Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–10

DOI: 10.1177/0962280217699996
journals.sagepub.com/home/smm
SAGE

Optimal quantile level selection for disease classification and biomarker discovery with application to electrocardiogram data

Yingchun Zhou,[1] Rong Huang,[1] Shanshan Yu[1] and Yanyuan Ma[2]

Abstract

Classification with a large number of predictors and biomarker discovery become increasingly important in biological and medical research. This paper focuses on performing classification of cardiovascular diseases based on electrocardiogram analysis which deals with many variables and a lot of measurements within variables. We propose an optimal quantile level selection procedure to reduce dimension by characterizing distributions with quantiles and combine with classification tools to produce sensible classification and biomarker discovery results. Simulation and an intensive study of a real data set are performed to illustrate the performance of the proposed method.

Keywords

Optimal quantile level, disease classification, biomarker identification, electrocardiogram analysis, quantile treatment difference

1 Introduction

With fast advancement of technology in medical devices, "big data" in medical area has been rapidly growing in the recent years. Various data related to a person's health are collected. For example, a wearable watch may collect a person's heart rate, blood pressure and walking speed over a long period of time; a DNA test may obtain information about a person's genes with relatively low cost. With such big data, much research focuses on making predictions and finding useful predictors for certain purposes. Prediction with a categorical response is classification, and useful predictors can be biomarkers (for diseases). Therefore, classification with a large number of predictors and biomarker discovery have become increasingly important in modern biological and medical research.

A lot of statistical methods and machine learning tools have been developed for classification with a large number of predictors, such as support vector machine (SVM),[1] neural network,[2] classification trees,[3–5] other deep learning methods, etc. These methods have sophisticated software packages and have been widely used in many areas. Although they perform reasonably well with a large number of predictors, the performance tend to get worse when the number of predictors becomes larger than or far exceeding the number of sample size. For the data that a portable device generates, there are usually many measurements recorded for each variable. In dealing with such data, it is not wise to put all the measurements of all the variables into the classifiers directly, since (1) there will be tens of thousands of input variables which add to the difficulty of producing accurate results and (2) the data have certain structures that can be utilized to reduce dimension. A natural way would be to find some representative measures for each variable and then put them into the classifiers.

The most commonly used summary measures are mean, median, wavelet coefficients, etc.[6] However, the means or medians only represent the central locations of the distributions of the variables, which may not be very

[1]Department of Statistics and Actuarial Sciences, East China Normal University, Shanghai, P.R. China
[2]Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA

Corresponding author:
Yingchun Zhou, Department of Statistics and Actuarial Sciences, East China Normal University, 500 Dongchuan Road, Shanghai 200241, P.R. China.
Email: yczhou@stat.ecnu.edu.cn

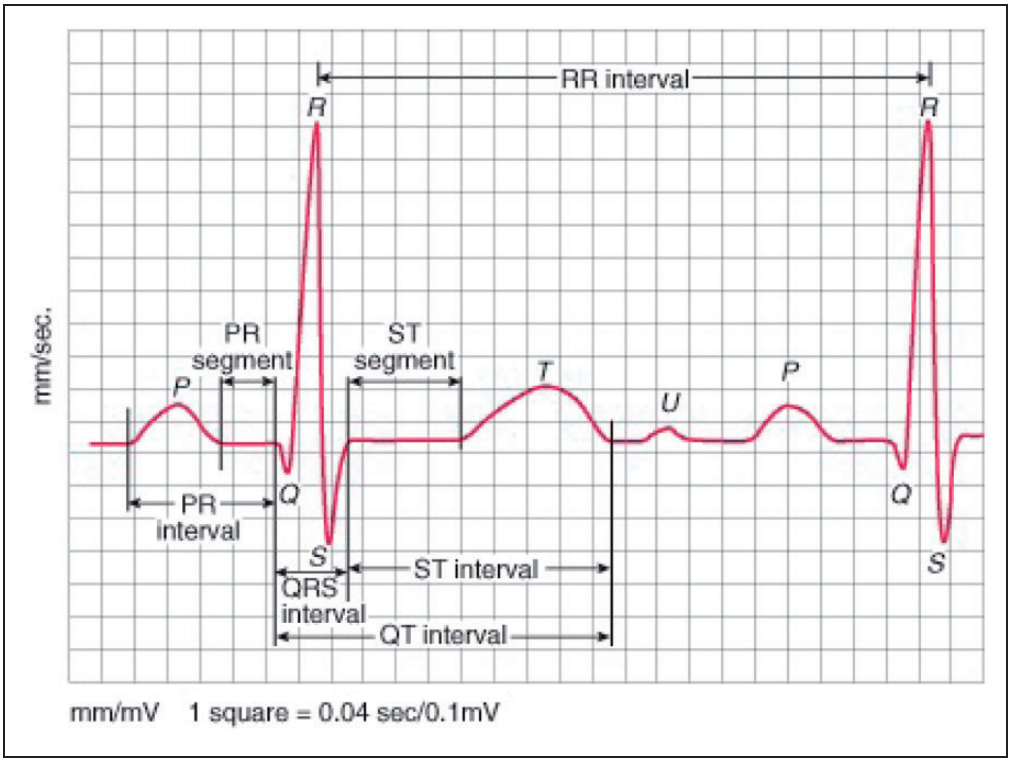

Figure 1. Various local wave forms and intervals/segments on ECG.

powerful in distinguishing two distributions when the distributions differ in other ways, as shown in the real data example. The wavelet coefficients may not be interpreted well from the physical background. In the present paper, we aim to find interpretable (and easy to understand for nonstatisticians) summary measures and focus on the situation when each variable has a large number of measurements, such as heartbeat interval measurements on electrocardiograms (ECGs). Our research is specifically focused on cardiovascular disease classification and biomarker identification problem based on ECG data.

