Statistics and Computing (2023) 33:55
https://doi.org/10.1007/s11222-023-10224-4

ORIGINAL PAPER

Quantile-distribution functions and their use for classification, with application to naïve Bayes classifiers

Edoardo Redivo[1] · Cinzia Viroli[1] · Alessio Farcomeni[2]

Received: 4 August 2022 / Accepted: 15 February 2023 / Published online: 24 March 2023

Abstract

We develop a flexible parametric framework for the estimation of quantile functions. This involves the specification of an analytical quantile-distribution function. It is shown to adapt well to a wide range of distributions under reasonable assumptions. We derive a least-square type estimator, leading to computationally efficient inference. By-products include a test for comparing two distributions, a variable selection method, and an innovative naïve Bayes classifier. Properties of the estimator, of the asymptotic test and of the classifier are investigated through theoretical results and simulation studies, and illustrated through a real data example.

Keywords Quantile estimation · Variable selection · Naïve Bayes

1 Introduction

Quantile functions, defined as the generalised inverse of cumulative distribution functions, have nice properties that make them a valuable inferential tool. For instance, sums and convex linear combinations of quantile functions are still quantile functions. As a consequence, it is possible to construct arbitrary new quantile functions that have great flexibility and a small number of parameters (see, for instance, Karvanen (2006)). Thus, we can obtain distributions with a wide range of different shapes and also the exact or approximate form of many common distributions, including the normal, Student's T and logistic distributions. See Gilchrist (2000) for a clear introduction to the use of quantile functions, their properties, and the main estimation methods.

Various flexible quantile functions have been proposed in the literature. The so-called *g-and-k* distribution (Haynes et al. 1997; Rayner and MacGillivray 2002) is defined as a generalization of the Gaussian distribution with additional skewness and kurtosis parameters. Freimer et al. (1988) introduced the quantile-based representation of the generalized Lambda distribution. Sankaran et al. (2016) proposed a new quantile function based on the sum of generalized Pareto and Weibull quantile functions.

Quantile functions that are linear in their parameters have desirable inferential properties, as will be shown in the following. Well-known examples are the flattened logistic distribution (Sharma and Chakrabarty 2019) and the generalized flattened logistic distribution (Chakrabarty and Sharma 2021).

Quantile functions can be estimated according to different strategies. Distributions that have analytical L-moments can be estimated by matching sample L-moments with their theoretical counterparts, in the same spirit as the method of moments (see, for instance, Chakrabarty and Sharma (2021)). Maximum likelihood estimation is possible as well; however, if the quantile function is not invertible - as is usually the case - then, for each observation of the data sample, say x, a numerical inversion needs to be carried out to find the correspondent percentile u, thus making the parameter estimation process numerically unstable and computationally expensive (Rayner and MacGillivray 2002). An alternative illustrated in Gilchrist (2000) is based on the minimization of the $L1$ norm between the ordered statistics and their theoretical median, leading to a least absolute deviation method. Without explicit

✉ Edoardo Redivo
edoardo.redivo@unibo.it

Cinzia Viroli
cinzia.viroli@unibo.it

Alessio Farcomeni
alessio.farcomeni@uniroma2.it

[1] Department of Statistical Sciences, University of Bologna, Bologna, Italy

[2] Department of Economics and Finance, University of Rome "Tor Vergata", Rome, Italy

density functions Bayesian estimation cannot be applied; however Allingham et al. (2009) and Drovandi and Pettitt (2011) developed an Approximate Bayesian Computation (ABC) strategy for the estimation of some classes of quantile functions.

In this work we show that the family of linear quantile functions can be efficiently estimated using least squares by exploiting the properties of the order statistics. We also develop the asymptotic distribution of a statistical test to check whether two estimated quantile functions have the same parameters. We also show how the procedure can be used for classification, by constructing a simple Naïve Bayes classifier based on quantile distributions, where the proposed testing procedure is used for variable selection and variable importance in a two-class problem. Empirical studies indicate that the proposed variable screening can help the classification task, and, in this perspective, it is alternative to variable weighting (see, for instance, Jiang et al. (2018) and Jiang et al. (2019)) or structure extensions by hidden variables Jiang et al. (2008). A completely different approach where quantile functions are used for classification is reported in Farcomeni et al. (2022).

The rest of the paper is organised as follows. In the next section we outline linear quantile functions and define our least squares estimator. Asymptotic results are given in Sect. 2.3, where we also derive the null distribution of relevant test statistics. In Sect. 3 we discuss how to use linear quantile functions for supervised classification and variable selection. Simulation studies are reported in Sect. 4 and the proposed strategy is illustrated on real data in Sect. 5. Some concluding remarks are given in Sect. 6.

2 Quantile-based distributions

Denote with $F(x;\boldsymbol{\theta})$ a distribution function that is right-continuous, depending on a vector of parameters $\boldsymbol{\theta}$ of length p. The quantile distribution function can be defined as in Parzen (1979):

$$F^{-1}(u;\boldsymbol{\theta}) = Q(u;\boldsymbol{\theta}) = \inf\{x : F(x;\boldsymbol{\theta}) \geq u\},$$

for $0 < u < 1$.

As in Tukey (1965), we call

$$q(u;\boldsymbol{\theta}) = Q'(u;\boldsymbol{\theta}),$$

the quantile density function, which is related to the density function as:

$$f(x;\boldsymbol{\theta}) = \frac{1}{q(F(x;\boldsymbol{\theta}))}. \tag{1}$$

For certain probability distributions the quantile function can be derived in analytical form through the inversion of the cumulative distribution function. Some examples are reported in Table 1. Most probabilistic densities do not admit closed-form quantile functions though. One notable example is the Gaussian distribution. The contrary is also true: a quantile function can be defined without making reference to an explicit probability distribution function.

An interesting family of quantile functions is given by the ones that are linear in their parameters. Starting from the symmetric quantile function of the logistic distribution:

$$Q(u;\boldsymbol{\theta}) = \alpha + \beta[\log u - \log(1-u)] \tag{2}$$

Sharma and Chakrabarty (2019) proposed the flattened version

$$Q(u;\boldsymbol{\theta}) = \alpha + \beta\left[\log\frac{u}{1-u} + \kappa u\right],$$

where the additional component indexed by the shape parameter κ regulates the flatness of the peak of the distribution. They derived classical and quantile-based properties of the distribution and compared its flexibility with respect to the logistic distribution in terms of fitting in empirical contexts.

More recently, Chakrabarty and Sharma (2021) proposed a generalization of the flattened logistic distribution (*fgld*):

$$Q(u;\boldsymbol{\theta}) = \alpha + \beta\left[(1-\delta)\log u - \delta\log(1-u) + \kappa u\right] \tag{3}$$

that proved to be very flexible and outperformed the existing strategies in terms of model fitting. Figures 1 and 2 show the range of shapes this distribution can take.

2.1 Least squares estimation

In order to estimate the quantile function $Q(u,\boldsymbol{\theta})$, different strategies can be applied. L-moments matching (Chakrabarty and Sharma 2021) requires the analytical form of L-moments for the quantile function, along the same lines of method of moments. Maximum likelihood is a possible alternative strategy but it requires the approximation of the percentiles for each observation and the inversion of the derivative of the quantile function, thus resulting in an computationally expensive method (Rayner and MacGillivray 2002). In a Bayesian perspective, an Approximate Bayesian Computation (ABC) method has been developed (Allingham et al. 2009; Drovandi and Pettitt 2011) for specific classes of quantile functions, but again at the price of computational burden.

In Gilchrist (2000) two estimation methods based on ‘lack of fit criteria’ are introduced, which are denoted as distributional least absolutes and distributional least squares. The first is based on the minimization of the $L1$ norm between

Table 1 Quantile functions of some probability distributions

Probability distribution	Density function	Quantile function
Exponential	$\theta e^{-\theta x}$	$-\frac{\log(1-u)}{\theta}$
Extreme value	$\frac{1}{\beta} e^{\frac{x}{\beta}} \exp\left[-e^{\frac{x}{\beta}}\right]$	$\beta \log \log(1-u)^{-1}$
Weibull	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$	$\lambda\{\log(1-u)^{-1}\}^{1/k}$
Logistic	$\frac{e^{-(x-\alpha)/\beta}}{\beta\left(1+e^{-(x-\alpha)/\beta}\right)^2}$	$\alpha + \beta \log \frac{u}{(1-u)}$
Double-exponential	$\frac{e^{-\|x\|}}{2}$	$\log 2u, \quad u < 0.5$ $-\log 2(1-u), \quad u > 0.5$
Cauchy	$\frac{1}{\pi(1+x^2)}$	$\tan \pi(u-0.5)$
Pareto	$\frac{\alpha\mu^\alpha}{x^{\alpha+1}}$	$\mu \log(1-u)^{-\frac{1}{\alpha}}$

