Biometrika (2015), **0**, 0, *pp*. 1–12

Printed in Great Britain

Quantile-based classifiers

BY C. HENNIG
Department of Statistical Science, University College London, London, WC1E 6BT, U.K.
c.hennig@ucl.ac.uk

AND C. VIROLI
Department of Statistical Sciences, University of Bologna, Bologna, 40126, Italy
cinzia.viroli@unibo.it

SUMMARY

Classification with small samples of high-dimensional data is important in many areas. Quantile classifiers are distance-based classifiers that require a single parameter, regardless of the dimension, and classify observations according to a sum of weighted component-wise distances of the components of an observation to the within-class quantiles. An optimal percentage for the quantiles can be chosen by minimizing the misclassification error in the training sample. It is shown that this choice is consistent for the classification rule with the asymptotically optimal quantile, and that, under some assumptions, when the number of variables goes to infinity, the probability of correct classification converges to unity. The role of skewness of the distributions of the predictor variables is discussed. The optimal quantile classifier gives low misclassification rates in a comprehensive simulation study and a real data set.

Some key words: high-dimensional data; median-based classifier; misclassification rate; skewness.

1. INTRODUCTION

Supervised classification has received wide interest in the scientific literature. Classification methods can be broadly divided into parametric methods, which make distributional assumptions (e.g., Hastie & Tibishirani, 1996; Bensmail & Celeux, 1996; Fraley & Raftery, 2002; Hand & Yu, 2001), and nonparametric methods, which concentrate on the local vicinity of the point to be classified (e.g., Cover & Hart, 1967; Mika et al., 1999). Implementing classification methods in high dimensions can be computationally demanding, because of the curse of dimensionality (Bellman, 1961). A way to address this problem is to rely on portions of the conditional distribution of the features given the class labels. Distance-based classifiers use the partial information of the class conditional distributions: centroid-based methods have been successfully used for gene expression data (Tibshirani et al., 2002; Dudoit et al., 2002; Dabney, 2005; Fan & Fan, 2008), and median-based classifiers (Jörnsten, 2004; Ghosh & Chaudhuri, 2005) represent a more robust alternative when distributions have heavy tails. Hall et al. (2009) proposed a component-wise median-based classifier that works well in high dimensions, by assigning a new observed vector to the class from which it has the smallest L_1-distance in the training set. All these methods consider the distance from the core of a distribution to be the major source of discriminatory information, but other quantities may contain information relevant for classification.

In this work we define and explore classifiers based on the quantiles of the class conditional distributions. More specifically, by defining a natural distance for quantiles, which includes the

L_1-distance to the component-wise median as special case, we will obtain the quantile classifier dependent on the θ quantile, $\theta \in [0, 1]$. The optimal θ chosen in the training set will define the empirically optimal quantile classifier. We establish the consistency of this choice for the θ that yields the optimal true correct classification probability as the sample size $n \to \infty$. We show that under certain assumptions the correct classification probability converges to unity as the number of variables p increases with the sample size, as Hall et al. (2009) did for the median classifier.

2. THE CLASSIFICATION RULE

2·1. *Distance-based classifiers*

We consider constructing a quantile distance-based discriminant rule for classifying new observations into one of two populations or classes. Let Π_0 and Π_1 be populations described by the random variables X and Y with probability densities P_0 and P_1 on $\mathbb{R}^p$d. Distance-based classifiers (Jörnsten, 2004; Tibshirani et al., 2003; Hall et al., 2009) assign a new datum $z = (z_1, \ldots, z_p)$ to the population to which it is closest. More specifically, the decision rule allocates z to Π_0 if

$$\sum_{j=1}^{p} \{d(z_j, \xi_Y) - d(z_j, \xi_X)\} > 0, \tag{1}$$

where $\xi_X = (\xi_{X1}, \ldots, \xi_{Xp})$ and $\xi_Y = (\xi_{Y1}, \ldots, \xi_{Yp})$ are p-variate moments of populations Π_0 and Π_1 and $d(\cdot)$ denotes a specific distance measure. The rule in (1) includes centroid classifiers (Tibshirani et al., 2002, 2003; Wang & Zhu, 2007), for which $\xi_X = \{E(X_1), \ldots, E(X_p)\}$, $\xi_Y = \{E(Y_1), \ldots, E(Y_p)\}$, and $d(\cdot)$ is the squared difference, so that the sum is the squared L_2-distance between z and the mean vector, and the median-based classifier (Hall et al., 2009) defined by choosing $\xi_X,\ \xi_Y$ as the component-wise medians of P_0 and P_1, respectively, and $d(\cdot)$ as the absolute value, so that the sum is the L_1-distance between z and the component-wise median. These definitions are population-based. For finite samples, population quantities are replaced by sample analogues.

The choice of the metric L_1 for medians, instead of L_2, is motivated by the fact that the mean vector is the statistic that minimizes the sum of L_2-distances of points to the centroid, whereas the median minimizes the sum of the corresponding L_1-distances.