ECGs are comparatively low cost and noninvasive in screening and diagnosing heart diseases. With the development of personal ECG monitors, large amounts of ECGs are recorded and stored; therefore, fast and efficient algorithms are called for to analyze the data and make diagnosis. Interpretable features that characterize ECGs include heartbeat interval features, amplitude features, slope features, etc. See Figure 1 for various waveforms and intervals on ECG. For each variable such as the QT interval (the time between the onset of the Q-wave and the offset of the T-wave on ECG), a large number of measurements are available since each subject may have thousands of beats recorded and each beat has one QT interval measurement. Therefore, summary statistics need to be chosen to characterize the distribution of the QT intervals. The most frequently used statistic in ECG analysis is the sample mean. However, it is observed that the sampling distributions of various measurement variables of the diseased subjects are often skewed, heavy-tailed or multi-modal, as compared to the symmetric, light-tailed and unimodal distributions for those of the healthy subjects, as shown in Figure 2. Indeed, the distributions of the PR and QT intervals for the diseased subjects have heavier tails than those of the healthy subjects; the distributions of the slopes of up-T and down-T waveforms are mixed for the diseased subjects and not mixed for the healthy subjects. It seems that the diseased subjects tend to have a small portion of abnormal beats, leading to heavy-tailed or mixture distributions (MDs) for some measurement variables. Therefore, sample quantiles characterizing tail behaviors of distributions seem to be better summary statistics to distinguish such distributions from normal distributions. To maximize the reduction of dimension, we propose to select an optimal quantile level to best distinguish distributions. This optimal quantile level together with variable selection procedures will help doctors to identify important measurement variables on ECGs and to interpret the cardiovascular behavior of patients with various diseases.

There has been much research in comparing two distributions in terms of quantiles. Doksum[8] introduced quantile treatment difference, which is to measure the treatment effect in terms of quantiles. Hollander and

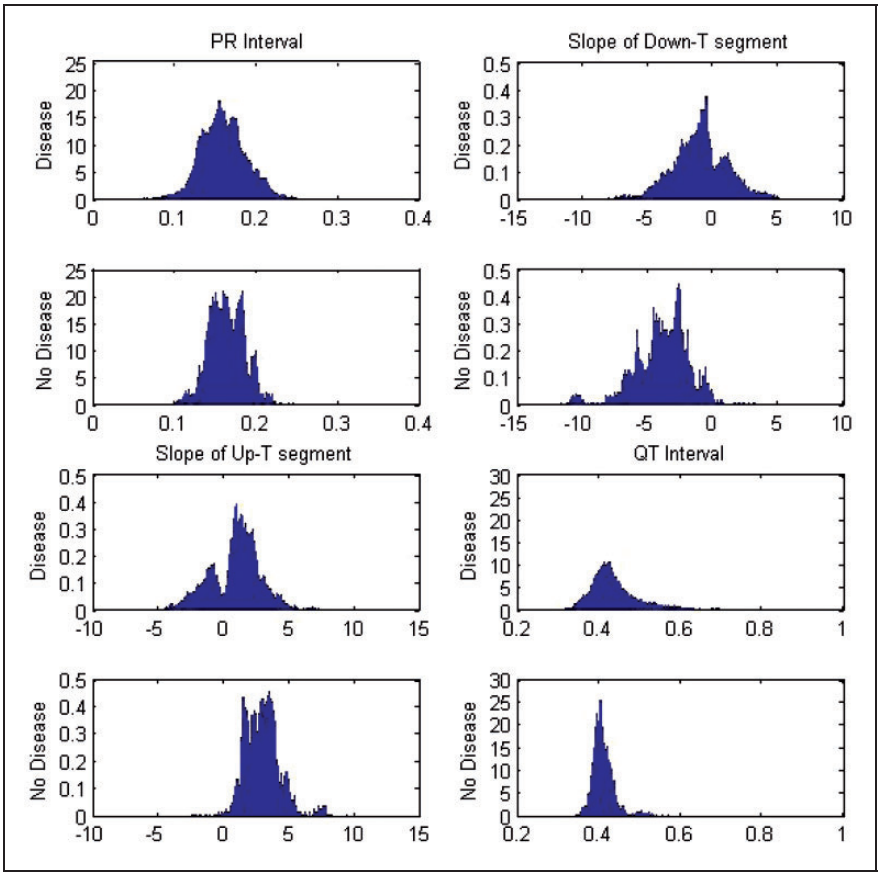

Figure 2. Histograms of the PR interval, the QT interval, the slope of the up-T waveform and the slope of down-T waveform of both healthy and diseased subjects.