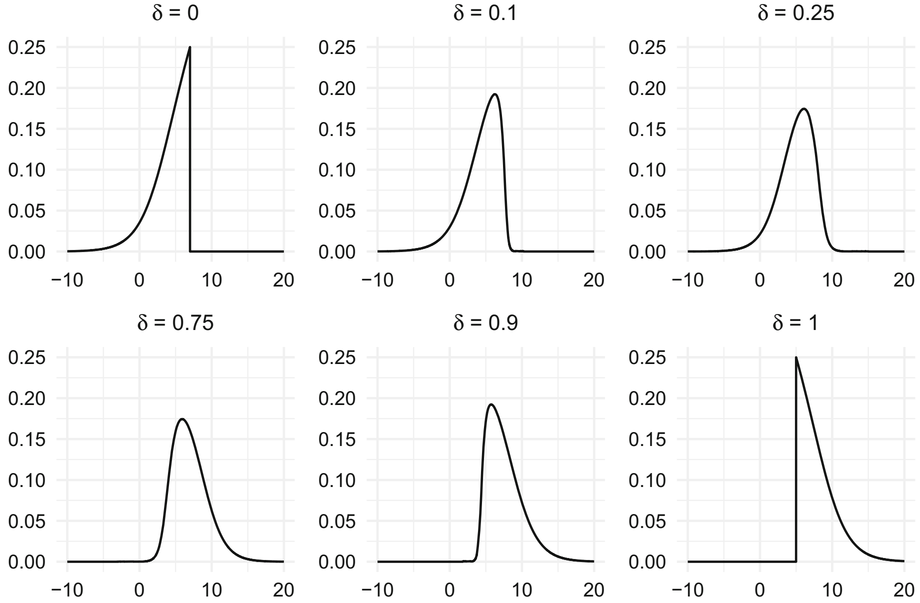

Fig. 1 *fgld* with $\alpha = 5,\ \beta = 2,\ \kappa = 1$ and varying δ

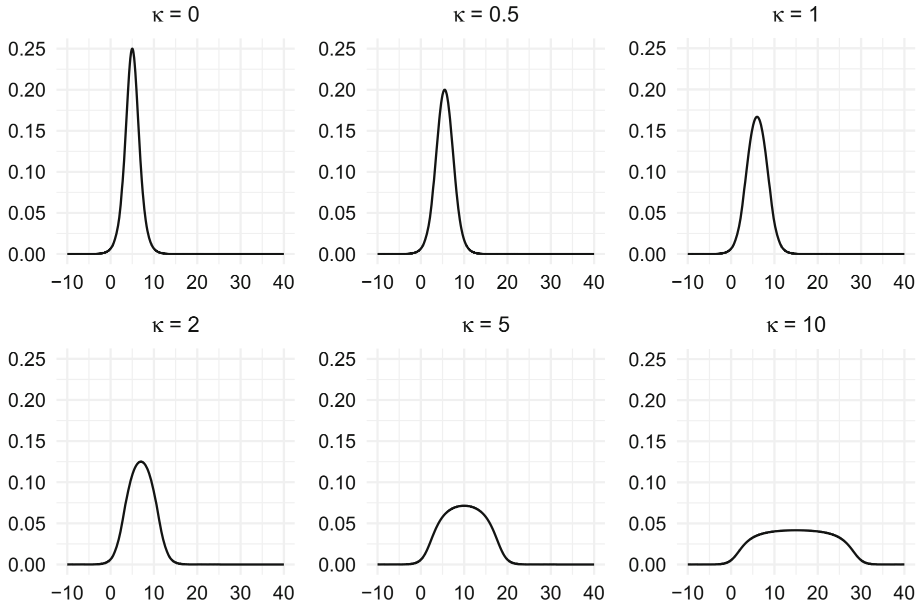

Fig. 2 *fgld* with $\alpha = 5,\ \beta = 2,\ \delta = 0.5$ and varying κ

the ordered statistics and their theoretical median. The second approach consists in minimizing the $L2$ norm between the expected and the observed ordered statistics. Gilchrist highlights that, if no analytical form for the expected order statistics is available, they need to be approximated by a Taylor series expansion. For this reason the author champions the approach of the L1 norm, which does not require such derivation. Here instead, we develop a framework under which the least squares approach can be effectively and efficiently used with a closed form solution, and we also derive some theoretical results.

In fact, there is a specific link between theoretical order statistics and quantile-based distributions (David and Nagaraja 2004). More specifically, the expected value of an order statistic can expressed in terms of the quantile distribution as follows:

$$E[X_{(i)}] = \frac{1}{B(i, n-i+1)} \int_0^1 Q(u;\boldsymbol{\theta}) u^{i-1} (1-u)^{n-i} du. \tag{4}$$

As stated in the following Lemma, if the quantile function is linear in its parameters, the expected value of the theoretical order statistics takes a similar linear form that simplifies the estimation method.

Lemma 1 *If a quantile distribution function is linear with respect to its parameters, then the expected order statistics of that distribution will also be linear with respect to those same parameters.*

The proof is shown in Appendix. Take for instance the simple quantile function $Q(u;\boldsymbol{\theta}) = \theta_0 + \theta_1 u$, with $\theta_1 > 0$ and $\boldsymbol{\theta} = (\theta_0, \theta_1)$. Then by solving the integral in (4) we easily get

$$E[X_{(i)}] = \theta_0 + \theta_1 \frac{i}{n+1} = \begin{bmatrix} 1 & \frac{i}{n+1} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \mathbf{b}_i^\top \boldsymbol{\theta}.$$

For a quantile function with a quadratic term in u, $Q(u;\boldsymbol{\theta}) = \theta_0 + \theta_1 u + \theta_2 u^2$, similarly we get

$$E[X_{(i)}] = \theta_0 + \frac{i}{n+1}\theta_1 + \frac{i\,(i+1)}{(n+2)(n+1)}\theta_2.$$

Thus, for any linear quantile function, the expected values of the order statistics can written as

$$E[X_{(i)}] = \mathbf{b}_i^\top \boldsymbol{\theta},$$

where $\mathbf{b}_i$ are p-dimensional vectors of known coefficients.

Now, given a sample of IID observations $(x_1, \ldots, x_n)$ from $X \sim F(\boldsymbol{\theta})$ denote with $x_{(i)}$ the observed i-th order statistics. We can minimize:

$$\phi(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(x_{(i)} - E[X_{(i)}]\right)^2 = \sum_{i=1}^{n} \left(x_{(i)} - \mathbf{b}_i^\top \boldsymbol{\theta}\right)^2 \tag{5}$$

with respect to $\boldsymbol{\theta}$.

The resulting least squares estimation method is very efficient, since it provides a closed-form solution for the parameters.

By defining $\mathbf{B}$ as the matrix of dimension $n \times p$ having as rows $\mathbf{b}_i$ and by $\mathbf{X}_{(\cdot)}$ the ordered random sample, the estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{X}_{(\cdot)}. \tag{6}$$

Furthermore we have:

$$E[\hat{\boldsymbol{\theta}}] = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top E[\mathbf{X}_{(\cdot)}] = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{B}\,\boldsymbol{\theta} = \boldsymbol{\theta} \tag{7}$$

and

$$V[\hat{\boldsymbol{\theta}}] = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top\, \boldsymbol{\Sigma}\, \mathbf{B} (\mathbf{B}^\top \mathbf{B})^{-1}$$

where $V[\mathbf{X}_{(\cdot)}] = \boldsymbol{\Sigma}$ is the covariance matrix of the order statistics. So the estimator $\hat{\boldsymbol{\theta}}$ is unbiased, but, given the correlation among order statistics, we can not invoke the BLUE property of the Gauss-Markov theorem.

2.2 An example: the flattened generalised logistic distribution

In this section, we derive the results needed for least squares parameter estimation of the flattened generalized logistic (*fgld*) quantile function defined in Eq. (3). To this aim it is convenient to re-parameterise the quantile function as follows:

$$\begin{cases} \alpha = \theta_0 \\ \beta\kappa = \theta_1 \\ \beta(1-\delta) = \theta_2 \\ \beta\delta = \theta_3 \end{cases} \qquad \begin{cases} \alpha = \theta_0 \\ \beta = \theta_2 + \theta_3 \\ \delta = \frac{\theta_3}{\theta_2+\theta_3} \\ \kappa = \frac{\theta_1}{\theta_2+\theta_3} \end{cases}$$

The quantile distribution function of the *fgld* becomes:

$$Q(u) = \theta_0 + \theta_1\, u + \theta_2\, \log u - \theta_3\, \log(1-u) \tag{8}$$

To estimate the parameters via least squares we need to derive the expected value of the order statistics.

Lemma 2 *The expected order statistic of the flattened generalised logistic distribution is equal to:*

$$E[X_{(i)}] = \theta_0 + \theta_1 \frac{i}{n+1} + \theta_2\,(\psi(i) - \psi(n+1)) \\ +\theta_3\,(\psi(n+1) - \psi(n-i+1)) \tag{9}$$

where $\psi(\cdot)$ indicates the digamma function, which is defined as the derivative of the logarithm of the gamma function.

Therefore, in this case we get $\mathbf{b}_i = \Big(1, \frac{i}{n+1}, \psi(i) - \psi(n+1), \psi(n+1) - \psi(n-i+1)\Big)$. For a proof see the Appendix.

In order to compute the variance of the estimator we also need to derive the covariance matrix for the order statistics of the *fgld*.

