Statistics and Computing (2025) 35:48
https://doi.org/10.1007/s11222-025-10566-1

ORIGINAL PAPER

Fuzzy *K*-expectiles clustering

Pierpaolo D'Urso[1] · Livia De Giovanni[2,3] · Lorenzo Federico[2,3] · Vincenzina Vitale[1]

Received: 3 July 2024 / Accepted: 8 January 2025

Abstract
In this paper the Fuzzy K-expectiles clustering model is proposed. The model takes into account the asymmetry inherent in the data distribution, extending its applicability to a broader spectrum of data than the Fuzzy K-means. To achieve this, the Fuzzy K-expectiles clustering model introduces the cluster centroid expectile, and assigns data points based on expectile distances. An adaptive asymmetry parameter is specified for each variable and for each cluster The performance of the adaptive Fuzzy K-expectiles model is compared with other clustering models suggested in the literature. To show the performances of the proposed model three simulation studies and three applications to real datasets are presented.

Keywords Asymmetric quadratic loss · Fuzzy clustering · Expectiles

1 Introduction

When dealing with skewed or asymmetrically distributed data, whose characteristics may not be fully represented by the first two moments, conventional methods may fall short for non-spherical clusters. Addressing within-cluster skewness, Hennig et al. (2019) introduce the K-quantile clustering algorithm based on asymmetric absolute discrepancy that depends on both the quantile level τ and an additional scale/penalty parameter λ, with τ and λ kept uniform across different clusters to simplify computations. A related approach to quantile-based clustering is presented in Zhang et al. (2019).

In response to these challenges, in Wang et al. (2022) the authors propose a novel model, the K-expectile clustering. Drawing inspiration from the concept of K-means, the approach aims to minimize a weighted quadratic loss that considers asymmetry. The authors explore two schemes: one with a pre-specified τ level and another with an adaptive τ that may vary across variables or clusters, allowing for either a fixed cluster shape or a data-driven cluster shape to capture heterogeneity.

In this paper the Fuzzy K-expectile clustering model with an adaptive τ that may vary across variables and clusters is proposed. The model takes into account the asymmetry inherent in the data distribution, extending its applicability to a broader spectrum of data than the Fuzzy K-means.

The paper is organized as follows. In Sect. 2 the Fuzzy K-expectiles clustering model with variable τ is introduced. In section 3 three simulation studies are presented. The Fuzzy K-expectiles model with variable τ is compared with the model with fixed $\tau = 0.50$ in Simulation 1 in the case of asymmetry by variable and cluster and in Simulation 2 in the case of asymmetry by variable. The Fuzzy K-expectiles model with variable τ is compared with the (crisp) K-expectiles model in Simulation 3. In all the simulation studies a comparison with a clustering model based on mixtures of asymmetric variables is also illustrated.

In section 4 the Fuzzy K-expectiles clustering model with variable τ is applied to three real datasets, i.e. wine data, banknotes data and sport data.

✉ Livia De Giovanni
ldegiovanni@luiss.it

Pierpaolo D'Urso
pierpaolo.durso@uniroma1.it

Lorenzo Federico
lfederico@luiss.it

Vincenzina Vitale
vincenzina.vitale@uniroma1.it

[1] Department of Social Sciences and Economics, Sapienza University, Piazzale Aldo Moro 5, 00185 Rome, Italy

[2] Department of Political Sciences, Luiss University, Viale Romania 32, 00197 Rome, Italy

[3] Data Lab, Luiss University, Viale Pola 12, 00198 Rome, Italy

Published online: 10 February 2025

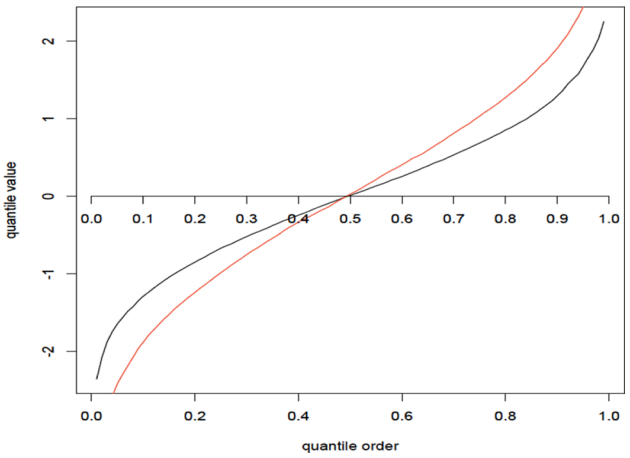

Fig. 1 The quantile (red) and expectile (black) function for a standard Normal distribution

2 Fuzzy *K*-expectiles clustering

2.1 The τ-expectile

The expectile has been introduced by Newey and Powell in the framework of expectile regression, that is, regression on a parameter that generalizes the mean and characterizes the tail behaviour of a distribution (Newey and Powell 1987). The τ-expectile of a real-valued random variable X with distribution function $F(X)$ for $\tau \in (0, 1)$ is the minimizer of the asymmetric quadratic loss:

$$e_\tau(X) = \underset{\epsilon\in\mathbb{R}}{argmin} \int_{\mathbb{R}} (x-\epsilon)^2(\tau\mathcal{I}_{(x>\epsilon)} + (1-\tau)\mathcal{I}_{(x<\epsilon)})dF(X) \tag{1}$$

Compared to quantiles, expectiles have the advantage that for any distribution with finite mean, the expectile is unique for each τ, and the expectile curve is always strictly increasing and continuous. require the existence of a first moment. On the other hand, the expectile is not properly defined for distributions that do not have finite mean (and obviously for distributions on ordered spaces other than real numbers) and it lacks the same intuitive interpretation of the quantile. While saying that x is the α-quantile of a probability distribution X can be written not only as the minimizer of an asymmetric linear loss function, but more explicitly as $\mathbb{P}(X \leq x) = \alpha$, the expectile can only be interpreted as the minimum of the asymmetric loss function in (1). In Fig. 1 the quantile and expectile function for a standard Normal distribution are shown.