2·2. *The quantile classifier*

We now introduce a family of the quantile classifiers that includes the median classifier as special case. By definition, the θ-quantile of a univariate random variable, say U, with probability distribution function F_U, denoted by $q_U(\theta)$, solves the equation $q_U(\theta) = F_U^{-1}(\theta) = \inf\{u : F_U(u) \geq \theta\}$, with $\theta \in [0, 1]$. The θ-quantile of F_U is the value q that minimizes

$$\theta \int_{u>q} |u - q| dF_U(u) + (1 - \theta) \int_{u<q} |u - q| dF_U(u), \tag{2}$$

where $|\cdot|$ denotes absolute value. Equation (2) takes its minimum for $F_U(q) = \theta$. For observations $u_1, \ldots, u_n$, the empirical θ-quantile of U minimizes the sample counterpart of (2):

$$\theta \sum_{u_i>q} |u_i - q| + (1 - \theta) \sum_{u_i \leq q} |u_i - q| = \sum_{i=1}^{n} \left\{\theta + (1 - 2\theta)\mathbb{1}_{(u_i \leq q)}\right\} |u_i - q|, \tag{3}$$

where the indicator function $\mathbb{1}_{(u_i \le q)}$ is unity if $u_i \le q$ and zero otherwise. Expression (3) is used to define the quantile-based classifier. Let

$$\Phi_{kj}(z,\theta) = \left[\theta + \{1-2\theta\}\mathbb{1}_{\left\{z_j \le q_{kj}(\theta)\right\}}\right] |z_j - q_{kj}(\theta)|, \quad j = 1,\ldots,p, \quad k = 0,1,$$
$$\Phi_j(z,\theta,q) = \left\{\theta + (1-2\theta)\mathbb{1}_{(z_j \le q)}\right\} |z_j - q|,$$

where $q_{0j}(\theta),\ q_{1j}(\theta)$ are the marginal quantile functions of $P_0,\ P_1$ evaluated at θ. Given two sets of observations from the two populations Π_0 and Π_1, $x_1,\ldots,x_{n_0}$ and $y_1,\ldots,y_{n_1}$ and a new observation $z = (z_1,\ldots,z_p) \in \mathbb{R}^p$, let

$$s(z,\theta) = \sum_{j=1}^{p} \{\Phi_{1j}(z,\theta) - \Phi_{0j}(z,\theta)\}. \tag{4}$$

Then z is assigned to Π_0 if $s(z,\theta) > 0$ and to Π_1 otherwise.

Remark 1. The application of (4) to more than $g = 2$ classes is straightforward. The quantile classifier rule for allocating an observation z to one of g populations $\Pi_1,\ldots,\Pi_g$ is to allocate z to the population which gives the lowest quantile distance $\sum_{j=1}^{p} \Phi_{kj}(z,\theta)$, $(k = 1,\ldots,g)$.

Remark 2. Expression (4) coincides with the median classifier for $\theta = 0{\cdot}5$.

Given the two populations, Π_0 and Π_1 with prior probabilities π_0 and π_1, respectively, the probability of correct classification of the quantile classifier, based on the true quantiles, is

$$\Psi(\theta) = \pi_0 \int \mathbb{1}_{\{s(z,\theta)>0\}} dP_0(z) + \pi_1 \int \mathbb{1}_{\{s(z,\theta)\le 0\}} dP_1(z). \tag{5}$$

In the Supplementary Material we show that there is a straightforward formula to compute (5) for $p = 1$, and that with θ maximizing (5) the quantile classifier equals the optimal Bayes classifier in many cases. Figure 1 shows a number of univariate examples for how the theoretical misclassification rates $1 - \Psi(\theta)$ change with θ, assuming $\pi_0 = \pi_1 = 0{\cdot}5$. The value $\theta = 0{\cdot}5$ is optimal for symmetric distributions of equal shape, but in case of distributions of different shapes or with skewness, other values of θ improve the misclassification rate. Experiments show that the shapes of the curves on the right hand side of Figure 1, including the location of the minimum, can be estimated fairly accurately by misclassification probabilities within the training sample if n is large enough.

2·3. *The empirically optimal quantile classifier*

We address the choice of the quantile value in the family of possible quantile classifiers by selecting the optimum θ based on misclassification rates in the training sample.

Let $(Z_1,C_1),(Z_2,C_2),\ldots$ be $\mathbb{R}^p \times \{0,1\}$-valued independent and identically distributed random variables. Let Z_1 be distributed according to a two-component mixture of distributions $P_0 = \mathcal{L}(Z_1 \mid C_1 = 0)$ and $P_1 = \mathcal{L}(Z_1 \mid C_1 = 1)$. Let $\pi_0 = \text{pr}(C_1 = 0)$ and $\pi_1 = \pi_0$. Let $P_{01},\ldots,P_{0p}$ denote the marginal distributions of P_0, analogously $P_{11},\ldots,P_{1p}$. For arbitrarily small $0 < \tau < 1/2$ define $T = [\tau, 1-\tau]$. For $\theta \in [0,1]$ $(j = 1,\ldots,p;\ k = 0,1)$ let $q_{kj}(\theta)$ denote the θ-quantile of P_{kj}. For given $(Z_1,C_1),\ldots,(Z_n,C_n)$ let $\hat{q}_{kjn}(\theta)$ be the empirical θ-quantile for the subsample defined by $C_i = k$ $(i = 1,\ldots,n)$. The notation $\Phi_{kj}(z,\theta)$ is used for $\Phi_j\{z,\theta,q_{kj}(\theta)\}$ and $\Phi_{kjn}(z,\theta)$ is used for $\Phi_j\{z,\theta,\hat{q}_{kjn}(\theta)\}$.