Lemma 3 *The n-dimensional covariance matrix of the order statistics, $\boldsymbol{\Sigma}$, of the flattened generalised logistic distribution has diagonal variances given by*

$$\begin{aligned} V[X_{(r)}] =& \theta_1^2 \frac{r(n-r+1)}{(n+1)^2(n+2)} + \theta_1\theta_2 \frac{2(n-r+1)}{(n+1)^2} + \\ &+ \theta_1\theta_3 \frac{2r}{(n+1)^2} + \theta_2^2 (\psi_1(r) - \psi_1(n+1)) + \\ &+ \theta_2\theta_3 \, 2\psi_1(n+1) + \theta_3^2 (\psi_1(n-r+1) \\ &- \psi_1(n+1)) \end{aligned}$$

with $r = 1, \ldots, n$ and where $\psi_1(\cdot)$ indicates the trigamma function, which is the derivative of digamma function $\psi(\cdot)$.

The covariance between any two order statistics of the flattened generalised logistic distribution is equal to:

$$\begin{aligned} & Cov[X_{(r)}, X_{(s)}] \\ & = \theta_1^2 \left[\frac{r(n-s+1)}{(n+1)^2(n+2)} \right] + \theta_1\theta_2 \left[\frac{(n-s+1)(r+s)}{(n+1)^2 s} \right] \\ & \quad + \theta_1\theta_3 \left[\frac{r(2n-r-s+2)}{(n+1)^2(n-r+1)} \right] + \theta_2^2 [\psi_1(s) - \psi_1(n+1)] \\ & \quad + \theta_2\theta_3 [(\psi(n+1) - \psi(n-r+1))(\psi(n+1) - \psi(s)) \\ & \quad + \psi_1(n+1)] \\ & \quad + \theta_3^2 [\psi_1(n-r+1) - \psi_1(n+1)] - \theta_2\theta_3 \xi(n, r, s) \end{aligned}$$

where

$$\begin{aligned} \xi(n, r, s) =& \Gamma(s-r)\,\Gamma(n-s+1) \\ & \sum_{h=1}^{\infty} \frac{1}{h} \frac{\Gamma(h+r)}{\Gamma(n+h+1)} (\psi(n+h+1) - \psi(h+s)) \end{aligned}$$

for $r, s = 1, \ldots, n$.

A sketch of the proof in given in the Appendix.

2.3 Asymptotic results

In this section we derive the asymptotic distribution of the estimator of the *fgld* defined in Eq. (6). First notice that this estimator can be expressed as a linear combination of the order statistics:

$$\hat{\boldsymbol{\theta}} = \sum_{i=1}^{n} \boldsymbol{c}_{in} X_{(i)},$$

where the coefficients $\boldsymbol{c}_{in}$ are vectors of the same length p as $\hat{\boldsymbol{\theta}}$.

Lemma 4 *The coefficients $\boldsymbol{c}_{in}$ for the least squares estimator of the* fgld *are continuous and bounded.*

The proof is given in the Appendix. Given this lemma we can derive the following theorem.

Theorem 1 *The least squares estimator for the parameters of the* fgld *linear quantile function has an asymptotically normal distribution:*

$$\hat{\boldsymbol{\theta}} \xrightarrow{d} N_p(\boldsymbol{\theta}, \boldsymbol{\Gamma}) \tag{10}$$

with $\boldsymbol{\Gamma} = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\boldsymbol{\Sigma}\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}$.

The proof of Theorem 1 is shown in the Appendix.

Given the previous result, the null hypothesis that the sample comes from a quantile function with parameters $\boldsymbol{\theta}_0$ can be tested as stated in the following theorem.

Theorem 2 *The null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ can be checked through the test statistic*

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N_p(\mathbf{0},\ \boldsymbol{\Gamma}),$$

where for fgld *quantile function the matrices $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are known quantities derived in Lemma 1 and 3.*

As a simple consequence we can also test the hypothesis that two observed samples come from the same population $H_0 : \mathbf{B}\boldsymbol{\theta}_1 = \mathbf{B}\boldsymbol{\theta}_2$ which is equivalent to $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Under the previous assumptions we get

$$(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2) \xrightarrow{d} N_p(\mathbf{0},\ 2\,\boldsymbol{\Gamma})$$

or alternatively

$$\frac{1}{2}(\hat{\boldsymbol{\theta}}_0 - \hat{\boldsymbol{\theta}}_1)^\top \boldsymbol{\Gamma}^{-1} (\hat{\boldsymbol{\theta}}_0 - \hat{\boldsymbol{\theta}}_1) \xrightarrow{d} \chi^2_p. \tag{11}$$

3 Application to supervised classification

Let Y be a categorical random variables taking values $y = \{1, \ldots, K\}$, where K denotes the total number of classes and let $\mathbf{X} = (X_1, \ldots, X_p)$ be a set of observed variables. One of the most used classification methods in the supervised setting is the so-called naïve Bayes classifier (John and Langley 1995; Hand and Yu 2001). Suppose you have a training data set in which both Y and $\mathbf{X}$ are known. According to the Bayesian rule, the posterior probability of belonging to a generic class k $(k = 1, \ldots, K)$ is

$$Pr(Y = k \mid \mathbf{X} = \mathbf{x}) = \frac{\pi_k f(\mathbf{x} \mid Y = k)}{f(\mathbf{x})}$$

$$= \frac{\pi_k f(\mathbf{x} \mid Y = k)}{\sum_{k'=1}^{K} \pi_{k'} f(\mathbf{x} \mid Y = k')}, \qquad (12)$$

where π_k denotes the proportion of units that belong to class k in the training set.

The naïve Bayes classifier assumes conditional independence of the variables given the categorical response

$$f(\mathbf{x} \mid Y = k) = \prod_{j=1}^{p} f_j(x_j \mid Y = k),$$

thus each variable is treated separately.

The class conditional distributions $f_j(x_j \mid Y = k)$ are usually assumed to be Gaussian. An alternative has been proposed by John and Langley (2013), who suggested the use of kernel density estimation as a tool to allow for more flexible distributional shapes. A further common method is the discretization of all continuous variables, that is estimating the density function via a step function. For this method the main issue is to choose the breaks that define the categories; a recent heuristic proposal is that of Yang and Webb (2009), the so-called proportional discretization. This method achieves (approximately) a discretization with bins having both equal width and equal frequency, with the added advantage that the tuning parameter is derived automatically and based on the sample size (n): $\texttt{width} = \texttt{frequency} \approx \sqrt{n}$.

Quantile-based distributions can be applied in this setting with the goal of taking advantage of their flexible and parsimonious specifications and the fast and reliable estimation given by the least squares method.

The application of quantile-based distributions in the naïve Bayes algorithm involves the estimation of $K \times p$ univariate distributions, similarly to the other methods. Each of the univariate samples is identified by a variable and a category of the response, and their quantile function can be estimated via least squares, provided we choose a linear quantile function. The output of the estimation phase is just a set of parameters: $\boldsymbol{\theta}_{jk}$, with $j = 1, \ldots, p$ and $k = 1, \ldots, K$. Given a single sample identified by a set of variables $\mathbf{x} = (x_1, \ldots, x_p)$, the class conditional distribution is evaluated as follows, for each variable j and categorical response k:

$$P(X_j = x_j \mid Y = k) = f_j(x_j; \boldsymbol{\theta}_{jk}) = \frac{1}{q_j(u_j; \boldsymbol{\theta}_{jk})},$$

where the density is evaluated based on the relationship shown in Eq. (1) and u_j is the inverse of $x_j = Q(u_j; \boldsymbol{\theta}_{jk})$ and needs to be computed numerically in the case of non-invertible quantile functions, such is the case of the *fgld*.

As a by-product of the least square fit, a simple distance measure between two quantile distributions can be derived. Imagine that $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are the estimates of the parameters of two quantile functions. For instance, the quantile function of the classes 1 and 2 of the training sample. Then for each variable we can measure:

$$\|\mathbf{B}\hat{\boldsymbol{\theta}}_1 - \mathbf{B}\hat{\boldsymbol{\theta}}_2\|_2$$

where $\|\ldots\|_2$ denotes the Euclidean distance. The formula can also be interpreted as the Euclidean distance between two vectors containing the expected order statistics for the two distributions.

The formula can be applied seamlessly in the case of two response classes with equal number of observations. When the latter differs between the classes, n can be chosen for instance as the minimum class frequency; when the classes are more than two, the distance can be computed for each pair and the maximum pairwise distance can be retained, meaning that the variable can at least discriminate between those two classes.

This measure can serve to rank variables in terms of their importance, of course limited to their application in the naïve Bayes algorithm. This can be useful in interpreting and explaining the model, in a similar way to the use of variable importance measures derived from algorithms such as random forests.

Moreover, it can serve as the basis of a variable selection procedure as explained in Sect. 2.3 (Theorem 2). Imagine we have $K = 2$ classes, then a variable is relevant for classification if the null $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ is rejected, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ denote the parameters in the two class-populations.

4 Simulation study

In this section we present some empirical studies to evaluate the goodness-of-fit of the illustrated quantile functions in different scenarios, their classification performance in the naïve Bayes algorithm and the behaviour of the asymptotic test.