The asymptotic properties of the sample expectiles are studied in Holzmann and Klar (2016). In contrast to the mean (which is a special case of the τ-expectile for $\tau = 0.5$), even under the assumption of a finite second moment the sample expectile, which, for a sample $\{x_a : a \in A)$ is defined as

$$\hat{e}_{\tau,A}(X) = \underset{\epsilon\in\mathbb{R}}{argmin} \frac{1}{|A|}\sum_{a\in A}(x_a-\epsilon)^2(\tau\mathcal{I}_{(x_a>\epsilon)} + (1-\tau)\mathcal{I}_{(x_a<\epsilon)}), \tag{2}$$

is proved to be only asymptotically normal if the distribution function F of the random variable X is continuous at its τ-expectile, otherwise, the limit distribution is non-normal.

Introducing the identification function of the expectile as

$$I_\tau(e_\tau, x) = \tau(x-e_\tau)\mathcal{I}_{(x>e_\tau)} + (1-\tau)(e_\tau - x)\mathcal{I}_{(x<e_\tau)}, \tag{3}$$

e_τ can be defined as the solution of the first-order condition:

$$E[I_\tau(e_\tau, x)] = 0, \ x \in \mathbb{R}. \tag{4}$$

Note that, given that the function $I_\tau(e_\tau, x)$ is strictly decreasing in e_τ, such solution is always unique.

In the same way, the empirical τ-expectile $\hat{e}_{\tau,A}$ is defined as the solution of the equation (Holzmann and Klar 2016):

$$\hat{e}_{\tau,A} : \quad \frac{1}{|A|}\sum_{a\in A} I_\tau(e_{\tau,A}, x_a) = 0. \tag{5}$$

We can extend the notion of empirical expectile that we have shown here to a situation where the sample is defined over a fuzzy set $\tilde{A}$, instead of a crisp one. That is, instead of having a binary indicator for each unit of whether that unit is part of the sample or not, we instead have for each unit $a \in \tilde{A}$ a membership function $u_a : \tilde{A} \to [0, 1]$ that indicates the degree of membership of a to the sample. Fixed a parameter $m \geq 1$ that tunes the way in which the memberships are weighted, we define the fuzzy sample expectile $\hat{e}_{\tau,\tilde{A},m}$ as the solution of the following equation:

$$\hat{e}_{\tau,\tilde{A},m} : \quad \frac{\sum_{a\in\tilde{A}} u_a^m I_\tau(e_{\tau,\tilde{A},m}, x_a)}{\sum_{a\in\tilde{A}} u_a^m} = 0. \tag{6}$$

Note that for every m this definition reduces to that of the crisp sample expectile if $u_a \in \{0, 1\}$ for all a, that is, if the set $\tilde{A}$, even if defined within a fuzzy framework, is actually crisp.

An alternative equation for e_τ, which is more useful when trying to invert the expectile function, is

$$\frac{\tau}{1-\tau} = \Big[\int_{-\infty}^{e_\tau}(e_\tau - x)dF(x)\Big]\Big[\int_{e_\tau}^{\infty}(x-e_\tau)dF(x)\Big]^{-1}, \tag{7}$$

which we will extend to the fuzzy sample expectile in Sect. 2.2.

2.2 Fuzzy *K*-expectiles model with variable τ

The Fuzzy K-expectiles model, based on a PAC (Partition Around Centroids) approach, is characterized as follows (in the following the subscript τ is omitted):

$$\begin{aligned}&\text{min: } J_{FKE}(\mathbf{X},\mathbf{U},\mathbf{E},\mathbf{T}) = \\ &\sum_{k=1}^{K}\sum_{i=1}^{I}u_{ik}^{m}\sum_{j=1}^{J}(x_{ij}-e_{jk})^2(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})} \\ &+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})}) \\ &\sum_{k=1}^{K}u_{ik}=1, u_{ik}\geq 0\end{aligned} \tag{8}$$

where x_{jk} indicates the value of the j-th variable for the i-th unit, u_{ik} indicates the membership degree of the i-th unit in the k-th cluster, e_{jk} the centroid expectile of the j-th variable in the k-th cluster. $\tau_{jk} \in (0,1)$ is the parameter of the expectile of the j-th variable in the k-th cluster and $\mathbf{X}$, $\mathbf{U}$, $\mathbf{E}$, $\mathbf{T}$ the related matrices.