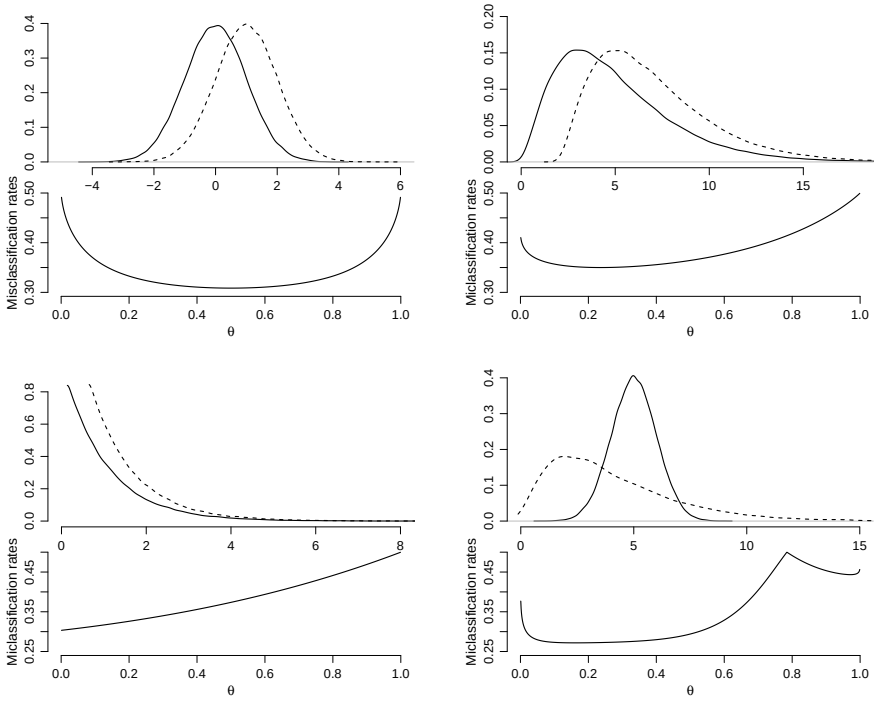

Fig. 1. Theoretical misclassification rates in four different scenarios with equal class prior probabilities. First panel: probability density functions of two location-shifted Gaussians and the corresponding misclassification rates as function of θ. Second panel: two location-shifted chi-squared distributions and the corresponding misclassification function. Third panel: misclassification rates for two location-shifted exponentials. Forth panel: misclassification rates for a Gaussian vs a chi-squared distribution.

The empirically optimal quantile classifier is defined by assigning Z to Π_0 if

$$s_n(Z, \hat{\theta}_n) = \sum_{j=1}^{p} \{\Phi_{1jn}(Z, \hat{\theta}_n) - \Phi_{0jn}(Z, \hat{\theta}_n)\} > 0, \tag{6}$$

where $\hat{\theta}_n = \arg\max_{\theta \in T} \Psi_n(\theta)$ is the estimated optimal θ from $(Z_1, C_1), \ldots, (Z_n, C_n)$; if the argmax is not unique, any maximizer can be chosen. The observed rate of correct classification for θ in data $(z_1, c_1), \ldots, (z_n, c_n)$ is

$$\Psi_n(\theta) = \frac{1}{n} \left[\sum_{i:\ c_i=0} \mathbb{1}_{\{s_n(Z,\theta)>0\}} + \sum_{i:\ c_i=1} \mathbb{1}_{\{s_n(Z,\theta)\leq 0\}} \right].$$

We look for the optimal value of θ in T, a closed interval not containing zero. In practice, a small nonzero τ needs to be chosen, and $\Psi_n(\theta)$ is evaluated on a grid of equispaced values between τ and $1-\tau$, where τ should be chosen large enough that there is sufficient information to estimate the τ-quantile. The parameter τ should not be seen as a crucial tuning parameter of the method; we recommend to choose it as small as possible in order to find the empirical optimum of θ, while ensuring that the estimated τ-quantile still is of some use.

In case of equal misclassification rates for different values of θ, which can arise for data sets with small n, we recommend fitting a square polynomial to the misclassification rate as function of θ and choosing the optimum θ by evaluating this polynomial at the empirically optimal values of θ.

3. ASYMPTOTIC THEORY

3·1. *Consistency of the quantile classifier*

The asymptotic probability of correct classification of the quantile classifier is defined in (5). Let $\tilde{\theta} = \arg\max_{\theta \in T} \Psi(\theta)$ be the optimal θ regarding the true model. We make the following assumptions.

A1. For all $j = 1, \ldots, p$ and $k = 0, 1$ let q_{kj} be a continuous function of $\theta \in T$;

A2. For all $\theta \in T$, pr $\left[\sum_{j=1}^{p} \{\Phi_{1j}(Z, \theta) - \Phi_{0j}(Z, \theta)\} = 0\right] = 0$.

THEOREM 1. *Under A1 and A2, for any* $\epsilon > 0$,

$$\lim_{n \to \infty} \text{pr}\{|\Psi(\tilde{\theta}) - \Psi(\hat{\theta}_n)| > \epsilon\} = 0. \tag{7}$$

This means that in large samples the empirically optimal $\hat{\theta}_n$ in the quantile classifier achieves the true correct classification probability for the true optimal θ, and is therefore at least as good as the median classifier. Proofs are given in the Supplementary Material.