4.1 Empirical bias

In this first simulation we investigate the goodness-of-fit of three different quantile-based distributions: the simple quantile function with a linear term in u (*linear*), the quantile function with a quadratic term in u (*quad*) and the *fgld*. In order to measure the empirical bias and the variability of the estimators of $\boldsymbol{\theta}$ we compare the observed order statistics with their expectation according to the three models, by computing this empirical bias measure:

$$\sqrt{\frac{\sum_{i=1}^{n}(x_{(i)} - \hat{E}[X_{(i)}])^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_{(i)} - \mathbf{B}\,\hat{\boldsymbol{\theta}})^2}{n}}.$$

Table 2 Average empirical bias over 100 replicates for 4 distributional scenarios (rows) and for 3 quantile-based distributions (columns). Standard deviations are reported in brackets

	Linear	Quad	fgld
Norm	0.22 (0.05)	0.21 (0.05)	0.08 (0.02)
t	0.86 (0.54)	0.84 (0.53)	0.35 (0.38)
Logabst	0.43 (0.13)	0.36 (0.12)	0.13 (0.06)
Exp	1.01 (0.27)	0.72 (0.26)	0.23 (0.1)

We simulated $n = 100$ observations from four different distributions: a standard normal, a T distribution with 3 degrees of freedom, an exponential distribution with rate parameter equal to 0.5, and a $\log(\mid T_{\nu=3} \mid)$, that is the logarithm of the absolute value of a t distribution (again with 3 degrees of freedom). For each scenario we generated 100 replicates. Table 2 shows the mean of the empirical bias across the replicates for each scenario and model. In brackets the standard deviations offer an indication of the variability of the estimates.

Results show that the *fgld* is by far the most flexible model, it being able to fit well in all scenarios.

4.2 Classification

We evaluated the performance of the quantile-based distributions in the naïve Bayes algorithm via a simulation study. We considered the *fgld* and the quantile function with a quadratic term in u (*quad*) described in Sect. 2.1. We generated p variables X_j $(j = 1, \ldots, p)$ of sample size n, according to the four different distributions described in the previous subsection.

We fixed $K = 2$ classes, of equal size $n/2$. Denote X_{j0} the variable X_j when $Y = 0$ and X_{j1} when $Y = 1$. In order to separate the classes we shifted each variable according to the rule

$$X_{j1} = X_{j0} + 0.3\,(-1)^j \quad j = 1, \ldots, p$$

Alteratively, we have applied a scaling as

$$X_{j1} = 0.8\, X_{j0} \quad j = 1, \ldots, p$$

Shifting has been applied to all distributional settings, while scaling has been applied only to the $\log(\mid t_{\nu=3} \mid)$ distribution; thus creating five different scenarios: (1) shifted $N(0, 1)$, (2) shifted $t_{\nu=3}$, (3) shifted $\mathrm{Exp}(\lambda = 0.5)$, (4) shifted $\log(\mid t_{\nu=3} \mid)$ and (5) scaled $\log(\mid t_{\nu=3} \mid)$. For each scenario we let $p = \{10, 50, 100\}$, $n = \{100, 500, 1000\}$, and correlated or independent variables.

The five distributional scenarios, three variable set sizes, three sample sizes and two correlation structures lead to ninety settings. For each setting we repeated data generation and estimation 100 times. Misclassification rates were evaluated on test sets generated in same way as the training samples, and we report the average over the replicates.

Table 3 Computational average times in seconds for training the naïve Bayes classifier and applying its prediction on a test set over the 100 replications for the 5 distributional scenarios. In brackets standard deviations are reported

	Method	$p = 10$	$p = 50$	$p = 100$
n = 100	Discrete	0.06 (0.01)	0.33 (2.54)	1.42 (29.54)
	fgld	0.15 (0.02)	4.62 (61.03)	5.42 (61.08)
	gev	0.04 (0.00)	1.16 (31.51)	1.32 (31.51)
	kde	1.00 (29.54)	0.31 (0.02)	0.62 (0.04)
	Normal	0.02 (0.00)	0.04 (0.01)	0.09 (0.01)
	Quad	0.11 (0.01)	2.46 (43.17)	1.08 (0.06)
n = 500	discrete	0.06 (0.01)	0.26 (0.04)	0.51 (0.03)
	fgld	0.64 (0.03)	3.25 (0.14)	6.54 (0.16)
	gev	0.08 (0.01)	0.41 (0.08)	0.80 (0.12)
	kde	0.34 (0.02)	1.59 (0.10)	3.13 (0.12)
	Normal	0.08 (0.01)	0.23 (0.03)	0.41 (0.03)
	Quad	0.54 (0.04)	2.74 (0.18)	5.48 (0.29)
n = 1000	discrete	0.06 (0.01)	0.27 (0.02)	0.48 (0.07)
	fgld	1.31 (0.04)	6.46 (0.19)	11.94 (1.31)
	gev	0.15 (0.03)	0.72 (0.16)	1.33 (0.31)
	kde	0.70 (0.03)	3.19 (0.10)	5.85 (0.60)
	Normal	0.16 (0.02)	0.46 (0.03)	0.76 (0.09)
	Quad	1.11 (0.07)	5.48 (0.27)	10.12 (1.21)

We compared with other choices for the class-conditional distributions; namely the normal, the kernel (kde), with default Silverman's rule for the bandwidth, the discrete method (with proportional discretization (Yang and Webb 2009)), the generalized extreme value distribution (*gev*) estimated via maximum likelihood by the R package `evd`.

Table 3 contains a summary of the computational times for this simulation. We can note that the time needed for the methods based on the least squares estimation of quantile functions is longer than for simpler methods such as the normal and the discrete, but it is manageable even for the larger data sets. Times are particularly affected by the increase in the number of independent variables (p).

Results for the classification are presented graphically in Fig. 3 for each data generating distribution, where we collapse over the 18 settings evaluated for each case. We show scaled differences with respect to a reference method for each setting; we choose *fgld* as the reference. The scaled differences are computed as follows:

$$d_{jk} = \frac{e_{jk} - e_{j1}}{\bar{e}_j}$$

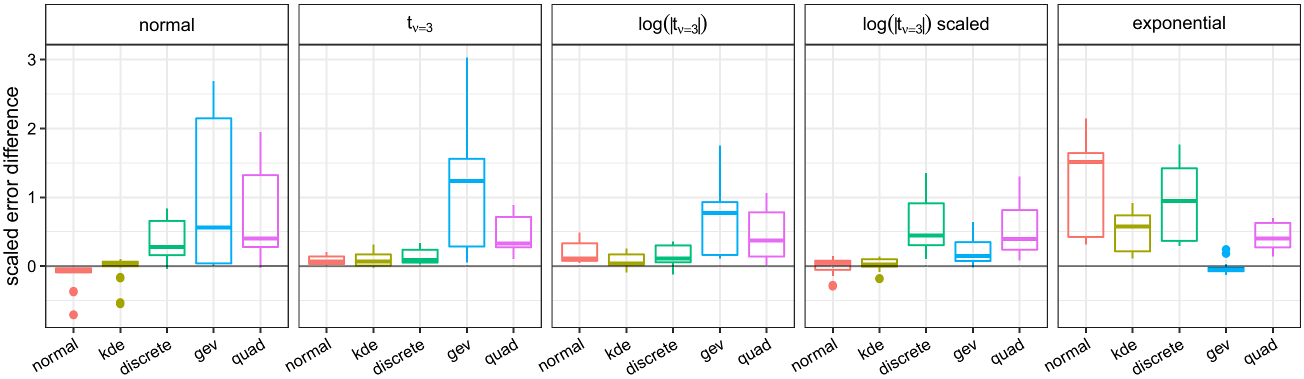

Fig. 3 Results from a simulation study comparing different methods for the naïve Bayes classifier. Each panel represents a distributional scenario under which the data was simulated. Results are presented as scaled differences from the *fgld*, where a value higher than 0 means that for a setting (combination of sample size, number of variable and correlation structure) the method had a larger mean misclassification error than the *fgld*

where $j = 1, \ldots, 18$ indicates the setting for fixed data generating distribution, $k = 1, \ldots, 5$ represents the method (with 1 being the reference method), and $\bar{e}_j$ being the average test error for that setting. From Fig. 3 we can see that *fgld* is very competitive: as expected it performs worse than the normal when the data are indeed normal, but the discrepancy is minimal; it is the best method otherwise with the exception of the exponential data when only *gev* performs better.

4.3 Testing procedure

In order to evaluate the performance of the test we assess the distribution of the test statistic for the *fgld* and for the *quad* quantile functions under the null hypothesis $H_0 : \theta_1 = \theta_2$, and the power of the test when the null hypothesis is not true. The variance of the order statistics of the *quad* quantile function, needed for the variance of the least squares estimator, is reported in the Appendix.

Type I error

Under the null hypothesis the two samples come from the same distribution. In order to evaluate the convergence of the test statistic to its null distribution we compare empirical type I errors with the nominal significance level that has been chosen in advance.