Korwar[9] and Henry et al.[10] extensively studied the two-sample horizontal quantile shift function; Li et al.[11] introduced and discussed the two-sample vertical quantile comparison function, both were mainly focused on testing equality of two distributions based on quantile measures of censored or truncated data. Gilchrist[12] discussed the statistical applications of quantile functions. Parzen[14] introduced further quantile-based measures to quantify the distance between two distributions. The quantile-based classification procedure is usually related to the receiver operating characteristic (ROC) curve, which is to assess the performance of a diagnostic test associated with a marker that yields continuous measurements.[15] However, these procedures mainly focused on classification with one continuously measured variable and identification of the best quantile to make diagnostics. In our problem, there are a number of continuously measured variables. One cannot use ROC or related methods to simultaneously choose the best quantiles for all the variables. In addition, to combine with the aforementioned machine learning classification tools, we need not only a measure to describe the distance between two distributions for each variable but also an easy-to-understand statistic to be conveniently put into the classifiers to produce sensible classification and biomarker identification results.

The proposed optimal quantile level is easily understood by medical doctors since the sample quantile with a certain level can be interpreted in the same way as the median (sample quantile with a 50% level). In addition, in diagnosing cardiovascular diseases, usually it is not the exact time that abnormal heart beats occur but rather the frequency of the abnormal heart beats during a period that helps to diagnose a heart disease. This frequency can be captured by the densities of certain variables on ECG. To select the optimal quantile level to achieve best classification between healthy and diseased subjects is similar to finding the frequency of abnormal beats. Therefore, we think this optimal quantile level is a good biomarker for cardiovascular diseases.

The organization of the paper is as follows. In Section 2, the optimal quantile level selection method is introduced. Section 3 contains a simulation study to compare the performance of the proposed method under different situations to the mean method. In Section 4, we perform a thorough real data analysis to demonstrate the powerfulness of the method. Discussion on several issues of this approach is given in Section 5.

2 The optimal quantile level selection procedure

The optimal quantile level is chosen to maximize the standardized difference between the sample quantiles of two distributions. Suppose we have observations

$$X_i,\ i = 1, \ldots, N \equiv \sum_i l_{1i}$$
$$Y_j,\ j = 1, \ldots, M \equiv \sum_j l_{2j}$$

where X_i is a random variable having $N = \sum_{i=1}^{n} l_{1i}$ iid observations following distribution F_X, n is the number of subjects of the first group and l_{1i} is the number of measurements of the ith subject in this group. Y_j is a random variable having $M = \sum_{j=1}^{m} l_{2j}$ iid observations following distribution F_Y, m is the number of subjects of the second group and l_{2j} is the number of measurements of the jth subject in this group. The optimal quantile level $\hat{\tau}$ is defined to maximize the standardized difference of the sample quantiles

$$\hat{\tau} = \text{argmax}_{\mathbf{0} \le \tau \le 1} |\hat{\theta}_{X\tau} - \hat{\theta}_{Y\tau}| / \widehat{se}(\hat{\theta}_{X\tau} - \hat{\theta}_{Y\tau}) \tag{1}$$

where $\hat{\theta}_{X_\tau} = \hat{F}_X^{-1}(\tau)$ and $\hat{\theta}_{Y_\tau} = \hat{F}_Y^{-1}(\tau)$ represent the sample quantiles at level τ of X and Y, respectively

$$\widehat{se}(\hat{\theta}_{X\tau} - \hat{\theta}_{Y\tau}) = \sqrt{\tau(1-\tau)} \sqrt{\frac{1}{N\hat{f}_X(\hat{\theta}_{X\tau})^2} + \frac{1}{M\hat{f}_Y(\hat{\theta}_{Y\tau})^2}}$$

represents the estimate of the standard deviation of $\hat{\theta}_{X\tau} - \hat{\theta}_{Y\tau}$ obtained from the asymptotic variance of $\hat{\theta}_{X\tau} - \hat{\theta}_{Y\tau}$, in which $\hat{f}_X(\cdot), \hat{f}_Y(\cdot)$ are kernel estimates of the probability density functions $f_X(\cdot), f_Y(\cdot)$.

Theorem 1. Let $\hat{\tau}$ be defined in equation (1), and let τ maximize the standardized absolute quantile difference $|\theta_{X_\tau} - \theta_{Y_\tau}| / \text{se}(\theta_{X_\tau} - \theta_{Y_\tau})$, i.e.

$$\tau = \text{argmax}_{0 \le \tau \le 1} |\theta_{X\tau} - \theta_{Y\tau}| / \text{se}(\theta_{X\tau} - \theta_{Y\tau})$$