Here, it is important to stress that we are looking for optima of the objective function with respect to $\mathbf{U}$, $\mathbf{E}$ for a given value of $\mathbf{T}$, with respect to which we do not require the final result to be a local minimum. This is because by the definition of the objective function, the global minimum with respect to $\mathbf{U}$, $\mathbf{E}$, $\mathbf{T}$ can always be obtained by setting $\tau_{jk} = 0$ and $e_{jk} \leq \min x_{ij}$ for all j,k. Indeed irrespective of the choice of $\mathbf{U}$, in this case $J_{FKE}(\mathbf{X},\mathbf{U},\mathbf{E},\mathbf{T}) = 0$. By solving the constrained quadratic minimization problem shown in equation (8) *via* the Lagrangian multiplier method, we obtain the optimal solutions u_{ik} and e_{jk}. In particular, by considering the following Lagrangian function:

$$\begin{aligned}L_m(\mathbf{U},\mathbf{E},\mathbf{T},\lambda) = &\sum_{k=1}^{K}\sum_{i=1}^{I}u_{ik}^{m}\sum_{j=1}^{J}(x_{ij}-e_{jk})^2(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})} \\ &+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})}) - \lambda\sum_{k=1}^{K}(u_{ik}-1)\end{aligned} \tag{9}$$

and setting the first partial derivatives with respect to u_{ik} and λ equal to zero, it follows:

$$\begin{aligned}\frac{\partial L_m(\mathbf{U},\mathbf{E},\mathbf{T},\lambda)}{\partial u_{ik}} = 0 \Leftrightarrow & mu_{ik}^{m-1}\sum_{j=1}^{J}(x_{ij}-e_{jk})^2 \\ &\times(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})} \\ &+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})}) - \lambda = 0\end{aligned} \tag{10}$$

$$\frac{\partial L_m(\mathbf{U},\mathbf{E},\mathbf{T},\lambda)}{\lambda} = 0 \Leftrightarrow \sum_{k=1}^{K}u_{ik} - 1 = 0. \tag{11}$$

From Equation (10), we obtain:

$$\begin{aligned}u_{ik} = &\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} \\ &\times\frac{1}{\sum_{j=1}^{J}(x_{ij}-e_{jk})^2(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})}+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})})^{\frac{1}{m-1}}}\end{aligned} \tag{12}$$

By considering Equation (11):

$$\begin{aligned}1 = \sum_{k=1}^{K}u_{ik} = &\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}}\sum_{k=1}^{K} \\ &\times\frac{1}{\sum_{j=1}^{J}(x_{ij}-e_{jk})^2(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})}+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})})^{\frac{1}{m-1}}}\end{aligned} \tag{13}$$

and by replacing Equation (13) in Equation (12), it follows:

$$u_{ik} = \frac{\frac{1}{\sum_{j=1}^{J}(x_{ij}-e_{jk})^2(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})}+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})})^{\frac{1}{m-1}}}}{\sum_{k=1}^{K}\frac{1}{\sum_{j=1}^{J}(x_{ij}-e_{jk})^2(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})}+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})})^{\frac{1}{m-1}}}} \tag{14}$$

Then setting the first derivative of the Lagrangian function with respect to e_{jk} equal to 0 it follows:

$$\begin{aligned}\frac{\partial L_m(\mathbf{U},\mathbf{E},\mathbf{T},\lambda)}{e_{jk}} = 0 \Leftrightarrow &\sum_{i=1}^{I}u_{ik}^{m}(x_{ij}-e_{jk})(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})} \\ &+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})}) = 0\end{aligned} \tag{15}$$

that leads to:

$$e_{jk} = \frac{\sum_{i=1}^{I}u_{ik}^{m}x_{ij}(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})}+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})})}{\sum_{i=1}^{I}u_{ik}^{m}(\tau_{jk}\mathcal{I}_{(x_{ij}>e_{jk})})+(1-\tau_{jk})\mathcal{I}_{(x_{ij}<e_{jk})})} \tag{16}$$

The solution to equation (16) is obtained recursively.

The optimal value of τ is obtained by the following recursive formula:

$$\frac{\tau_{jk}}{1-\tau_{jk}} = \frac{\sum_{i=1}^{I}u_{ik}^{m}\mathcal{I}_{(x_{ij}>e_{jk})}\sum_{i=1}^{I}(e_{jk}-x_{ij})\mathcal{I}_{(x_{ij}<e_{jk})}}{\sum_{i=1}^{I}u_{ik}^{m}\mathcal{I}_{(x_{ij}<e_{jk})}\sum_{i=1}^{I}(x_{ij}-e_{jk})\mathcal{I}_{(x_{ij}>e_{jk})}} \tag{17}$$

Equation (17) is derived from equation (7).

Algorithm 1 Fuzzy K-expectiles algorithm

1: Fix $K, max.iter, conv$ and generate randomly the degree matrix $\mathbf{U}$;
2: Set $iter = 0$;
3: Set $\tau_{jk} = 0.5$ for all j, k.
4: Pick K initial J-dimensional centroids: $\mathbf{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$;
5: **repeat**
6: Set $\mathbf{U_{OLD}} = \mathbf{U}$, $\mathbf{T_{OLD}} = \mathbf{T}$
7: Update iteratively τ_{jk} and e_{jk} $j = 1, \ldots, J; k = 1, \ldots, K$ by using (17) until $\|\tau_{jk} - \tau_{jkOLD}\|_1 < conv$;
8: Compute $\mathbf{u}_i$ $i = 1, \ldots, I$ by using (14);
9: Compute $\mathbf{e}_k$ $k = 1, \ldots, K$ by solving (16);
10: $iter \leftarrow iter_{OLD} + 1$;
11: **until** $\|\mathbf{U}_{OLD} - \mathbf{U}\|_1 + \|\mathbf{T}_{OLD} - \mathbf{T}\|_1 < conv$ or $iter = max.iter$

The fuzzy clustering procedure is illustrated in Algorithm 1.