A total of 200 sets of parameters have been randomly generated, and for each of them 1,000 two-group samples have been simulated. From each of these 1,000 data sets the test statistic can be computed and the empirical type I error corresponds to the proportion of test statistics above the critical value (the 95$^{\text{th}}$ quantile of the $\chi^2_{df=4}$ distribution for the *fgld* and the $\chi^2_{df=3}$ for the *quad*). This procedure has been repeated for different group sample sizes, with the same parameter sets, and the results are shown in Fig. 4. As could be expected, the empirical type I error converges to the nominal one as the sample size increases in both cases.

ROC curves

To evaluate the power of the test we have simulated data sets of 1,000 variables, with half of those variables having a different distribution between the two balanced groups, and half having the same distribution. For each variable the p-value associated with the test statistic is computed.

This problem can be re-framed as a classification problem in which the response is whether or not the variable is useful (having a different or equal distribution across the two groups).

In the simulation we know whether the variable is useful or not, so we can evaluate it with the metrics of a classification model, such as a ROC curve. This is particularly suited to the test because the different thresholds (and subsequent classifications) can be interpreted as significance levels.

In Fig. 5 we report the ROC curves for the *fgld* and *quad* that evaluate whether test statistics are able to identify correctly useful and not useful variables. In both cases we can see that as n increases the curves move more and more towards the top left corner. Even with low sample sizes there are cut-off points for which the test performs extremely well both in terms of sensitivity and of specificity.

5 Real data examples

5.1 Benchmark datasets

We have compared the different methods for the naïve Bayes classifier used in Sect. 4.2 on some real datasets commonly used for benchmarking. The chosen datasets are all publicly

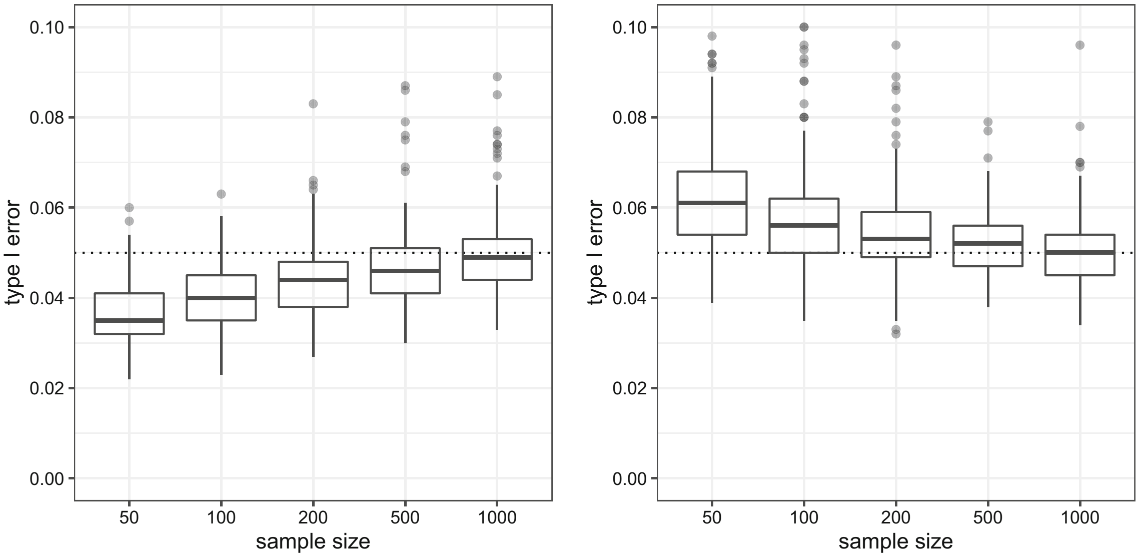

Fig. 4 Distribution of empirical type I errors across 200 parameter sets for the *fgld* (left panel) and *quad* (right panel) for different group sample sizes. As the sample size increases, empirical type I errors get closer to their nominal 5% value. The left panel refers to the *fgld*, the right panel to the *quad*

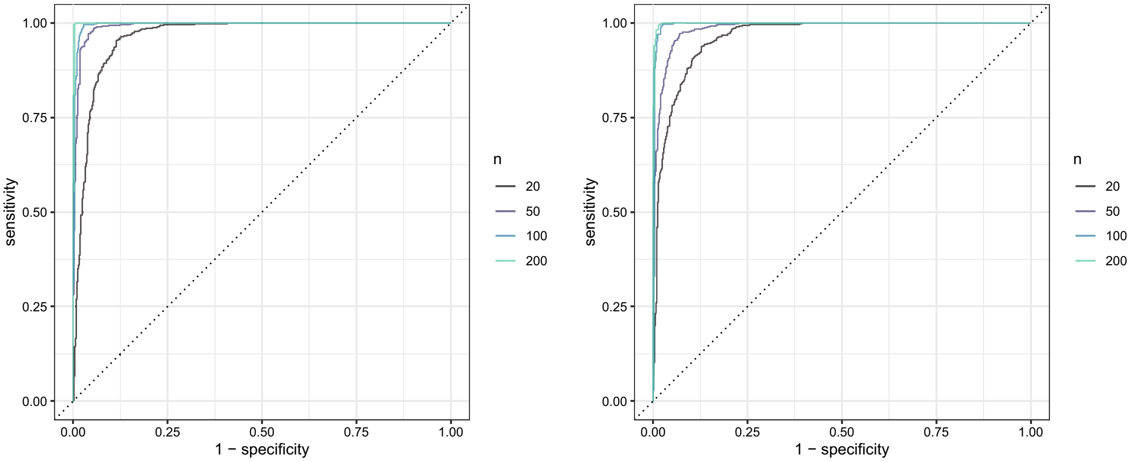

Fig. 5 ROC curves based on the identification of whether a variable has the same distribution across two groups. Results are obtained by computing the hypothesis test across 1,000 variables, of which only half have the same distribution across the two groups

available from the UCI machine learning repository (Dua and Graff 2019). When available we used the preprocessed version from the R package `mlbench` (Leisch and Dimitriadou 2021). In Table 4 some basic information of the datasets used is provided: we can note the general adaptability of the naïve Bayes classifier, being able to deal with both numerical and categorical variables at the same time and with multi-class response variables.

On these data we fitted the models that performed the best in the simulation study (Sect. 4.2), namely the *fgld*, the normal, the kde and the discrete. Results in terms of accuracy from tenfold cross-validation are presented in Table 5. We can note that no method is uniformly superior to the others. In general, the additional flexibility given by the *fgld*, the kde and the discrete, with respect to the normal, proves advantageous. We can note the *fgld* performs comparatively well and there are multiple datasets where it achieves the maximum accuracy.

5.2 Variable selection

In this section we illustrate the proposed strategy for variable selection on a real dataset. We revisit data from Altman (1968), available in the R package `MixGHD` (Tortora et al. 2021), by adding noise variables. The original dataset contains information about $n = 66$ companies that have filed for bankruptcy. Our task is to predict the status of the firms (0 for ‘bankruptcy’ or 1 for ‘financially sound’). The original

Table 4 Datasets from the UCI Machine learning repository used for comparing naïve Bayes methods, with some information regarding data size and type

	Sample size	Numerical variables	Categorical variables	Response classes
Cleveland	297	6	7	2
Credit	653	6	9	2
Diabetes	768	8	0	2
Glass	214	9	0	6
Heart	270	6	7	2
Ionosphere	351	32	2	2
Letter	20000	16	0	26
Sonar	208	60	0	2
Thyroid	2751	6	21	2
Vehicle	752	18	0	4
Waveform	5000	40	0	3
wbcd	569	30	0	2

Table 5 Accuracy from different naïve Bayes methods (columns) applied on 12 benchmark datasets (rows). The results are obtained from tenfold cross-validation

	fgld	Normal	kde	Discrete
Cleveland	80.79	80.13	80.46	82.15
Credit	80.36	73.94	76.28	84.36
Diabetes	76.05	75.39	75.01	65.24
Glass	57.58	45.84	54.55	53.23
Heart	82.96	81.48	81.11	82.22
Ionosphere	73.23	82.35	91.75	88.07
Letter	65.39	64.28	70.48	51.57
Sonar	70.18	67.63	75.49	74.90
Thyroid	93.20	93.42	95.02	92.08
Vehicle	59.32	44.92	57.15	62.50
Waveform	80.28	80.00	79.86	75.24
wbcd	94.73	92.95	93.67	88.90

predictors are two measurements related to the earnings of the firm. On top these two relevant variables we added 198 irrelevant variables sampled from a standard normal distribution, for a total $p = 200$. The goal is to check whether the variable selection procedure developed in Sect. 3 is able to identify the two real variables, and then to compare the accuracy of various naïve Bayes classification algorithms in the complete dataset and with some other values of p.

To this aim we considered the naïve Bayes classifiers, with the previously used methods for estimating the distribution (normal, kde, discretization and *fgld*). We also compare these classifiers to other commonly used ones: k-nearest neighbors with $k = 3$, logistic regression and linear discriminant analysis.

First we computed the p values associated with the test for each variable, and by using a procedure for controlling the false discovery rate (the Benjamini-Hochberg procedure), we correctly reject the null hypothesis only for the two original variables. Next, we re-ordered variables in ascending order by the obtained p values and we compare the classifiers in datasets with an increasing number of variables, where variables with progressively higher p values are included. Results are shown in Table 6 for values of $p = 2, 50, 100, 150, 200$. A visual representation of the naïve Bayes with the *fgld* is shown in Fig. 6, where the first 7 variables in terms of p-value are visualised, separated by class, with a histogram and the density from the estimated *fgld*. It can be noted how the *fgld* can capture the skewness present in the first two original variables.