Then, $\hat{\tau} - \tau$ has the leading term $A(\tau)^{-1} T$ as N and M approach infinity.
Here

$$\begin{aligned} A(\tau) = & \frac{f_X^3(\theta_{X\tau}) f_Y'(\theta_{Y\tau}) - f_Y^3(\theta_{Y\tau}) f_X'(\theta_{X\tau})}{f_X^3(\theta_{X\tau}) f_Y^3(\theta_{Y\tau})(\theta_{X\tau} - \theta_{Y\tau})} - \frac{\{f_X(\theta_{X\tau}) - f_Y(\theta_{Y\tau})\}^2}{f_X^2(\theta_{X\tau}) f_Y^2(\theta_{Y\tau})(\theta_{X\tau} - \theta_{Y\tau})^2} + \frac{1/2 - \tau + \tau^2}{(\tau - \tau^2)^2} \\ & + \frac{M^{-1} f_X^5(\theta_{X\tau}) f_Y''(\theta_{Y\tau}) + N^{-1} f_Y^5(\theta_{Y\tau}) f_X''(\theta_{X\tau})}{f_X^3(\theta_{X\tau}) f_Y^3(\theta_{Y\tau})\{M^{-1} f_X^2(\theta_{X\tau}) + N^{-1} f_Y^2(\theta_{Y\tau})\}} - \frac{2\{M^{-1} f_X^4(\theta_{X\tau}) f_Y'(\theta_{Y\tau}) - N^{-1} f_Y^4(\theta_{Y\tau}) f_X'(\theta_{X\tau})\}^2}{f_X^4(\theta_{X\tau}) f_Y^4(\theta_{Y\tau})\{M^{-1} f_X^2(\theta_{X\tau}) + N^{-1} f_Y^2(\theta_{Y\tau})\}^2} \\ & - 4(NM)^{-1} \frac{f_Y^4(\theta_{Y\tau})\{f_X'(\theta_{X\tau})\}^2 + f_X^4(\theta_{X\tau})\{f_Y'(\theta_{Y\tau})\}^2}{f_X^2(\theta_{X\tau}) f_Y^2(\theta_{Y\tau})\{M^{-1} f_X^2(\theta_{X\tau}) + N^{-1} f_Y^2(\theta_{Y\tau})\}^2} \end{aligned}$$

and

$$\begin{aligned} T = & \frac{1}{f_X(\theta_{X\tau}) f_Y(\theta_{Y\tau})\{f_X^2(\theta_{X\tau}) + f_Y^2(\theta_{Y\tau})\}} \times \left[\frac{f_Y^3(\theta_{Y\tau})(Nh^2)^{-1} \sum_{i=1}^{N} K'\{(\theta_{X\tau} - X_i)/h\}}{f_X(\theta_{X\tau})} \right. \\ & \left. - \frac{f_Y^3(\theta_{Y\tau}) f_X'(\theta_{X\tau})}{f_X(\theta_{X\tau})} + \frac{f_X^3(\theta_{X\tau})(Mh^2)^{-1} \sum_{j=1}^{M} K'\{(\theta_{Y\tau} - Y_j)/h\}}{f_Y(\theta_{Y\tau})} - \frac{f_X^3(\theta_{X\tau}) f_Y'(\theta_{Y\tau})}{f_Y(\theta_{Y\tau})} \right] \end{aligned}$$

where f' and f'' are the first and second derivatives of the density f, K is the kernel function and h is the bandwidth of the kernel estimator. Assume N/M is bounded between two constants, $T = O\{h^2 + (Nh^3)^{-1/2}\}$.

The results in Theorem 1 essentially tell us that the proposed procedure is estimating the target quantile level τ, which is the quantile level at which the standardized quantile difference of the two distributions is maximized. As an estimator, $\hat{\tau}$ is consistent, and $\hat{\tau}$ approaches τ at the rate of $h^2 + (Nh^3)^{-1/2}$. This is the order of a nonparametric estimation of the derivative of a density estimation, which agrees with our intuition due to the inclusion of the

estimated standard error term. When we set $h = N^{-1/7}$, the difference between $\hat{\tau}$ and τ will go to zero at the best rate of $N^{-2/7}$ when the sample size approaches infinity. The proof of Theorem 1 is in Appendix A of the Supplementary Materials.

Remark 1. To facilitate the proof of Theorem 1, we assume that the observations are independent and identically distributed. This condition can be relaxed to the assumption that the observations are weakly correlated.

In the quantile regression setting for comparison of treatment effects

$$\hat{\delta}(\tau) = \hat{F}_Y^{-1}(\tau) - \hat{F}_X^{-1}(\tau)$$

is called quantile treatment effect estimate.[8] In fact, the optimal quantile level defined here is to maximize the standardized quantile treatment effect. The reason for the standardization is that without standardization, the maximum tends to be achieved at the most extreme quantile levels and the estimator has a large variance. Hence, a balance between achieving the largest difference and the largest estimation accuracy needs to be considered.

Note that the optimal quantile proposed here is for two-category classification, more general procedures for multi-category classification will be presented in future work.

3 Simulation study

The proposed approach consists of two steps. The first step is to reduce dimension by characterizing the distributions of the measurements with ''optimal quantile levels'' that have good classification properties based on distance, and the second step is to combine with machine learning tools to further reduce dimension and perform classification. The simulation here focuses on the first step since this is the novel method proposed in the paper.