Some remarks on the proposed methods and algorithms are in order.

Remark 1 Convergence. The local and global convergence results for the objective function (8), and the rate of convergence, are discussed in Hathaway and Bezdek (1986), Hathaway and Bezdek (1988).

Following Bezdek (2013), the membership matrix $\mathbf{U}$ can be initialized randomly. However, as remarked in Coppi et al. (2010), it is recognized that fuzzy K-means clustering algorithms present a minor tendency of hitting local optima with respect to their traditional counterparts (see e.g., Bezdek et al. (1999)). In Algorithm 1 we make the stopping condition dependent on both $\mathbf{U}$ and $\mathbf{T}$ stabilizing at a given value. This is necessary even if we are only requiring the solution to be a minimum in $\mathbf{U}$ for the given $\mathbf{T}$ and not a global minimum over both (which would be trivially achieved choosing $\mathbf{U} = \underline{0}$ and all values in $\mathbf{E}$ smaller than the minimum of the corresponding variable). This way we ensure that the final values in $\mathbf{U}$ are computed using values of $\mathbf{T}$ that are almost identical to those used to compute $\mathbf{U}_{OLD}$ and thus they are a local minimum for the objective function for those fixed values of $\mathbf{T}$. Given that $\partial L_m / \partial u_{ik}$ from (15) has continuous and bounded derivatives with respect to $\mathbf{T}$ we can know that the position of a local minimum in u_{ik} is almost always continuous for small perturbations in $\mathbf{T}$.

Remark 2 The fuzziness parameter. The fuzziness parameter m, chosen in advance, plays an important role. A discussion on possible procedures for selecting m can be found in D'Urso (2015). As in the traditional PAC algorithms, it is required that $m \geq 1$ and a higher value of m increases the fuzziness of the partition, with $m = 1$ corresponding to the crisp clustering algorithm. We thus follow the same heuristic guidelines regarding the best choice of m as for the PAC approach. Pal and Bezdek (1995) gave some estimates suggesting that the level of fuzziness should be between 1.5 and 2.5. In this paper, m = 1.5 is used.

Remark 3 Determining the optimal number of clusters.

The Fuzzy Silhouette (FS) index (Campello and Hruschka 2006) is a popular measure computed as the weighted average of the individual silhouette widths, s_i and is defined as follows:

$$FS = \frac{\sum_{i=1}^{I}(u_{ip} - u_{iq})^{\alpha} \cdot s_i}{\sum_{i=1}^{I}(u_{ip} - u_{iq})^{\alpha}}, \qquad s_i = \frac{(b_i - a_i)}{\max\{b_i, a_i\}} \tag{18}$$

where p and q are the first and second best clusters (accordingly to the membership degree), respectively, with which the i-th unit is associated. Here, a_i is the average distance between the i-th unit and the units belonging to the cluster p (p = 1,...,C) with i being associated with the cluster with the highest membership degree; b_i is the minimum (over clusters) average distance of the i-th unit to all units belonging to the cluster q with $q \neq p$; $(u_{ip} - u_{iq})^{\alpha}$ is the weight of each s_i calculated upon the fuzzy partition matrix $\mathbf{U} = \{u_{ic}; i = 1, \ldots, I, \ c = 1, \ldots, C\}$, $\alpha \geq 0$ is an optional user defined weighting coefficient. The higher the value of FS, the better the assignment of the units to the clusters simultaneously obtaining the minimisation of the intra-cluster distance and the maximisation of the inter-cluster distance. The properties of the FS as a validity index are described in Rousseeuw (1987).

In the computation of the distances for the Fuzzy K-expectiles clustering distance in (1) is used.

Remark 4 Comparison of partitions.

The Fuzzy Rand Index (FRI) (Hüllermeier et al. (2012), Anderson et al. (2010)) is adopted to compare different partitions and/or to compare a given partition with a reference one. The FRI ranges from 0 (total disagreement) to 1 (complete agreement) and is a fuzzy extension of the Rand index (RI), Campello (2007), based on agreements and disagreements in two partitions. The RI is based on the contingency matrix associated with two crisp partitions P and Q, defined as $R = P^T Q$. If P consists of k clusters and Q consists of l clusters, then $R = [r_{i,j}]$ is a $k \times l$-matrix. In the non-fuzzy case, the entry $r_{i,j}$ corresponds to the number of objects that belong to the i-th cluster in P and to the j-th cluster in Q. In the case of fuzzy partitions P and Q are the matrices $units \times clusters$ of the fuzzy memberships ranging in [0, 1], and summing up 1 by row (Anderson et al. (2010)).