We can note that the naïve Bayes with the *fgld* reaches its maximum with $p = 2$, that is with the original variables. This is the best accuracy obtained in a leave-one-out cross validation scheme, and the method is the best strategy together with logistic regression. As more and more noise variables are included the performance of all methods deteriorates, with the naïve Bayes classifiers being pretty robust. This robustness, in particular of the normal and KDE naïve Bayes classifiers, has also been noted by the fact that it can happen that they retain or improve their accuracy even in presence of a moderate number of noisy variables, probably due random changes related to the small number of units n. However, the improvement given by the selection is sizable for all methods, and most of them benefit from the selection given by the *fgld* test, reaching very high accuracies when only the two original variables are included.

6 Concluding remarks

We focused on the family of linear quantile functions and in particular on the so-called generalized flattened logistic distribution (*fgld*). We showed a least squares estimation procedure and we derived its properties. The resulting estimators

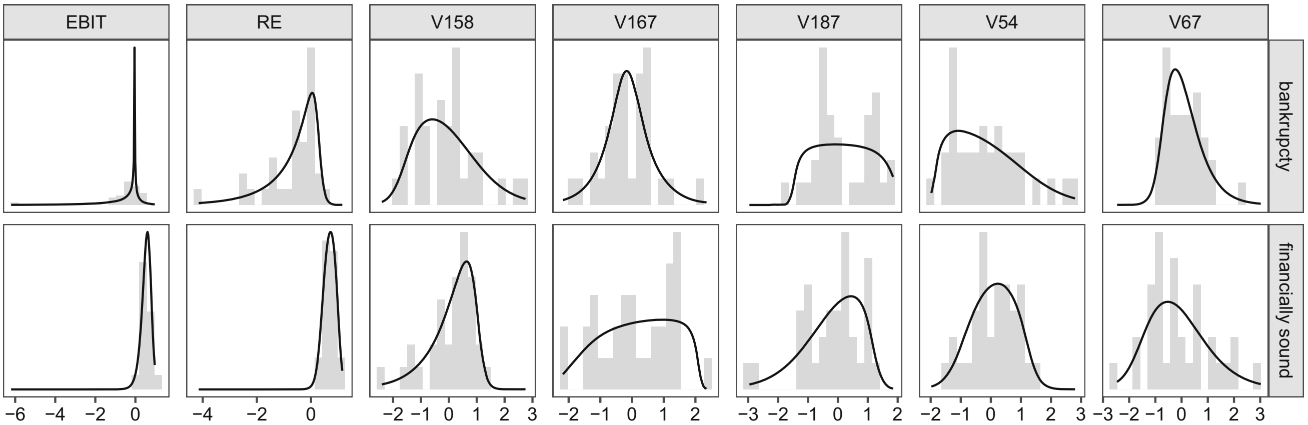

Fig. 6 Naïve Bayes classifier with *fgld* applied to the `bankruptcy` dataset with added noise variables. The 7 variables with the lowest p values are shown on the columns, while the rows identify the response class. The visualisation includes a histogram and the density function from the estimated *fgld*

Table 6 Leave-one-out cross validation accuracy for different classification algorithms applied to the `bankruptcy` dataset with added noise variables. The columns are for different numbers of variables (p), being the ones with the lowest p values for the *fgld* test

	$p = 2$	$p = 50$	$p = 100$	$p = 150$	$p = 200$
KNN k = 3	92.42	65.15	54.55	43.94	43.94
LDA	90.91	72.73	54.55	57.58	46.97
Logistic regression	95.45	56.06	53.03	53.03	53.03
Naïve Bayes discrete	84.85	87.88	84.85	71.21	63.64
Naïve Bayes *fgld*	95.45	87.88	78.79	69.70	60.61
Naïve Bayes KDE	93.94	92.42	95.45	81.82	69.70
Naïve Bayes normal	93.94	92.42	90.91	84.85	77.27

are unbiased and asymptotically normal, thus allowing us to derive a testing procedure. In the numerical experiments we have investigated the performance of the asymptotic test with different sample sizes. Results show that even with low sample sizes the asymptotic test has some acceptable power.

In principle one could consider any other linear quantile function, provided that the first and second moment of the order statistics can be derived, which are necessary respectively for the least square estimator and the testing procedure. We remark though that in our simulation and real data experiments the *fgld* distribution, characterized by four parameters, seems to be flexible enough to capture a wide range of shapes. The theoretical results about the *fgld* distribution have been then used to propose a novel naïve Bayes classifier, based on the quantile distribution rather than the conventional Gaussian density. The *fgld* quantile function performed very well in all the empirical studies. As by-products, strategies for variable importance and variable selection have been obtained by the simple application of the testing procedure developed in the first part of the work. Notice that the naïve Bayes classifier under the assumption of conditional independence requires univariate densities, and for this reason the *fgld* quantile distribution represents a useful and flexible tool. A challenging extension for future work is to develop an inferential framework for multivariate quantile functions, in the spirit of Farcomeni et al. (2022), with potentially different applications and statistical purposes. One could also consider an extension to quantile regression, where we speculate that the evaluation of the impact of changes in explanatory variables on marginal distributions of an outcome could be straightforward within the family of linear quantile functions (Firpo et al. 2009).

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Declarations

Conflict of interest No conflit of interest or funding to declare.

Appendix

Proof of Lemma 1

We assume that the quantile distribution function is linear with respect to parameters $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_p\right)$.

$$Q(u) = \theta_1 \, h_1(u) + \cdots + \theta_p \, h_p(u).$$

The expected value of the i-th order statistic can be written as follows, where $g(u)$ is the density of a Beta distribution with parameters equal to i and $n - i + 1$.

$$\begin{aligned}
E(X_{(i)}) &= \int_0^1 Q(u)\, g(u)\, du = \\
&= \int_0^1 \left[\theta_1 \, h_1(u) + \cdots + \theta_p \, h_p(u)\right] g(u)\, du = \\
&= \int_0^1 \left[\theta_1 \, h_1(u) g(u) + \cdots + \theta_p \, h_p(u) g(u)\right]\, du = \\
&= \theta_1 \left[\int_0^1 h_1(u) g(u)\, du\right] + \ldots \\
&\quad + \theta_p \left[\int_0^1 h_p(u) g(u)\, du\right] = \\
&= \theta_1 \, b_{1i} + \cdots + \theta_p \, b_{pi}
\end{aligned}$$

This shows that the expected value of a generic order statistic is linear with respect to those same parameters. Alternatively we can think of the proof in terms of maps, the quantile distribution function

$$Q : \boldsymbol{\theta} \to Q(\boldsymbol{\theta})$$

is a linear map by hypothesis with the following two defining properties:

$$\begin{aligned}
&Q(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) = Q(\boldsymbol{\theta}_1) + Q(\boldsymbol{\theta}_2) \\
&Q(\alpha \, \boldsymbol{\theta}_1) = \alpha \, Q(\boldsymbol{\theta}_1)
\end{aligned}$$

the expected value

$$E : Q \to E(X_{(i)})$$

is also a linear map (a definite integral is a linear map from the space of all real-valued integrable functions to $\mathbb{R}$). The composition of linear maps is linear, so $E \circ Q$ is linear.

Proof of Lemma 2

To obtain the expected value of the i-th order statistic of a sample of size n we need to solve the following integral:

$$\begin{aligned}
E[X_{(i:n)}] = \frac{1}{B(i,\, n-i+1)} \int_0^1 &[\theta_0 + \theta_1 \, u + \theta_2 \, \log u \\
&- \theta_3 \, \log(1-u)] \, u^{i-1} (1-u)^{n-i} \, du
\end{aligned}$$

The first two additive terms are easily solvable by recognizing the beta function:

$$\begin{aligned}
&\int_0^1 u^{i-1} (1-u)^{n-i} \, du = B(i, n-i+1) \\
&\int_0^1 u^{i} (1-u)^{n-i} \, du = B(i+1, n-i+1)
\end{aligned}$$

For solving the third term we can use the following rule, in which a and b are two positive real numbers.

$$\begin{aligned}
&\int_0^1 \log x \, x^{a-1} \, (1-x)^{b-1} \, dx \\
&= \int_0^1 \frac{\partial}{\partial a} x^{a-1} \, (1-x)^{b-1} \, dx \\
&= \frac{\partial \, B(a,\, b)}{\partial \, a} = \frac{\partial}{\partial \, a} \, \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)} \\
&= \frac{\Gamma'(a)\Gamma(b)\Gamma(a+b) - \Gamma(a)\Gamma(b)\Gamma'(a+b)}{\Gamma(a+b)^2} \\
&= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \left[\frac{\Gamma'(a)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)}\right] \\
&= B(a,\, b) \, (\psi(a) - \psi(a+b))
\end{aligned}$$

In a similar way it can be shown that:

$$\begin{aligned}
&\int_0^1 \log(1-x) \, x^{a-1} \, (1-x)^{b-1} \, dx \\
&\quad = B(a, b)(\psi(b) - \psi(a+b))
\end{aligned}$$

Thus the third and fourth term are equal respectively to:

$$\begin{aligned}
&\int_0^1 \log(u) u^{i-1} (1-u)^{n-i} \, du \\
&\quad = B(i,\, n-i+1) \, (\psi(i) - \psi(n+1)) \\
&\int_0^1 \log(1-u) u^{i-1} (1-u)^{n-i} \, du \\
&\quad = B(n-i+1,\, i) \, (\psi(n-i+1) - \psi(n+1))
\end{aligned}$$

By adding together the terms multiplied by their respective parameters and simplifying the beta functions the final result is obtained.