To simulate the situations that are observed in real data as shown in Figure 2, we consider classification between a standard normal distribution and a heavy-tailed distribution or a MD. For visual comparison of the distributions, Supplementary Figure 1 shows the histograms of 10,000 random numbers generated from a standard normal distribution and two heavy-tailed distributions: t distributions with degrees of freedom 10 and 3 ($t(10)$and $t(3)$), respectively, and Supplementary Figure 2 shows the histograms of 10,000 random numbers generated from a standard normal distribution and five MDs. The five MDs are MD1: $0.8N(0,1) + 0.2N(2,1)$, MD2: $0.8N(0,1) + 0.2N(4,1)$, MD3: $0.6N(0,1) + 0.4N(2,1)$, MD4: $0.8N(0,1) + 0.2N(2,4)$ and MD5: $0.8N(0,1) + 0.2(t(3) + 2)$. MD1 is a mixture of two normal distributions, with the main component $N(0, 1)$ representing the normal beats, and a minor component $N(2, 1)$ representing the abnormal beats for a subject with disease. Based on MD1, we increased the mean, the weight and the variance of the minor component in MD2, MD3 and MD4, respectively. In MD5, we replaced the minor normal distribution with a more heavy-tailed t-distribution. These changes make the minor component in the MD manifest more clearly in the histogram. Results of classification between a standard normal and all the five MDs are compared.

To illustrate the procedure, the objective functions computed from 10,000 randomly generated numbers are shown in Figure 3, where the left graph shows the case of distinguishing $N(0, 1)$ from $t(3)$, and the right graph shows the case of distinguishing $N(0, 1)$ from MD1. The left graph is close to symmetric since both $N(0, 1)$ and $t(3)$ are symmetric about 0. There is a peak on each side of the objective function. The right graph is skewed since the main difference between $N(0, 1)$ and MD1 is on the right side due to the minor component shifting some mass to the right of the main component in MD1. Naturally, the peak is therefore located on the right side at which the optimal quantile level is selected.

To compare the performance of the optimal quantile level method (referred to as ''quantile method'' in the sequel) with the mean method (classification using the means of the variables) under various scenarios, Table 1 shows the sensitivity and specificity of classification using these two methods from 5000 runs. Here, $n_1 = N = M$ is the sample size of the training data belonging to one category; hence, $2n_1$ is the sample size of the training data used to compute the optimal quantile level $\hat{\tau}$, and n_2 is the number of measurements within one observation/subject in the testing data set. Here, we fix $n_1/n_2 = 10$. This mimics the data structure when there are 21 subjects, each has the same number of measurements n_2, 20 of the 21 subjects are used for training (10 subjects for each category) and 1 subject is used for testing. Thus, $n_1 = 10n_2$.

After $\hat{\tau}$ is obtained from the training data, sample quantile at level $\hat{\tau}$ for the testing data is compared with the sample quantiles at level $\hat{\tau}$ of the two empirical distributions generated from the training data. The empirical distribution that has its sample quantile closer to the testing sample quantile is chosen to be the class where the

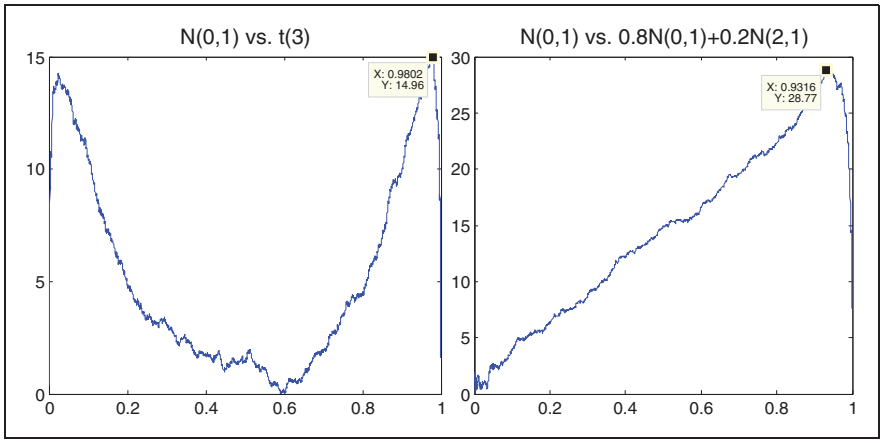

Figure 3. The objective functions of selecting optimal quantile levels comparing (0; 1) with t(3) (left graph) and comparing $N(0; 1)$ with MD1 (right graph).

testing data belong to. From Table 1, it is clear that the quantile method performs much better than the mean method in comparing normal with heavy-tailed distributions. As the sample size increases, the quantile method performs better while the mean method stays about the same. This is because the sample mean cannot differentiate two distributions with the same mean. In the MD cases, both methods perform well, while the quantile method generally outperforms the mean method.

To reflect the sample size of the real data set analyzed in Section 4, Supplementary Table 1 shows the classification results when $n_1 = 70,000$ and $n_2 = 300$. The quantile method outperforms the mean method in all cases.

To study further how the optimal quantile level $\hat{\tau}$ changes with the underlying distributions, $\hat{\tau}$ is obtained with one distribution being a standard normal, and the other distribution varies in different ways. Below are observations for three typical cases.

Case 1: When the tail of the other distribution becomes heavier, $\hat{\tau}$ becomes closer to 0.5. To illustrate this, we used t-distributions with decreasing degrees of freedom to represent distributions with heavier tails. Note that the objective function to distinguish a normal distribution from a t-distribution is symmetric and centered at 0.5, so the optimal quantiles have equal chance to be on either side of 0.5. To make the comparison easier, we always select the optimal quantiles larger than 0.5 and show the mean and standard deviation of $\hat{\tau}$ from 5000 runs, as shown in the top part of Table 2. The sample size for estimating τ is set to be $n_1 = 100,000$ for more accurate estimation. To provide a more complete picture of the heavy-tail effect, results for both heavy-tail distributions and mixture of normal and heavy-tailed distributions are listed, and they show similar trends.