Remark 5 Computational aspects: The time complexity of the proposed FKE Algorithm 1 is evaluated assuming unit cost for all operations. The time complexity depends on three parameters, the number of data units I, the number of variables (scales) J and the number of clusters C. We consider the complexity for each algorithm iteration. The complexity of Step 6 is $O(IC)$, that of Steps 8 and 9 is $O(ICJ)$, and the most complex portion of the algorithm is Step 7 where every iteration of the inner cycle has complexity $O(ICJ)$. The

Table 1 Beta distributions of the clusters

	Beta j=1	Mean	γ	*Beta* j=2	Mean	γ
Cluster 1	$Beta(2, 6)$+0.65	0.9000	0.69	$Beta(6, 2)$+0.35	1.1000	− 0.69
Cluster 2	$Beta(2, 12)$+0.35	0.4929	0.99	$Beta(12, 2)$+0.65	1.5100	− 0.99
Cluster 3	$Beta(2, 18)$	0.1000	1.11	$Beta(18, 2)$+1.00	1.9000	− 1.11

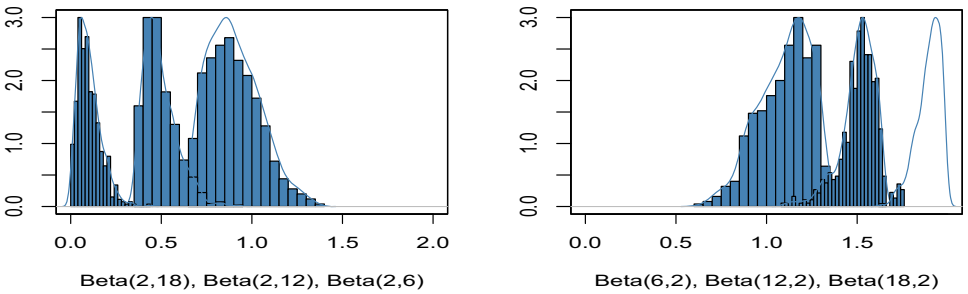

Fig. 2 Beta distributions of the clusters ($j = 1$, left; $j = 2$, right)

complexity of Step 7 can be bounded by setting a fixed maximum number of iterations after which the cycle is stopped and Steps 8 and 9 are performed anyway to prevent the algorithm from getting stuck in specific pathological configurations.

3 Simulation studies

The simulation studies intend to show the performances of the Fuzzy K-expectiles model in clustering data characterized by asymmetry by variable and by cluster and variable. The Fuzzy K-expectiles model with variable τ is compared with the model with fixed $\tau = 0.50$ in Simulation 1 in the case of asymmetry by variable and cluster and in Simulation 2 in the case of asymmetry by variable. The Fuzzy K-expectiles model with variable τ is compared with the (crisp) K-expectiles model in Simulation 3. In the three simulations a comparison with a clustering model based on mixtures of skewed factor analyzers (Parsimonious Mixtures of Skew-t Factor Analyzers PMSTFA) is also presented Murray et al. (2014). The alternating expectation-conditional maximization algorithm is used for model parameter estimation and the Bayesian information criterion is used for model selection. The factor analysis model assumes that a p-dimensional vector of observed variables can be modeled by a q-dimensional vector of latent factors. The model-based clustering employs finite mixture models to estimate the group memberships of a given set of unlabeled observations.

Three clusters ($K = 3$) and two variables ($J = 2$) were considered. The value of the fuzziness parameter m was set to 1.5. Each simulation was repeated 100 times.

Table 2 Adjusted Rand index and Fuzzy Silhouette index

	ARI	*FS*
Variable τ	0.9517	0.8125
Fixed τ=0.50	0.9436	0.8027
PMSTFA clustering	0.8913	0.7967

3.1 Simulation study 1

In Simulation study 1 the Fuzzy K-expectiles model with variable τ is compared with the model with fixed $\tau = 0.50$ in the case of asymmetry by variable and cluster.

Three shifted $Beta$ distributions with increasing asymmetry were generated for each variable. For $j = 1$, $X_{1,j} \sim Beta(2, 6)$ shifted by 0.65, $X_{2,j} \sim Beta(2, 12)$ shifted by 0.35 and $X_{3,j} \sim Beta(2, 18)$. For $j = 2$, $X_{1,j} \sim Beta(6, 2)$ shifted by 0.35, $X_{2,j} \sim Beta(12, 2)$ shifted by 0.65 and $X_{3,j} \sim Beta(18, 2)$ shifted by 1.0 (Table 1, Fig. 2). The moments of the random variable $X \sim Beta(\alpha, \beta)$ $(\alpha, \beta > 0)$ depend on the parameters α, β by the relations $E[X] = \frac{\alpha}{\alpha+\beta}$, $Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, $\gamma = 2\frac{(\beta-\alpha)}{(\alpha+\beta+2)}\sqrt{\frac{(\alpha+\beta+1)}{(\alpha\beta)}}$, where γ is the skewness parameter defined as the expected value of the third power of the standardized variable.

The performances of the two models are presented in Table 2.

The average values of τ_{jk} and the centroid expectiles for each cluster are shown in Tables 3 and 4.

The Fuzzy K-expectiles model with variable τ performs better than the model with fixed τ=0.50 and the model-based PMSTFA clustering either in terms of clustering accuracy or of compactness of the clusters, as measured by the indexes ARI and FS, respectively.