3·2. *A result for* $p \to \infty$

Theorem 1 refers to $n \to \infty$ for fixed finite p. In many modern applications p is larger than n, so that results for $p \to \infty$ seem more appealing.

Hall et al. (2009) proved under certain conditions that the misclassification probability of the median classifier converges to zero for $n, p \to \infty$. Their proof adapts in a more or less straightforward manner to classifiers based on any fixed quantile. The result presented here requires the assumptions of Hall et al. (2009) to hold uniformly for a range of quantiles. This is stronger than in Hall et al. (2009), and reflects the fact that the quantile classifier is more difficult theoretically than the median classifier. The arguments of Hall et al. (2009) then carry over to classifiers based on whatever quantile selection rule is chosen, including selecting the empirically optimal one. In this sense, regarding $p \to \infty$, we only establish that the quantile classifier is as good as the median classifier, under stronger conditions. For finite n it can improve on the median classifier, as shown empirically in the simulation and in Theorem 1.

Again let $T = [\tau, 1 - \tau]$ for arbitrarily small $0 < \tau < 1/2$. Let $U = (U_1, U_2, \ldots)$ denote an infinite sequence of random variables, each U_i with θ-quantiles $q_i(\theta)$ for all $\theta \in T$ and median zero. Assume that there is at most one value x with $F_{U_i}(x) = \theta$ for all $\theta \in T$. For infinite sequences of constants $(\nu_{X_1,\frac{1}{2}}, \nu_{X_2,\frac{1}{2}}, \ldots)$, $(\nu_{Y_1,\frac{1}{2}}, \nu_{Y_2,\frac{1}{2}}, \ldots)$, assume that for each p, the p-vectors $X_1, \ldots, X_m$ are identically distributed as $(\nu_{X_1,\frac{1}{2}} + U_1, \ldots, \nu_{X_p,\frac{1}{2}} + U_p)$, and the p-vectors $Y_1, \ldots, Y_n$ are identically distributed as $(\nu_{Y_1,\frac{1}{2}} + U_1, \ldots, \nu_{Y_p,\frac{1}{2}} + U_p)$. For $i \geq 1$ define the quantiles $\nu_{X_i,\theta} = \nu_{X_i,\frac{1}{2}} + q_i(\theta)$, $\nu_{Y_i,\theta} = \nu_{Y_i,\frac{1}{2}} + q_i(\theta)$. Let C be a $[0, 1]$-valued random variable and assume Z to be distributed as X_1 if $C = 0$ and as Y_1 if $C = 1$, and $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ and (Z, C) as totally independent.

The following assumptions are needed:

B1. $\lim_{\lambda \to \infty} \sup_{k \geq 1} E\{|U_k| \mathbb{1}_{(|U_k| > \lambda)}\} = 0$;

B2. for each $c > 0$,

$$\inf_{k\geq 1}\inf_{|x|\geq c}\inf_{\theta\in T}\left(E[\Phi_k\{U,\theta,q_k(\theta)+x\}]-E[\Phi_k\{U,\theta,q_k(\theta)\}]\right)>0;$$

B3. for each $\epsilon > 0$,

$$\inf_{k\geq 1}\inf_{\theta\in T}\left(\min[\theta-\text{pr}\{U_k\leq q_k(\theta)-\epsilon\},1-\theta-\text{pr}\{U_k\geq q_k(\theta)+\epsilon\}]\right)>0;$$

B4. with $\mathcal{B}$ denoting the class of Borel subsets of the real line,

$$\lim_{k\to\infty}\sup_{k_1,k_2:\ |k_1-k_2|\geq k}\sup_{B_1,B_2\in\mathcal{B}}|\text{pr}(U_{k_1}\in B_1,U_{k_2}\in B_2)-\text{pr}(U_{k_1}\in B_1)\text{pr}(U_{k_2}\in B_2)|=0;$$

B5. the differences $|\nu_{Xk,\theta}-\nu_{Yk,\theta}|$ are uniformly bounded; and

B6. for sufficiently small $\epsilon > 0$, the proportion of values $k \in [1, p]$ for which $|\nu_{Xk,\theta}-\nu_{Yk,\theta}| > \epsilon$ for all $\theta \in T$ is bounded away from zero as p diverges.

Assumptions B1 and B4 are identical to (4.1) and (4.4) in Hall et al. (2009). Assumptions B2, B3, B5 and B6 are (4.2), (4.3), (4.5), (4.6) in Hall et al. (2009), enforced uniformly for $\theta \in T$. Assumption B4 is a standard α-mixing condition, which implies that variables with very different index numbers are approximately independent, and B6 implies that there is an infinite amount of variables relevant for telling the classes apart. Assumptions B1 and B5 are needed, given B6, to prevent classification from being dominated by a single or a finite number of variables, and B2 and B3 concern uniform continuity and well-definedness of the quantiles. See Hall et al. (2009) for further discussion.

Let $R:\ \mathbb{N}\mapsto T$ be any quantile selection rule. Let $\mathcal{R}_{m,n,i}$ $(i \in \mathbb{N})$ be the sequence of $\{0,1\}$-valued $R(i)$-quantile classifiers computed from $\{(X_1,0),\ldots,(X_m,0),(Y_1,1),\ldots,(Y_n,1)\}$.