Case 2: When the mean and standard deviation of the minor distribution in the MD become bigger, i.e., the separation of the two component distributions becomes more obvious, $\hat{\tau}$ becomes closer to 1, as shown in the middle part of Table 2. Also note that Std ($\hat{\tau}$) becomes smaller since it is easier to distinguish the two distributions.

Case 3: When the weight of the minor distribution in the MD becomes bigger, $\hat{\tau}$ becomes closer to 0.5, as shown in the bottom part of Table 2. Std ($\hat{\tau}$) becomes bigger as it is harder to distinguish the standard normal with the MDs. Also note that $\hat{\tau}$ is smaller for $0.8N(0, 1) + 0.2(t(2) + 2)$ than that for $0.8N(0, 1) + 0.2N(2, 1)$, since the tail of the minor component is heavier, which agrees with the observations in Case 1.

4 Real data analysis

We now analyze the ECG data from the Department of Cardiology of University Clinic Benjamin Franklin in Berlin, Germany. The ECGs were collected from healthy volunteers and patients with different heart diseases. The data set contains ECG records of 290 subjects, each subject has several 1- to 2-min long records of standard 12-lead ECGs, accompanied with his/her gender, age and clinical diagnosis results. Among the 290 subjects, 44 subjects have missing information in their records, so 246 subjects with 498 ECG records are used in classification. The data contain subjects belonging to five health status categories: healthy (52 subjects), myocardial infarction (149 subjects), cardiomyopathy (17 subjects), atrioventricular bundle branch block

Table 1. Sensitivity (above) and specificity (below) for both the mean and the quantile methods.

		$n_1 = 1000$ $n_2 = 100$	$n_1 = 10{,}000$ $n_2 = 1000$	$n_1 = 100{,}000$ $n2 = 10{,}000$
	Quantile	57.10%	89.34%	99.98%
$N(0, 1)$ vs. t(10)		74.40%	96.28%	100%
	Mean	43.76%	48.72%	45.10%
		54.54%	52.16%	55.92%
	Quantile	70.18%	99.72%	100%
$N(0; 1)$ vs. t(3)		98.94%	100%	100%
	Mean	46.72%	48.98%	48.48%
		53.60%	51.88%	53.64%
	Quantile	95.78%	100%	100%
$N(0, 1)$ vs. MD1:		99.62%	100%	100%
$0.8N(0, 1) + 0.2N(2, 1)$	Mean	98.98%	100%	100%
		95.48%	100%	100%
	Quantile	100%	100%	100%
$N(0, 1)$ vs. MD2:		100%	100%	100%
$0.8N(0, 1) + 0.2N(4, 1)$	Mean	100%	100%	100%
		99.98%	100%	100%
	Quantile	99.98%	100%	100%
$N(0, 1)$ vs. MD3:		100%	100%	100%
$0.6N(0, 1) + 0.4N(2, 1)$	Mean	100%	100%	100%
		99.98%	100%	100%
	Quantile	88.92%	100%	100%
$N(0, 1)$ vs. MD4:		100%	100%	100%
$0.8N(0, 1) + 0.2N(2, 4)$	Mean	89.22%	99.98%	100%
		94.38%	100%	100%
	Quantile	98.62%	100%	100%
$N(0, 1)$ vs. MD5:		98.54%	100%	100%
$0.8N(0, 1) + 0.2(t(3) + 2)$	Mean	96.98%	100%	100%
		96.38%	100%	100%

(14 subjects) and rhythm disorders (14 subjects). Since the sample sizes of the last three diseases are too small, we combine the disease groups together to form a ''Disease'' category. The sampling frequency of the data is 1000 Hz.

Single-lead data (MLII) are used for classification, with the notion that the method can be applied to 12-lead data as well. The ECGPUWAVE function in the WFDB package, available at http://www.physionet.org/physiotools/ecgpuwave/, is applied to mark the start, peak and end points of the P-wave, the QRS complex and the T-wave. This function also provides the T-wave type of each heartbeat which is one of the features used in classification.

Measurements of various variables on ECGs are obtained based on the annotations by the ECGPUWAVE function. Three general types of features are considered: time span measurement variables, amplitude measurement variables and the slopes of waveforms. Below are detailed descriptions of these types of features:

- *Time span measurements*
  Six commonly used time span measurements are considered: the lengths of the RR interval, PR interval, QT interval, P-wave, QRS complex and T-wave.
- *Amplitude measurements*
  The amplitudes of P-wave, QRS complex and T-wave are considered. To measure the P-wave amplitude, we first estimate the baseline by taking the mean of the values in the PR segment, ST segment and TP segment (from the end of the T-wave to the start of the P-wave of the next heartbeat), then subtract the maximum and minimum values of the P-wave by the estimated baseline, and take the one with a bigger absolute value as the amplitude of P-wave. Other amplitude measurements are obtained similarly.

Table 2. Mean and standard deviation of the optimal quantile level for distinguishing a N(0, 1) from various heavy-tailed distributions (top); distinguishing a N(0, 1) from mixed distributions with changed means and standard deviations (middle); distinguishing a N(0, 1) from mixed distributions with changed weights (bottom).