Proof of Lemma 3

The covariance between the r-th and s-th order statistics is given by the following integral (David and Nagaraja 2004):

$$
\begin{aligned}
Cov[X_{(r)}, X_{(s)}] = &\frac{n!}{(r-1)!(s-r-1)!(n-s)!} \\
&\int_0^1 \int_0^v (Q(u) - E[X_{(r)}])(Q(v) - E[X_{(s)}]) \\
&u^{r-1}(v-u)^{s-r-1}(1-v)^{n-s}\, du\, dv
\end{aligned}
$$

Denoting the product of factorials before the double integral as $C_{n,r,s}$, the expected values of the order statistics as μ_r and μ_s, and carrying out the product of the first two terms in the integral, the formula can be rewritten as:

$$
\begin{aligned}
&Cov[X_{(r)}, X_{(s)}] \\
&= C_{n,r,s} \int_0^1 \int_0^v Q_u\, Q_v\, u^{r-1}(v-u)^{s-r-1} \\
&\quad \times (1-v)^{n-s}\, du\, dv - \mu_r \mu_s
\end{aligned}
$$

Given that the quantile function for the *fgld* has 4 terms, the product $Q_u\, Q_v$ will have 16 terms, so the integral can be split into 16 parts that can be tackled one at the time. For instance, the solution of one of these 16 terms, up to the multiplicative constant $-\theta_2\,\theta_3$, is shown below:

$$
\begin{aligned}
&C_{n,r,s} \int_0^1 \int_0^v \log(u)\, \log(1-v)\, u^{r-1}(v-u)^{s-r-1} \\
&\quad \times (1-v)^{n-s}\, du\, dv \\
&= C_{n,r,s} \int_0^1 \log(1-v)(1-v)^{n-s} \\
&\quad \int_0^v \log(u)\; u^{r-1}(v-u)^{s-r-1}\, du\, dv \\
&= C_{n,r,s} \int_0^1 \log(1-v)(1-v)^{n-s} v^{s-1} \\
&\quad \int_0^1 \log(vt)\; t^{r-1}(1-t)^{s-r-1}\, dt\, dv \\
&= C_{n,r,s} \int_0^1 \log(1-v)(1-v)^{n-s} v^{s-1}\, B(r, s-r) \\
&\quad \left[\log(v) + \psi(r) - \psi(s)\right] dv \\
&= C_{n,r,s} B(r, s-r) \int_0^1 \log(1-v)(1-v)^{n-s} v^{s-1} \\
&\quad \left[\log(v) + \psi(r) - \psi(s)\right] dv \\
&= [\psi(n-s+1) - \psi(n+1)]\,[\psi(r) - \psi(n+1)] \\
&\quad - \psi_1(n+1)
\end{aligned}
$$

The only integral that, to our understanding, has no easy expression through the identification of special functions is the following (up to the constant $-\theta_2\,\theta_3$), whose solution involves a series:

$$
\begin{aligned}
&C_{n,r,s} \int_0^1 \int_0^v \log(1-u)\; \log(v)\, u^{r-1} \\
&\quad (v-u)^{s-r-1}(1-v)^{n-s}\, du\, dv \\
&= C_{n,r,s} \int_0^1 \log(v)\,(1-v)^{n-s}\, v^{s-1} \int_0^1 \log(1-vt)\, t^{r-1} \\
&\quad (1-t)^{s-r-1}\, dt\, dv \\
&= C_{n,r,s} \int_0^1 \log(v)\,(1-v)^{n-s}\, v^{s-1} \int_0^1 \\
&\quad \sum_{h=1}^{\infty} \frac{-(vt)^h}{h}\, t^{r-1}(1-t)^{s-r-1}\, dt\, dv \\
&= -C_{n,r,s} \sum_{h=1}^{\infty} \frac{B(h+r, s-r)}{h} \int_0^1 \log(v)\,(1-v)^{n-s}\, v^{h+s-1} dv \\
&= -C_{n,r,s} \sum_{h=1}^{\infty} \frac{B(h+r, s-r)}{h} \frac{\partial}{\partial h} \int_0^1 (1-v)^{n-s}\, v^{h+s-1} dv \\
&= -C_{n,r,s} \sum_{h=1}^{\infty} \frac{B(h+r, s-r)}{h} \\
&\quad B(n-s+1, h+s)(\psi(h+s) - \psi(n+h+1)) \\
&= \frac{\Gamma(n+1)}{\Gamma(r)} \sum_{h=1}^{\infty} \frac{1}{h} \frac{\Gamma(h+r)}{\Gamma(n+h+1)} (\psi(n+h+1) - \psi(h+s))
\end{aligned}
$$

After solving the 16 integrals and getting the 16 terms from the product $\mu_r\,\mu_s$, terms with the same parameters can be collected: all of the terms involving θ_0 cancel out in the difference and the 6 combinations that are left make up the terms shown in the resulting expression.

Proof of Lemma 4

The least squares estimator for the *fgld* distribution is given by Eq. (6). The coefficients $\boldsymbol{c}_{in}$ that form the linear combination of order statistics are defined as follows:

$$
\hat{\boldsymbol{\theta}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{x}_{(\cdot)} = \left[\boldsymbol{c}_{1n}\; \boldsymbol{c}_{2n}\; \cdots\; \boldsymbol{c}_{1nn}\right] \begin{bmatrix} x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(n)} \end{bmatrix},
$$

that is they constitute the columns of the $p \times n$ matrix $(\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top$. To prove that they are bounded it is enough to prove that each of the elements in the matrix $(\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top$ is bounded.

We start by expanding matrix $\mathbf{B}$:

$$\mathbf{B} = \begin{bmatrix} 1 & \frac{1}{n+1} & \psi(1) - \psi(n+1) & \psi(n+1) - \psi(n) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \frac{i}{n+1} & \psi(i) - \psi(n+1) & \psi(n+1) - \psi(n-i+1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \frac{n}{n+1} & \psi(n) - \psi(n+1) & \psi(n+1) - \psi(1) \end{bmatrix}$$

The product $\mathbf{B}^\top\mathbf{B}$ can be analytically defined up to the 4 entries that involve the summations involving the digamma functions. For them we can only define an asymptotic order, which we will denote as k. In the following it will be shown that for any $k > 1$ the boundedness of the coefficients is preserved:

$$\mathbf{B}^\top\mathbf{B} = \begin{bmatrix} n & \frac{n}{2} & -n & n \\ \frac{n}{2} & \frac{n(1+2n)}{6(1+n)} & \frac{-3n-n^2}{4(n+1)} & \frac{3n^2+n}{4(n+1)} \\ -n & \frac{-3n-n^2}{4(n+1)} & \mathcal{O}(n^k) & \mathcal{O}(n^k) \\ n & \frac{3n^2+n}{4(n+1)} & \mathcal{O}(n^k) & \mathcal{O}(n^k) \end{bmatrix}$$

Next we need to compute the inverse of $\mathbf{B}^\top\mathbf{B}$. To this aim we will use the formula for a block diagonal matrix in order to reframe the problem in terms of the inversion 2×2 matrices (Petersen and Pedersen 2012). First we identify four 2×2 blocks in $\mathbf{B}^\top\mathbf{B}$:

$$\mathbf{B}^\top\mathbf{B} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

then the inverse is defined as:

$$(\mathbf{B}^\top\mathbf{B})^{-1} = \begin{bmatrix} \mathbf{C}_1^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{C}_2^{-1} \\ -\mathbf{C}_2^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{C}_2^{-1} \end{bmatrix},$$

where

$$\mathbf{C}_1 = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$
$$\mathbf{C}_2 = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}.$$

In the following we derive the submatrices and their combinations needed for the inverse, we will assume that the determinants written in big O notation are not zero, so that the inverse can be computed.