Table 3 Estimated τ and centroids expectiles

	τ_{jk} j=1	Expectile	τ_{jk} j=2	Expectile
Cluster 1	0.3743	0.8833	0.6125	1.1136
Cluster 2	0.3429	0.4692	0.6462	1.5314
Cluster 3	0.3410	0.0822	0.6532	1.9166

Table 4 Estimated centroids expectiles, fixed τ =0.50

	τ_{jk} j=1	Expectile	τ_{jk} j=2	Expectile
Cluster 1	0.5000	0.9030	0.5000	1.0999
Cluster 2	0.5000	0.4922	0.5000	1.5081
Cluster 3	0.5000	0.0986	0.5000	1.9009

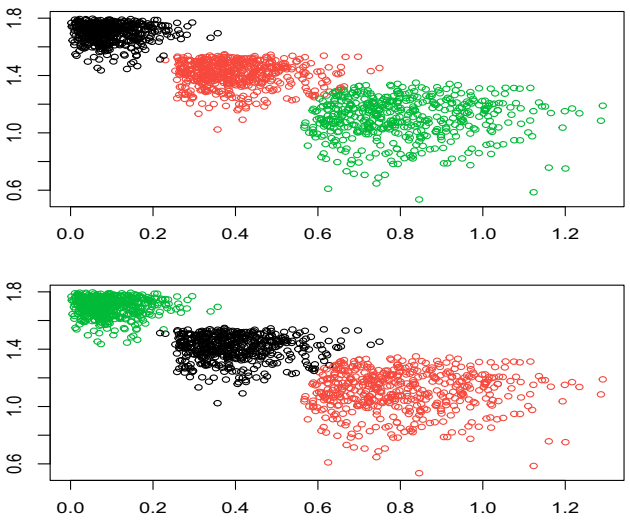

Fig. 3 Fuzzy K-expectiles clustering variable τ (top), fixed τ=0.50 (bottom)

As expected, the values of the estimated τ are smaller than 0.50 and decreasing for increasing positive asymmetry in the case of $j = 1$, are greater than 0.50 and increasing for increasing negative asymmetry in the case of $j = 2$. The Fuzzy K-expectiles model with variable τ tunes the value of τ by variable and cluster.

The values of the expectiles with variable τ are smaller than the values with fixed $\tau = 0.50$ in case of positive asymmetry (j=1); greater in case of negative asymmetry (j=2).

The model with fixed $\tau = 0.50$ behaves like a K-means.

The better performances of the Fuzzy K-expectiles with variable τ are also shown in Fig. 3.

The average error has been computed between expectiles computed according to formula (5) and centroids expectiles computed by the model (formula 16). They are illustrated in Table 5.

Table 5 Estimated error between expectiles and centroids expectiles

		j=1	j=2
Variable τ	Cluster 1	0.0198	− 0.0143
	Cluster 2	− 0.0017	0.0031
	Cluster 3	− 0.0045	0.0009

3.2 Simulation study 2

In Simulation study 2 the Fuzzy K-expectiles model with variable τ is compared with the model with fixed $\tau = 0.50$ in the case of asymmetry by variable.

Three shifted $Beta$ distributions with same asymmetry were generated for each variable. In a first scenario, for $j = 1$, $X_{1,j} \sim Beta(2, 6)$, $X_{2,j} \sim Beta(2, 6)$ shifted by 0.30 and $X_{3,j} \sim Beta(2, 6)$ shifted by 0.60 and for $j = 2$, $X_{1,j} \sim Beta(6, 2)$, $X_{2,j} \sim Beta(6, 2)$ shifted by 0.30 and $X_{2,j} \sim Beta(6, 2)$ shifted by 0.60 were generated.

In a second scenario, for $j = 1$, $X_{1,j} \sim Beta(2, 12)$, $X_{2,j} \sim Beta(2, 12)$ shifted by 0.30 and $X_{3,j} \sim Beta(2, 12)$ shifted by 0.60 and for $j = 2$, $X_{1,j} \sim Beta(12, 2)$, $X_{2,j} \sim Beta(12, 2)$ shifted by 0.30 and $X_{2,j} \sim Beta(12, 2)$ shifted by 0.60 were generated.

In a third scenario, for $j = 1$, $X_{1,j} \sim Beta(2, 18)$, $X_{2,j} \sim Beta(2, 18)$ shifted by 0.30 and $X_{3,j} \sim Beta(2, 18)$ shifted by 0.60 and for $j = 2$, $X_{1,j} \sim Beta(18, 2)$, $X_{2,j} \sim Beta(18, 2)$ shifted by 0.30 and $X_{2,j} \sim Beta(18, 2)$ shifted by 0.60 were generated (Table 6 and Fig. 4).

The performances of the two models are presented in Table 7.

The average values of τ_{jk} and the centroid expectiles for each cluster are shown in Tables 8 and 9.

The Fuzzy K-expectiles model with variable τ performs better than the model with fixed τ=0.5 and the model-based PMSTFA clustering either in terms of clustering accuracy or of compactness of the clusters, as measured by the indexes ARI and FS, respectively (Table 7).

As expected, the values of the estimated τ are smaller than 0.50 and equal in the three clusters in the case of positive asymmetry ($j = 1$); are greater than 0.5 and equal in the three clusters in the case of negative asymmetry ($j = 2$). The Fuzzy K-expectiles model with variable τ tunes the value of τ by variable.