(Top) $N(0; 1)$ vs.	Change of heavy-tail			Mixed + change of heavy-tail		
	t(10)	t(4)	t(2)	$0.8N(0, 1)+$ $0.2(t(10)+2)$	$0.8N(0, 1)+$ $0.2(t(4)+2)$	$0.8N(0, 1)+$ $0.2(t(2)+2)$
$E(\tau)$	0.9894	0.9797	0.9574	0.9405	0.9328	0.9209
$\text{Std}(\tau)$	0.0033	0.0026	0.0028	0.0054	0.0051	0.0044
(Middle) $N(0, 1)$ vs.	Change of mean			Change of standard deviation		
	$0.8N(0, 1)$ $+0.2N(1, 1)$	$0.8N(0, 1)$ $+0.2N(2, 1)$	$0.8N(0, 1)$ $+0.2N(4, 1)$	$0.8N(0, 1)$ $+0.2N(2, 1)$	$0.8N(0, 1)$ $+0.2N(2, 2)$	$0.8N(0, 1)$ $+0.2N(2, 4)$
$E(\tau)$	0.8399	0.9445	0.9463	0.9445	0.9826	0.9873
$\text{Std}(\tau)$	0.0330	0.0054	0.0024	0.0054	0.0016	0.0015
(Bottom) $N(0, 1)$ vs.	Change of weight			Heavy-tail + change of weight		
	$(0.8N(0, 1)$ $+0.2N(2, 1)$	$0.7N(0, 1)$ $+0.3N(2, 1)$	$0.6N(0, 1)$ $+0.4N(2, 1)$	$0.8N(0, 1)+$ $0.2(t(2)+2)$	$0.7N(0, 1)+$ $0.3(t(2)+2))$	$0.6N(0, 1)+$ $0.4(t(2)+2))$
$E(\tau)$	0.9444	0.9121	0.8747	0.9207	0.8835	0.8433
$\text{Std}(\tau)$	0.0054	0.0063	0.0074	0.0044	0.0047	0.0050

$n_1 = 100{,}000$.

- *The slopes of waveforms*
  The slopes of waveforms are also considered to measure the dynamic features of a heartbeat. Each heartbeat is split into nine segments, and the slope of the waveform in each segment is estimated by simple linear regression. Supplementary Table 2 lists the nine segments with definitions.

The optimal quantile levels $\hat{\tau}$'s are then selected for all the 19 measurement variables, and the sample quantiles at level $\hat{\tau}$'s are used as input variables of the classifiers. Although in this case, the number of variables is not that large compared to the number of subjects, yet there are several hundred measurements within each variable for each subject. If one considers all the measurements as input variables for classification, the number of input variables would be huge. Therefore, the main purpose of using this real data example is to illustrate the use of the optimal quantile level method. One may find other applications when both the number of variables and the number of measurements within variables are big.

To further reduce the number of input variables and account for correlations that might exist among predicting variables, we also consider using principal component analysis (PCA), where the largest principal components that contribute to a total of 90% of the variability are extracted. Therefore, the classification performances of four sets of input variables are compared: (a) the means of the variables; (b) the principal components of the means of the variables; (c) the quantiles of the variables and (d) the principal components of the quantiles of the variables. In addition, three frequently used classifiers are compared: stepwise discriminant analysis (SDA),[16] SVM[1] and LASSO logistic regression (LLR).[17] All three methods are popular classification methods and are based on different principles and procedures.

Table 3 shows the sensitivity, specificity and accuracy of classification with leave-one-out cross-validation. Note that quantile-based methods are generally better than mean-based methods in terms of accuracy, which takes into account both sensitivity and specificity. Comparing PCA to non-PCA procedures, the PCA procedure is generally helpful in improving classification performance. Comparing across the three classifiers, we find that both LLR and SVM have high sensitivities but low specificities, while SDA has higher specificity with relatively low sensitivity. Overall SDA has the best balance between sensitivity and specificity. Therefore, considering good balance between sensitivity and specificity, the best performance is achieved by "Quantile + PCA + SDA." Considering only the accuracy, "Quantile + PCA + SVM" has the best performance.

Through the PCA procedure and variable selection procedure in the classifiers, one is able to identify important variables as well as their optimal quantile levels as biomarkers for disease classification: the slope of the up-T-wave

Table 3. Classification results of the different methods using leave-one-out cross-validation.

Classification Method	Feature selection Method	Sensitivity	Specificity	Accuracy
SDA	Mean	88.05%	80.77%	84.14%
	Quantile	81.44%	90.38%	82.93%
	Mean+PCA	72.16%	90.38%	76.02%
	Quantile+PCA	**80.93%**	**96.15%**	**84.15%**
LLR	Mean	94.33%	48.08%	84.55%
	Quantile	90.72%	75%	87.40%
	Mean+PCA	94.33%	46.15%	84.15%
	Quantile+PCA	90.72%	55.77%	83.33%
SVM	Mean	92.78%	38.46%	81.30%
	Quantile	91.24%	50.00%	82.52%
	Mean+PCA	94.33%	48.08%	84.55%
	Quantile+PCA	92.78%	69.23%	**87.80%**

SDA: stepwise discriminant analysis; SVM: support vector machine; LLR: LASSO logistic regression; PCA: principal component analysis.
The bold values are computed based on leave-one-out cross-validation of the real data, they are percentages.

