0.001 (0.002)
CQC	0.330 (0.055)	0.178 (0.048)	0.103 (0.045)	0.066 (0.030)	0.340 (0.048)	0.168 (0.042)	0.103 (0.051)	0.072 (0.031)
EQC	0.320 (0.044)	0.164 (0.038)	0.079 (0.027)	0.001 (0.003)	0.326 (0.049)	0.161 (0.043)	0.081 (0.032)	0.001 (0.004)
LDA	0.284 (0.044)	0.190 (0.045)	0.379 (0.077)	0.058 (0.029)	0.301 (0.047)	0.178 (0.043)	0.368 (0.090)	0.055 (0.031)
KNN	0.344 (0.051)	0.175 (0.041)	0.103 (0.029)	0.002 (0.005)	0.357 (0.044)	0.176 (0.039)	0.109 (0.035)	0.003 (0.005)
PLR	0.289 (0.044)	0.138 (0.037)	0.049 (0.022)	0.000 (0.000)	0.299 (0.046)	0.128 (0.034)	0.052 (0.022)	0.000 (0.000)
SVM	0.279 (0.043)	0.126 (0.032)	0.039 (0.019)	0.000 (0.000)	0.305 (0.043)	0.111 (0.030)	0.041 (0.022)	0.000 (0.000)
Bayes	0.328 (0.048)	0.180 (0.039)	0.092 (0.030)	0.002 (0.004)	0.337 (0.050)	0.167 (0.040)	0.093 (0.033)	0.002 (0.004)
<i>Sample size <math>n = 500</math></i>								
DQC	0.263 (0.020)	0.079 (0.011)	0.022 (0.006)	0.000 (0.000)	0.261 (0.022)	0.078 (0.013)	0.024 (0.007)	0.000 (0.000)
Centroid	0.268 (0.019)	0.084 (0.011)	0.025 (0.007)	0.000 (0.000)	0.266 (0.021)	0.083 (0.012)	0.026 (0.007)	0.000 (0.000)
Median	0.303 (0.020)	0.126 (0.015)	0.053 (0.010)	0.000 (0.000)	0.300 (0.020)	0.124 (0.014)	0.053 (0.012)	0.000 (0.000)
CQC	0.308 (0.018)	0.132 (0.016)	0.056 (0.012)	0.006 (0.006)	0.304 (0.025)	0.132 (0.017)	0.056 (0.012)	0.005 (0.004)
EQC	0.306 (0.019)	0.131 (0.015)	0.057 (0.011)	0.000 (0.001)	0.305 (0.023)	0.130 (0.016)	0.055 (0.011)	0.000 (0.001)
LDA	0.270 (0.019)	0.095 (0.013)	0.041 (0.010)	0.358 (0.051)	0.268 (0.022)	0.097 (0.014)	0.041 (0.009)	0.364 (0.061)
KNN	0.325 (0.023)	0.141 (0.016)	0.069 (0.012)	0.003 (0.002)	0.326 (0.022)	0.173 (0.016)	0.075 (0.013)	0.003 (0.003)
PLR	0.270 (0.019)	0.099 (0.014)	0.038 (0.010)	0.000 (0.000)	0.268 (0.022)	0.102 (0.015)	0.039 (0.009)	0.000 (0.000)
SVM	0.267 (0.020)	0.090 (0.012)	0.030 (0.007)	0.000 (0.000)	0.268 (0.021)	0.097 (0.015)	0.030 (0.008)	0.000 (0.000)
Bayes	0.287 (0.020)	0.112 (0.013)	0.044 (0.009)	0.000 (0.000)	0.285 (0.022)	0.113 (0.015)	0.045 (0.010)	0.000 (0.000)

Table 3: Leave-one out estimates of the misclassification rates for the Crohn’s disease dataset ( $p = 5$ ) and its noisy version ( $p = 50$ ) using ten classifiers (DQC, directional quantile classifier; Centroid, centroid classifier; Median, median classifier; CQC, componentwise quantile classifier; EQC, ensemble quantile classifier; LDA, linear discriminant analysis; KNN,  $k$ -nearest neighbour; PLR, penalised logistic regression; SVM, support vector machines; Bayes, naïve Bayes).

	$p = 5$	$p = 50$
DQC	0.229	0.229
Centroid	0.286	0.286
Median	0.400	0.400
CQC	0.314	0.343
EQC	0.314	0.314
LDA	0.257	0.543
KNN	0.371	0.343
PLR	0.286	0.343
SVM	0.257	0.257
Bayes	0.286	0.257