The model with fixed $\tau = 0.50$ behaves like a K-means.

3.3 Simulation study 3

In Simulation study 3 the Fuzzy K-expectiles model with variable τ is compared with the (crisp) K-expectiles model. Three shifted $Skew\ Normal$ distributions were generated for each variable.

Table 6 Beta distributions of the clusters

	Beta j=1	Mean	γ	Beta j=2	Mean	γ
Cluster 1	$Beta(2, 6)$	0.2500	0,69	$Beta(6, 2)$	0.7500	− 0.69
Cluster 2	$Beta(2, 6)$+0.30	0.5500	0,69	$Beta(6, 2)$+0.30	1.0500	− 0.69
Cluster 3	$Beta(2, 6)$+0.60	0.8500	0,69	$Beta(6, 2)$+0.60	1.3500	− 0.69
Cluster 1	$Beta(2, 12)$	0.1429	0,99	$Beta(12, 2)$	0.8571	− 0.99
Cluster 2	$Beta(2, 12) + 0.30$	0.4429	0,99	$Beta(12, 2)$+0.30	1.1571	− 0.99
Cluster 3	$Beta(2, 12) + 0.60$	0.7429	0,99	$Beta(12, 2)$+0.60	1.4571	− 0.99
Cluster 1	$Beta(2, 18)$	0.1000	1.11	$Beta(18, 2)$	1.1000	− 1.11
Cluster 2	$Beta(2, 18) + 0.30$	0.4000	1.11	$Beta(18, 2)$+0.30	1.4000	− 1.11
Cluster 3	$Beta(2, 18) + 0.60$	0.7000	1.11	$Beta(18, 2)$+0.60	1.7000	− 1.11

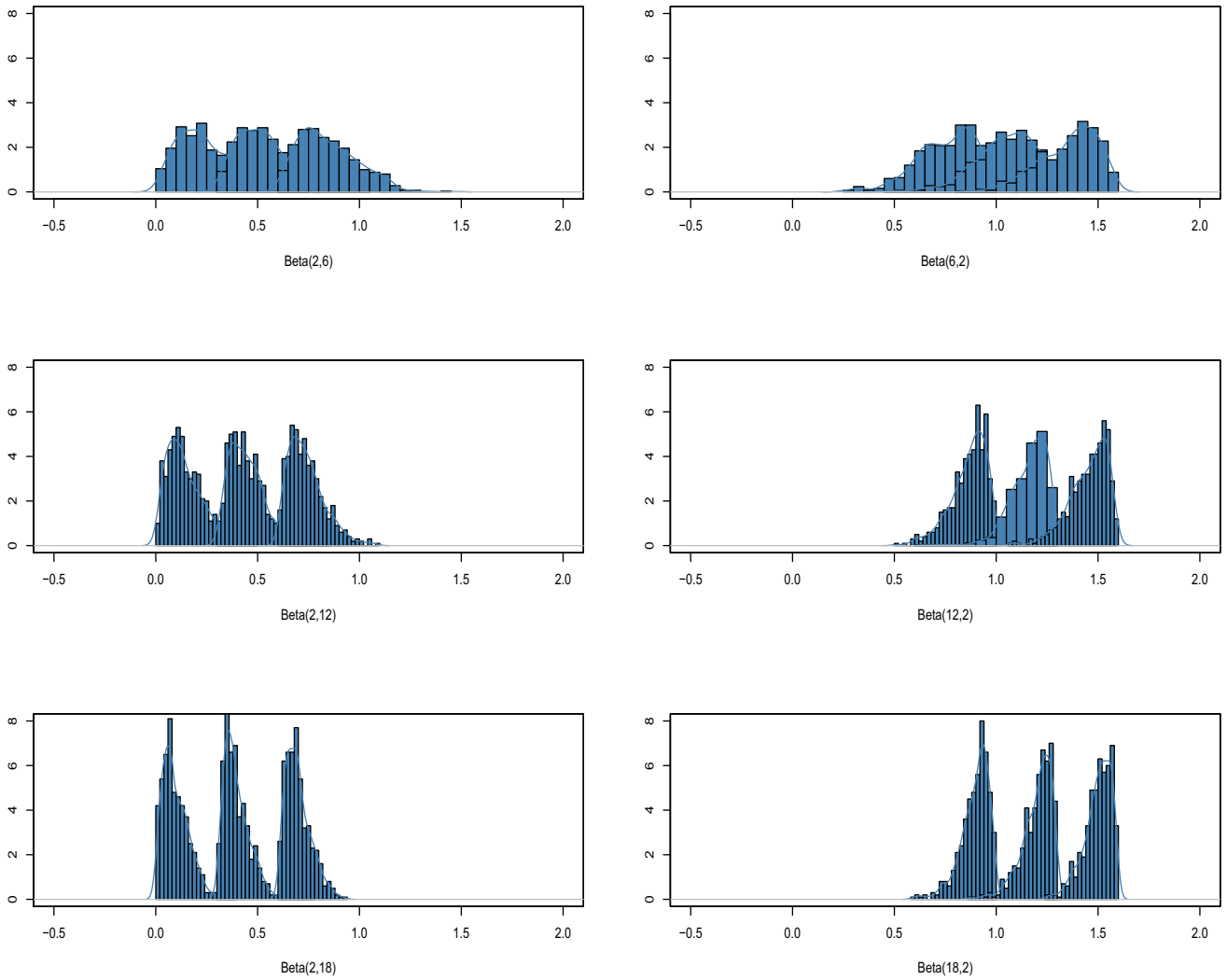

Fig. 4 Beta distributions of the clusters (($j = 1$, left; $j = 2$, right). From top to bottom scenario 1, scenario 2, scenario 3