(14% quantile), the slope of the down-T-wave (78% quantile), the slope of the up-R-wave (18% quantile), the length of the QRS interval (73% quantile) and the amplitude of the T-wave (4% quantile). These are most frequently selected and have highest weights in the classification procedure.

5 Discussion

The optimal quantile level selection is a good dimension reduction method when each variable has a large number of measurements. In the real data analysis, this efficient dimension reduction method allows us to include all commonly used measurement variables on ECGs without pre-selection. This is useful when no prior information is available regarding which variables are more important.

Three frequently used classifiers are applied using the optimal quantile levels of the measurement variables, and their results are compared. It is found that SDA on dimension-reduced features by PCA is the most stable and effective procedure. LASSO type of methods are well known for variable selection and classification when the number of independent variables is large (even larger than the sample size) and only a few of them are related to the response variable (sparsity).[18] In this particular application, the number of independent variables is large but still can be handled well by other methods. In addition, model sparsity may not be satisfied, because many variables may be related to the response and they are correlated. In this case, PCA and stepwise procedures are more appropriate dimension reduction methods.

It is likely that the measurements of different beats on ECGs are correlated, especially for diseased subjects. However, the correlation may vary from subject to subject and one may need to introduce a complex model for such correlation. In the present paper, we allow the observations to be weakly correlated without looking deeper into the structure of the correlation. Further research on correlated observations will be addressed in future work.

Due to small sample sizes in disease categories such as Bundle branch block, cardiomyopathy and dysrhythmia, only two-category classification is performed in the paper. However, the proposed method can be extended to multiple-disease classification when more data are available. Also note that with bigger sample sizes, multi-lead analysis is preferred to single-lead analysis, since different diseases may show abnormality in different leads. For example, Shen et al.[19] and Zimmerman and Syeda-Mahmood[20] performed multi-lead classification of ECG waveforms. The proposed quantile-based measures can be combined with those procedures to achieve better classification results.

Acknowledgments

The authors would like to thank Professor Michael Oeff at the Department of Cardiology of University Clinic Benjamin Franklin for providing the ECG data set and doctors at Shanghai Ruijin Hospital for helpful discussion about clinically meaningful variables on ECGs.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work was supported by Natural Science Foundation of Shanghai (project number: 14ZR1412900), the 111 Project (B14019), Program of Shanghai Subject Chief Scientist (14XD1401600) and an NSF grant DMS-1608540.

References

1. Chang CC and Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011; **2**: 1–27.
2. Bishop CM. *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press, 1995.
3. Loh WY. Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov* 2011; **1**: 14–23.
4. Fayn J. A classification tree approach for cardiacischemia detection using spatiotemporal information from three standard ECG leads. *IEEE Trans Biomed Eng* 2011; **58**: 95–102.
5. Mair J, Smidt J, Lechleitner P, et al. A decision tree for the early diagnosis of acute myocardial infarction in nontraumatic chest pain patients at hospital admission. *Chest* 1995; **108**: 1502–1509.
6. Ghongade R and Ghatol A. A robust and reliable ECG pattern classification using QRS morphological features and ANN. In: *TENCON 2008 – 2008 IEEE Region 10 Conference*, Hyderabad, India, 19–21 November 2008, pp. 1–6. IEEE.
7. Doksum K. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann Stat* 1974; **2**: 267–277.
8. Hollander M and Korwar RM. Nonparametric Bayesian estimation of the horizontal distance between two populations. In: Gnedenko BV, Puri ML and Vincze I (eds) *Nonparametric statistical inference I*. New York: North-Holland, 1982, pp.409–416.
9. Henry HS, Wells MT and Tiwari RC. Inference for shift functions in the two-sample problem with right-censored data, with applications. *J Am Stat Assoc* 1994; **89**: 1017–1026.
10. Li G, Tiwari RC and Wells MT. Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *J Am Stat Assoc* 1996; **91**: 689–698.
11. Gilchrist W. *Statistical modelling with quantile functions*. London, UK: Chapman and Hall/CRC, 2000.
12. Parzen E. Quantile probability and statistical data modelling. *Stat Sci* 2004; **19**: 652–662.
13. Wieand HS, Gail MH, Barray RJ, et al. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; **76**: 585–592.
14. Michael CC and Afifi AA. Comparison of stopping rules in forward stepwise discriminant analysis. *Journal of the American Statistical Association* 1979; **74**: 777–785.
15. Lee SL, et al. Efficient L1 regularized logistic regression. In: *Proceedings of the 21th national conference on artificial intelligence (AAAI-06)*, Boston, USA, 16–20 July 2006.
16. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B* 1996; **58**: 267–288.
17. Shen M, et al. Multi-lead ECG classification based on independent component analysis and support vector machine. In: *3rd International Conference on Biomedical Engineering and Informatics (BMEI)*, vol. 3, Yantai, China, 16–18 October 2010, pp. 960–964. IEEE.
18. Zimmerman TG and Syeda-Mahmood T. Automatic detection of heart disease from twelve channel electrocardiogram waveforms. *Comput Cardiol* 2007; **2007**: 809–812.