THEOREM 2. *Assume B1–B6 hold and that both n and m diverge as $p\to\infty$. Then, with probability converging to 1 as p increases, the classifier $\mathcal{R}_{m,n,p}$ makes the correct decision, i.e.,*

$$\text{pr}\{\mathcal{R}_{m,n,p}(Z)=1\mid C=0\}+\text{pr}\{\mathcal{R}_{m,n,p}(Z)=0\mid C=1\}\to 0. \tag{8}$$

4. SOME ISSUES AND EXTENSIONS

4·1. *Standardization*

Like other classifiers, the quantile classifier depends on the scaling of the variables. This dependence can be removed by standardizing them. Standardization can be seen as implicit reweighting of the variables. Optimally, variables are treated in such a way that their relative weights reflect their relative information content for classification.

In practice standardization may be inadvisable when variables have the same measurement units and there are reasons to expect that the information content of the variables for classification may be indicated by their variation. This will play a role in the example in Section 5·2. Cross-validation could also help to decide if standardization is beneficial.

Where variables are standardized, transformation to unit pooled within-class variance as estimated from the training data can be expected to improve classification performance compared with the overall variance, because the separation between classes may contribute strongly to the overall variance. Thus variables with a strong separation between classes and hence a large amount of classification information will be implicitly downweighted, whereas standardization to unit pooled within-class variance will downweight variables for which the classes are heterogeneous and which are therefore not so useful for classification.

4·2. *Individual treatment of variables*

The empirically optimal quantile classifier is based on a single θ that is optimal for all variables simultaneously. We tried out ways to choose individual θ-values for each variable, but none of these improved on the quantile classifier based on a single θ on independent test data. However, we found a simple method to increase adaptation to the individual variables, which led to a clear improvement in some situations while not making things clearly worse elsewhere. As in the univariate setting, the optimal θ depends on the skewness of the distributions involved. In practice, a set of $p > 1$ measurements could be skewed in different directions, giving conflicting messages about what values of θ are to be preferred. In order to overcome this, we recommend changing the direction of skewness of variables by applying sign changes so that all the variables have the same direction of skewness.

More specifically, compute a skewness measure separately for each variable, such as the conventional third standardized empirical moment or, alternatively, a measure from the family of the robust quantile-based quantities (Hinkley, 1975),

$$\tau(u) = \frac{F^{-1}(u) + F^{-1}(1-u) - 2F^{-1}(1/2)}{F^{-1}(u) - F^{-1}(1-u)},$$

where F denotes the marginal cumulative distribution function and u a fixed value in the interval $[0{\cdot}5, 1]$. When $u = 3/4$, this corresponds to Galton's skewness measure, and $u = 0{\cdot}1$ gives the less robust Kelley skewness measure (Johnson et al., 1994). Evaluate the amount of skewness of each variable separately within classes, and then summarize by averaging all the within-class measures with equal weights. The signs of variables with negative aggregated within-class skewness are then changed, so that finally the variables used for the quantile estimator all have positive skewness. The adjustment takes into account the individuality of the variables in a rather rough and only empirically founded way. In general, the connection between skewness and optimal θ is not straightforward, so there is little hope of employing skewness in a more sophisticated way. The results in Sections 3·1 and 3·2 carry over if the skewness of all variables is estimated correctly with probability 1 for large enough n.

4·3. *Asymmetric loss functions*

In many applications of supervised classification, different losses are associated with misclassifications from or into different classes. The current work focuses on the misclassification probability with symmetric loss, i.e., equal loss for all types of misclassification, but can be adapted to other loss functions.

One approach is to choose the optimal quantile minimizing the appropriate within-sample loss. A better option may be to choose the decision boundary different from 0, i.e., $s_n(Z, \hat{\theta}_n) > c$ instead of $s_n(Z, \hat{\theta}_n) > 0$ in (6), after having chosen $\hat{\theta}_n$ to optimize the misclassification rate with symmetric loss.

The latter approach seems reasonable because the boundary $s_n(Z, \hat{\theta}_n) > 0$ implicitly treats the two classes symmetrically. The role of $\hat{\theta}_n$ is to account for the distributional shapes, whereas c accounts for asymmetric loss. Preliminary simulations show that under asymmetric loss the quantile classifier with this approach competes with other methods as well as in the simulations in Section 5·1.

4·4. *Dependence*

Distance-based classifiers defined according to (1), including the quantile classifier, aggregate information from the variables without taking dependence between them into account. The simulations in Section 5·1 show that there can be good classification results in case of dependence.

Distance-based classifiers do not require independence, but only that the information that separates the classes can be picked up from the original variables. Using marginal information from all variables is a way to avoid overfitting with high-dimensional data different from dimension reduction. This strategy is superior to dimension-reduction approaches in some situations and inferior in others; no classifier can be expected to be universally optimal. In situations in which the classification information is poorly represented in the original variables, one could apply the quantile classifier to principal components or independent components (Hyvärinen et al., 2001). Preliminary simulations show that using principal components may improve results with approximately normal distributions, but may give unstable results for high-dimensional data with skew or heavy-tailed marginal distributions.