$$\mathbf{A}_{22}^{-1} = \det(\mathbf{A}_{22})^{-1} \begin{bmatrix} \mathcal{O}(n^k) & \mathcal{O}(n^k) \\ \mathcal{O}(n^k) & \mathcal{O}(n^k) \end{bmatrix} = \mathcal{O}(n^{-2k})$$

$$\begin{bmatrix} \mathcal{O}(n^k) & \mathcal{O}(n^k) \\ \mathcal{O}(n^k) & \mathcal{O}(n^k) \end{bmatrix} = \begin{bmatrix} \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \\ \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \end{bmatrix}$$

$$\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} = \begin{bmatrix} \frac{7n^2+n}{4n+4} & -\frac{n(n+7)}{4(n+1)} \\ -\frac{n(n+7)}{4(n+1)} & \frac{7n^2+n}{4n+4} \end{bmatrix} = \begin{bmatrix} \mathcal{O}(n) & \mathcal{O}(n) \\ \mathcal{O}(n) & \mathcal{O}(n) \end{bmatrix}$$

$$\mathbf{C}_2 = \begin{bmatrix} \mathcal{O}(n^k) & \mathcal{O}(n^k) \\ \mathcal{O}(n^k) & \mathcal{O}(n^k) \end{bmatrix} - \begin{bmatrix} \mathcal{O}(n) & \mathcal{O}(n) \\ \mathcal{O}(n) & \mathcal{O}(n) \end{bmatrix} = \begin{bmatrix} \mathcal{O}(n^k) & \mathcal{O}(n^k) \\ \mathcal{O}(n^k) & \mathcal{O}(n^k) \end{bmatrix}$$

$$\mathbf{C}_2^{-1} = \begin{bmatrix} \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \\ \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \end{bmatrix}$$

$$\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} = \begin{bmatrix} \mathcal{O}(n) & \mathcal{O}(n) \\ \mathcal{O}(n) & \mathcal{O}(n) \end{bmatrix} \begin{bmatrix} \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \\ \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \end{bmatrix}$$

$$\begin{bmatrix} \mathcal{O}(n) & \mathcal{O}(n) \\ \mathcal{O}(n) & \mathcal{O}(n) \end{bmatrix} = \begin{bmatrix} \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-1}) \\ \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-1}) \end{bmatrix}$$

$$\mathbf{C}_1 = \begin{bmatrix} \mathcal{O}(n) & \mathcal{O}(n) \\ \mathcal{O}(n) & \mathcal{O}(n) \end{bmatrix} - \begin{bmatrix} \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-1}) \\ \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-1}) \end{bmatrix} = \begin{bmatrix} \mathcal{O}(n) & \mathcal{O}(n) \\ \mathcal{O}(n) & \mathcal{O}(n) \end{bmatrix}$$

$$\mathbf{C}_1^{-1} = \begin{bmatrix} \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-1}) \\ \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-1}) \end{bmatrix}$$

$$\mathbf{A}_{11}^{-1}\mathbf{A}_{12} = \left[\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\right]^\top = \begin{bmatrix} -\frac{5}{2} & -\frac{1}{2} \\ 3 & 3 \end{bmatrix}$$

$$(\mathbf{B}^\top\mathbf{B})^{-1} = \begin{bmatrix} \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \\ \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-1}) & \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \\ \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \\ \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) & \mathcal{O}(n^{-k}) \end{bmatrix}$$

The final step is to multiply the inverse we have just derived by the transpose of $\mathbf{B}$, which we will write in asymptotic notation:

$$\mathbf{B}^\top = \begin{bmatrix} \mathcal{O}(1) & \cdots & \mathcal{O}(1) \\ \mathcal{O}(1) & \cdots & \mathcal{O}(1) \\ \mathcal{O}(k) & \cdots & \mathcal{O}(k) \\ \mathcal{O}(k) & \cdots & \mathcal{O}(k) \end{bmatrix}$$

The final matrix $(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top$ will contain terms of order 1 ($\mathcal{O}(1)$), that is bounded, or below (from n^{-1} to n^{-k}), so all the entries of the coefficients $\boldsymbol{c}_{in}$ are bounded.

Moreover, to prove that the functions that produce the coefficients $\boldsymbol{c}_{in}$ are continuous it is enough to note that although no analytical form for the functions is available, they are the result of products and sums of the continuous functions that define the columns of $\mathbf{B}$, so they will also be continuous.

Proof of Theorem 1

The theorem is based on the application of an asymptotic result regarding the linear combinations of order statistics (David and Nagaraja 2004, Theorem 11.4). The linear combination is denoted as:

$$L_n = \frac{1}{n}\sum_{i=1}^{n} J\left(\frac{i}{n}\right) X_{(i)},$$

where the coefficients are $\boldsymbol{c}_{in} = \frac{1}{n}J\left(\frac{i}{n}\right)$. In our case L_n is the vector of the least squares estimator $\hat{\boldsymbol{\theta}}$. The conditions for the asymptotic normality of L_n are that the variance of the distribution X is finite, which is true for the *fgld*, and that the functions $J(u)$ that define coefficients $\boldsymbol{c}_{in}$ are bounded and continuous, which is shown in Lemma 4. The expected value and variance of the limiting normal distribution are given by the ones of the linear combination. In our case these are equal respectively to the theoretical value of the parameters and the variance of the least squares estimator, for which–in the case of the *fgld*–we have an exact result, thanks to Lemmas 2 and 3.

Variance of the order statistics for the *quad* quantile function

$$
\begin{aligned}
&Cov[X_{(r)},\ X_{(s)}]\\
&\quad=\theta_1^2\ \frac{r(n-s+1)}{(n+1)^2(n+2)}+\\
&\qquad+\theta_1\,\theta_2\ \frac{2r(n-s+1)(r+s+2)}{(n+1)^2(n+2)(n+3)}\\
&\qquad+\theta_2^2\ \frac{2r(r+1)(n-s+1)(n(2s+3)+5s+6)}{(n+1)^2(n+2)^2(n+3)(n+4)}
\end{aligned}
$$

References

Allingham, D., King, R.A.R., Mengersen, K.L.: Bayesian estimation of quantile distributions. Statist. Comput. **19**(2), 189–201 (2009)

Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Financ. **23**(4), 589–609 (1968)

Chakrabarty, T.K., Sharma, D.: A generalization of the quantile-based flattened logistic distribution. Ann. Data. Sci. **8**(3), 603–627 (2021)

David, H.A., Nagaraja, H.N.: Order Statistics. Wiley, Hoboken (2004)

Drovandi, C.C., Pettitt, A.N.: Likelihood-free Bayesian estimation of multivariate quantile distributions. Comput. Stat. Data Anal. **55**(9), 2541–2556 (2011)

Dua, D., Graff, C.: UCI machine learning repository (2019)

Farcomeni, A., Geraci, M., Viroli, C.: Directional quantile classifiers. J. Comput. Graph. Stat. **31**, 907–916 (2022)

Firpo, S., Fortin, N.M., Lemieux, T.: Unconditional quantile regressions. Econometrica **77**, 953–973 (2009)

Freimer, M., Kollia, G., Mudholkar, G.S., Lin, C.T.: a study of the generalized Tukey lambda family. Commun. Statist. Theory Methods **17**(10), 3547–3567 (1988)

Gilchrist, W.: Statistical Modelling with Quantile Functions. Taylor and Francis, Andover (2000)

Hand, D., Yu, K.: Idiot's Bayes–Not so Stupid After All? Int. Stat. Rev. **69**, 385–398 (2001)

Haynes, M.A., MacGillivray, H.L., Mengersen, K.L.: Robustness of ranking and selection rules using generalised g-and-k distributions. J. Statist. Plan. Inference **65**(1), 45–66 (1997)

Jiang, L., Zhang, H., Cai, Z.: A novel Bayes model: hidden Naive Bayes. IEEE Trans. Knowl. Data Eng. **21**(10), 1361–1371 (2008)

Jiang, L., Zhang, L., Li, C., Wu, J.: A correlation-based feature weighting filter for Naive Bayes. IEEE Trans. Knowl. Data Eng. **31**(2), 201–213 (2018)

Jiang, L., Zhang, L., Yu, L., Wang, D.: Class-specific attribute weighted naive Bayes. Pattern Recognit. **88**, 321–330 (2019)

John, G. H., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers (2013). arXiv:1302.4964 [cs, stat]

John, G., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345 (1995)

Karvanen, J.: Estimation of quantile mixtures via L-moments and trimmed L-moments. Comput. Statist. Data Anal. **51**(2), 947–959 (2006)

Leisch, F., Dimitriadou, E.: mlbench: Machine Learning Benchmark Problems pp. 1–3 (2021)

Parzen, E.: Nonparametric statistical data modeling. J. Am. Stat. Assoc. **74**(365), 105–121 (1979)

Petersen, K.B., Pedersen, M.S.: The Matrix Cookbook, (2012)

Rayner, G.D., MacGillivray, H.L.: Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. Statist. Comput. **12**(1), 57–75 (2002)

Sankaran, P.G., Nair, N.U., Midhu, N.N.: A new quantile function with applications to reliability analysis. Commun. Statist. Simul. Comput. **45**(2), 566–582 (2016)

Sharma, D., Chakrabarty, T.K.: The quantile-based flattened logistic distribution: some properties and applications. Commun. Statist. Theory Methods **48**(14), 3643–3662 (2019)

Tortora, C., Browne, R.P., ElSherbiny, A., Franczak, B.C., McNicholas, P.D.: Model-based clustering, classification, and discriminant analysis using the generalized hyperbolic distribution: MixGHD R package. J. Stat. Softw. **98**(3), 1–24 (2021)

Tukey, J.W.: Which part of the sample contains the information? Proc. Natl. Acad. Sci. **53**(1), 127 (1965)

Yang, Y., Webb, G.I.: Discretization for naive-Bayes learning: managing discretization bias and variance. Mach. Learn. **74**(1), 39–74 (2009)