5. NUMERICAL RESULTS

5·1. *Simulation study*

We evaluated the performance of the quantile classifier by simulation. We generated p vectors from $g = 2$ populations in four scenarios. In the first scenario we considered symmetric Student *t*-distributed variables W_j ($j = 1, \ldots, p$) with 3 degrees of freedom. We simulated two location-shifted populations from W_j as $X_j \sim W_j$ and $Y_j \sim W_j + 0{\cdot}5$. In the second scenario we tested the behavior of the classifiers in highly skewed data, by generating identically distributed vectors W_j ($j = 1, \ldots, p$) from a multivariate Gaussian distribution, transforming them using the exponential function, $X_j \sim \exp(W_j)$ and $Y_j \sim \exp(W_j) + 0{\cdot}2$. In the third scenario we considered different distributions for the p variables. We first generated W_j from a multivariate Gaussian distribution and then we split p in 5 balanced blocks to which we applied different transformations: (a) $X_j \sim W_j$ and $Y_j \sim W_j + 0{\cdot}2$, (b) $X_j \sim \exp(W_j)$ and $Y_j \sim \exp(W_j) + 0{\cdot}2$, (c) $X_j \sim \log|W_j|$ and $Y_j \sim \log|W_j| + 0{\cdot}2$, (d) $X_j \sim W_j^2$ and $Y_j \sim W_j^2 + 0{\cdot}2$, (e) $X_j \sim |W_j|^{0{\cdot}5}$ and $Y_j \sim |W_j|^{0{\cdot}5} + 0{\cdot}2$. In the fourth scenario we simulated different distributional shapes and levels of skewness even for different classes within the same variable. Within each class, data were generated according to Beta distributions with parameters a and b in the interval $(0{\cdot}1, 10)$ randomly generated for each class within each variable. Within each class data were centered about 0, so that information about class differences was only in the distributional shape.

For each of the four scenarios we evaluated the combination of $p = 50, 100, 500$, $n = 50, 100, 500$, different percentages of relevant variables for classification, i.e., 100%, 50%, and 10%, independent or dependent variables, and, in the fourth scenario, balanced and unbalanced classes, with class weights $0{\cdot}25$ and $0{\cdot}75$, for a total of 216 different settings. The dependence structure between the variables was introduced by generating varying correlated variables $W_1, \ldots, W_p$ from a Gaussian distribution with random correlation matrix based on the method proposed by Joe (2006), so that correlation matrices are uniformly distributed over the space of positive definite correlation matrices, with each correlation marginally distributed as Beta($p/2$,$p/2$) on $(-1, 1)$. The irrelevant noise variables were generated independently of each other from a Gaussian distribution. Variables were standardized to unit within-class pooled variance in the third scenario but not standardized in the three others, because in the third scenario the scales of the variables seem incompatible, whereas in for datasets like those from the other scenarios the reasons against standardization given in Section 4·1 may apply. For each setting we simulated 100 data sets as training sets and 100 test sets. In the case of balanced classes, the pairs of data sets were split into the two populations with sample size $n/2$ each; in the fourth scenario we also considered the unbalanced setting with $n/4$ and $3n/4$ observations in each class.

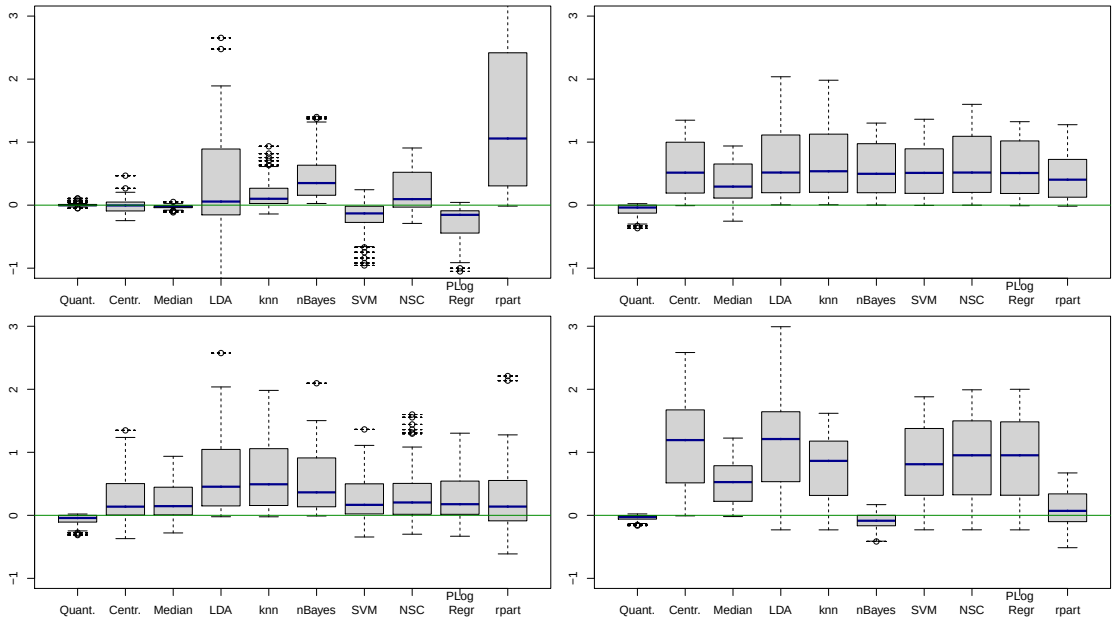

Fig. 2. Relative performance of the classifiers with respect to the quantile classifier with Galton skewness correction. The x-axis labels refer to different methods: LDA denotes linear discriminant analysis, knn is the k-nearest neighbor, SVM refers to support vector machines, NSC to the nearest shrunken centroid, PLog Regr to the penalized logistic regression and rpart to regression trees. The four panels show the distribution of the misclassification rates for (a) identically distributed symmetric variables, (b) asymmetric variables, (c) different distributions of variables and (d) different distributions of classes within variables in balanced and unbalanced populations.

The quantile based classifier has been implemented in the R package `quantileDA`, available from CRAN. Data were preprocessed by the skewness correction discussed in Section 4·2 using the conventional skewness measure and Galton's measure. In each setting we evaluated the classifier on a grid of equispaced values θ in $T = [\tau, 1-\tau]$ with $\tau = 0{\cdot}02$. In practice, τ could be tuned to the sample size n as, say, $\tau = 5/n$. The optimal θ was chosen in each training set.

We compared the quantile classifier's misclassification rates with those for nine other classifiers: the centroid and the median classifier, Fisher's linear discriminant, the k-nearest neighbor classifier (Cover & Hart, 1967), the naive Bayes classifier (Hand & Yu, 2001), the support vector machine (Cortes & Vapnik, 1995; Wang et al., 2008), the nearest shrunken centroid method (Tibshirani et al., 2002), penalized logistic regression (Park & Hastie, 2008), and classification trees (Breiman et al., 1984). Details about the implementation and parameter tuning of these methods are in the Supplementary Material. This includes results for the median classifier on Box–Cox transformed data in order to deal with skewness, which did not improve the median classifier's results by much. We computed the relative performance of each classifier with respect to the Galton quantile classifier's misclassification rates. More specifically, we evaluated the misclassification rates of each classifier as error rate minus the Galton quantile error rate divided by the average error rate in the given setting. The aggregated distribution of these rescaled results for

different choices of p, n, dependence/independence and the percentage of relevant variables is represented in the boxplots of Figure 2; see also the Supplementary Material.

For all methods, the misclassification rates decrease as the sample size increases. With reference to the quantile classifier, the the choice of the optimal θ appears more consistent as the sample size increases, and consequently the discriminative power of the method increases. Not surprisingly, classification performance worsens as the number of irrelevant variables increases. For fixed sample size and percentage of relevant variables, the methods seem to perform better as p increases in almost all settings.

The quantile classifier performs very well in most situations compared to the other classifiers. In the scenarios with equal distributional shapes and symmetric variables, the performance of the quantile classifiers is similar to those of the centroid and the median classifiers. The chosen optimal value of θ is on average close to the midpoint $0{\cdot}5$. Penalized logistic regression and support vector machines outperform the other methods in this scenario. In the settings with equal distributional shapes and asymmetric variables, the quantile classifiers outperform all other methods clearly and more or less uniformly. Here, the skewness correction according to the conventional third standardized moment seems to produce a slightly better classification performance compared to the Galton correction. However, the Galton skewness correction is preferable when analyzing real data more sensitive to outliers, as will be shown in Section 5·2. With different distributions of variables, the quantile classifiers again show excellent results. The overall results of support vector machines, nearest shrunken centroid, and penalized logistic regression are not much worse than those of the quantile classifier, but they are rarely significantly better and sometimes clearly worse. The fourth scenario with Beta distributions differing between variables and classes within variables is again generally dominated by the quantile classifiers, with only the naive Bayes classifier achieving better results overall. Overall, the methods that compete well with the quantile classifiers in one or two scenarios fall clearly behind in some others. Generally, the rankings of the methods do not strongly change with dependence, and the quantile classifier is still best where it was best under independence, although k-nearest neighbor, support vector machines, nearest shrunken centroid and penalized logistic regression are less affected by dependence.

5·2. *Real data example*

For illustration, we apply the quantile classifier to data collected testing a new method to detect bioaerosol particles based on gaseous plasma electrochemistry. The presence of such particles in air has a big impact on health, but monitoring bioaerosols poses great technical challenges. Sarantaridis et al. (2012) attempted to tell several different bioaerosols apart based on voltage changes over time on eight different electrodes when particles passed a premixed laminar hydrogen/oxygen/nitrogen flame. The resulting data are eight time series with 301 observations for each particle. Sarantaridis et al. (2012) discussed how the relevant information in every time series can be summarized in six characteristic features, namely maximum voltage in series, minimum voltage in series, maximum voltage change caused by electrode, difference between final and initial voltage, length of positive change caused by the electrode, length of negative change caused by the electrode. A seventh variable used in Sarantaridis et al. (2012) is omitted here based on recommendation of the chemists. We are therefore left with 48 variables.

We apply a variable standardization scheme driven by subject knowledge, which is motivated by the expectation of the chemists that the size of variation in voltage and length of effect is informative and that electrodes and variables for which the electrode causes stronger variation are actually more important for discrimination. The first four variables related to voltage on one hand and the variables related to the lengths on the other hand do not have comparable measure-

Table 1. *Leave-one-out cross-validated misclassification rates (%) of the bioaerosol particle data. In brackets are standard errors.*

Method	Misclassification rates
Quantile Classifier (no skewness correction)	13·3 (4·4)
Quantile Classifier (Galton correction)	3·3 (2·3)
Quantile Classifier (Skewness correction)	11·7 (4·2)
Centroid Classifier	21·7 (5·4)
Median Classifier	26·7 (5·8)
Linear Discriminant Analysis	6·7 (3·2)
k-nearest neighbor	15·0 (4·6)
Naive Bayes	15·0 (4·6)
Support vector machines	10·0 (3·9)
Nearest shrunken centroid	26·7 (5·8)
Penalized logistic regression	10·0 (3·9)
Classification trees	40·0 (6·4)

ment units. Therefore we computed one standard deviation from all 8×4 voltage variables and standardized all these variables by the same standard deviation. The 8×2 effect length variables were also standardized by the standard deviation computed from all of them combined.

We confine ourselves to distinguishing between two bioaerosols, Bermuda Smut Spores and Black Walnut Pollen, with data from 30 particles. The quantile classifier has been applied on non-preprocessed data and on data with sign adjustments according to the conventional and Galton skewness. We used leave-one-out cross-validation to assess the performance of the classifier. Within each fold we selected the optimal θ in the training set. Table 1 contains the misclassification rates of the quantile classifier according to the different preprocessing strategies, and of the discriminant methods of Section 5·1. The quantile classifier with Galton skewness correction is particularly effective for classifying the two bioaerosols. Only two particles are misclassified.

The sign adjustment preprocessing step is particularly relevant here. Without sign adjustment, the choice of the optimal quantile value is more variable across the cross validated sets and closer to the midpoint on average because of the possible different directions of skewness in the observed variables. In this case, when data are preprocessed using Galton skewness, the selected optimal θ across the cross-validated sets is always very small, with average $0{\cdot}04$, so more discriminant information between the two bioaerosols is contained in the left tail of the distributions than in their core.

SUPPLEMENTARY MATERIAL

Proofs and detailed simulation results can be found in the Supplementary Material for this paper.

ACKNOWLEDGEMENT

We are grateful to the editor, associate editor and the referees for their helpful comments and suggestions.

REFERENCES

BELLMAN, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.

BENSMAIL, H. & CELEUX, G. (1996). Regularized Gaussian Discriminant Analysis through eigenvalue decomposition. *Journal of the American Statistical Society* **91**, 1743–1748.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. & STONE, C. J. (1984). *Classification and Regression Trees*. Belmont CA: Wadsworth.

CORTES, C. & VAPNIK, V. (1995). Support-Vector Networks. *Machine Learning* **20**, 273–297.

COVER, T. M. & HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13**, 21–27.

DABNEY, A. R. (2005). Classification of microarrays to nearest centroids. *Bioinformatics* **21**, 4148–4154.

DUDOIT, S., FRIDLYAND, J. & SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Society* **97**, 77–87.

FAN, J. & FAN, Y. (2008). High dimensional classification using feature annealed independence rules. *The Annals of Statistics* **36**, 2605–2637.

FRALEY, C. & RAFTERY, A. E. (2002). Model-based clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Society* **97**, 611–631.

GHOSH, A. K. & CHAUDHURI, P. (2005). On Data Depth and Distribution-Free Discriminant Analysis Using Separating Surfaces. *Bernoulli* **11**, 1–27.

HALL, P., TITTERINGTON, D. M. & XUE, J. H. (2009). Median-based classifiers for high-dimensional data. *Journal of the American Statistical Society* **104**, 1597–1608.

HAND, D. J. & YU, K. (2001). Idiot's Bayes — Not so Stupid After All? *International Statistical Review* **69**, 385–398.

HASTIE, T. J. & TIBISHIRANI, R. J. (1996). Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society, Series B* **58**, 155–176.

HINKLEY, D. V. (1975). On power transformations to symmetry. *Biometrika* **62**, 101–111.

HYVÄRINEN, A., KARHUNEN, J. & OJA, E. (2001). *Independent Component Analysis*. New York: Wiley.

JOE, H. (2006). Generating Random Correlation Matrices Based on Partial Correlations. *Journal of Multivariate Analysis* **97**, 21772189.

JOHNSON, N. L., KOTZ, S. & BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions*. New York: Wiley.

JÖRNSTEN, R. (2004). Clustering and Classification based on the L_1 Data Depth. *Journal of Multivariate Analysis* **91**, 67–89.

MIKA, S., RÄTSCH, G., WESTON, J., SCHÖLKOPF, B. & MÜLLER, K. (1999). Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX*, Y. H. Hu, J. Larsen, E. Wilson & S. Douglas, eds. IEEE.

PARK, M. Y. & HASTIE, T. J. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50.

SARANTARIDIS, D., HENNIG, C. & CARUANA, D. J. (2012). Bioaerosol detection using potentiometric tomography in flames. *Chemical Science* **3**, 2210–2216.

TIBSHIRANI, R. J., HASTIE, T. J., NARASIMHAN, B. & CHU, G. (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.

TIBSHIRANI, R. J., HASTIE, T. J., NARASIMHAN, B. & CHU, G. (2003). Class Prediction by Nearest Shrunken Centroids, with application to DNA Microarray. *Statistical Science* **18**, 104–117.

WANG, L., ZHU, J. & ZOU, H. (2008). Hybrid Huberized Support Vector Machines for Microarray Classification and Gene Selection. *Bioinformatics* **24**, 412–419.

WANG, S. & ZHU, J. (2007). Improved Centroids Estimation for the Nearest Shrunken Centroid Classifier. *Bioinformatics* **23**, 972–979.