Table 7 Adjusted Rand Index and Fuzzy Silhouette index

	Scenario 1		Scenario 2		Scenario 3	
	ARI	*FS*	*ARI*	*FS*	*ARI*	*FS*
Variable τ	0.6485	0.6876	0.8980	0.8304	0.9661	0.9023
Fixed τ=0.5	0.6480	0.6775	0.8974	0.8136	0.9660	0.8922
PMSTFA clustering	0.6299	0.6799	0.8799	0.8141	0.9545	0.8599

Table 8 Estimated τ and centroids expectiles

	τ_{jk} j=1	Expectile	τ_{jk} j=2	Expectile
	$Beta(2, 6)$		$Beta(6, 2)$	
Cluster 1	0.4090	0.2090	0.5634	0.7525
Cluster 2	0.4123	0.5222	0.5930	1.0669
Cluster 3	0.4194	0.8422	0.5970	1.3946
	$Beta(2, 12)$		$Beta(12, 2)$	
Cluster 1	0.3830	0.1206	0.6400	0.8448
Cluster 2	0.3641	0.4223	0.6158	1.1230
Cluster 3	0.3805	0.7288	0.6376	1.3978
	$Beta(2, 18)$		$Beta(18, 2)$	
Cluster 1	0.3531	0.0836	0.6252	0.9118
Cluster 2	0.3914	0.3859	0.6336	1.2136
Cluster 3	0.3604	0.6901	0.6358	1.5150

Table 9 Estimated centroids expectiles, fixed τ =0.5000

	τ_{jk} j=1	Expectile	τ_{jk} j=2	Expectile
	$Beta(2, 6)$		$Beta(6, 2)$	
cluster 1	0.5000	0.2356	0.5000	0.7387
cluster 2	0.5000	0.5418	0.5000	1.0474
cluster 3	0.5000	0.8254	0.5000	1.3672
	$Beta(2, 12)$		$Beta(18, 2)$	
cluster 1	0.5000	0.1413	0.5000	0.8578
cluster 2	0.5000	0.4393	0.5000	1.1588
cluster 3	0.5000	0.7434	0.5000	1.4578
	$Beta(2, 18)$		$Beta(18, 2)$	
cluster 1	0.5000	0.0987	0.5000	0.8996
cluster 2	0.5000	0.3995	0.5000	1.2001
cluster 3	0.5000	0.7011	0.5000	1.5027

For $j = 1$, $X_{1,j} \sim SN(\frac{\delta}{\sqrt{1-\delta^2}})$, $X_{2,j} \sim SN(\frac{\delta}{\sqrt{1-\delta^2}})$ shifted by 2 and $X_{3,j} \sim SN(\frac{\delta}{\sqrt{1-\delta^2}})$ shifted by 4, $\delta = 0.99$. For $j = 2$, $X_{1,j} \sim SN(\frac{\delta}{\sqrt{1-\delta^2}})$, $X_{2,j} \sim SN(\frac{\delta}{\sqrt{1-\delta^2}})$ shifted by -2 and $X_{3,j} \sim SN(\frac{-\delta}{\sqrt{1-\delta^2}})$ shifted by -4, $\delta = -0.99$ (Table 10, Fig. 5).

The parameter δ satisfies $|\delta| < 1$. The parameters of the random variable $X \sim SN(\eta)$ distribution, $\eta = \frac{\delta}{\sqrt{1-\delta^2}}$, depend on the parameter δ by the relations $E[X] = b\delta$; $Var([X] = 1 - (b\delta)^2$, $\gamma = \frac{4-\pi}{2} sign(\eta) \frac{E(X]^3}{Var[X]^{\frac{3}{2}}}$, where γ is the skewness parameter. The choice of δ corresponds to the maximum possible asymmetry, either positive and negative, equal to about 0.995.

The performances of the two models are presented in Table 11.

The average values of τ_{jk} and the centroid expectiles for each cluster are shown in Tables 12 and 13.

The ternary plot is shown in Fig. 6. The ternary plot visually illustrates how each unit is positioned within the triangle, based on the memberships indicating its fuzzy belonging to each of the three clusters. As expected, fuzzy memberships are observed between cluster 2 and cluster 1, cluster 2 and cluster 3.

The Fuzzy K-expectiles model with variable τ performs better than the crisp K-expectiles model with variable τ and the model-based PMSTFA clustering in terms of compactness of the clusters, as measured by the index FS, being comparable in terms of accuracy, as measured by the index ARI.

As expected, the values of the estimated τ are smaller than 0.50 and almost equal in the three clusters in the case of positive asymmetry ($j = 1$); are greater than 0.50 and almost equal in the three clusters in the case of negative asymmetry ($j = 2$). The (Fuzzy) K-expectiles model with variable τ tunes the value of τ.

4 Applications

The Fuzzy K-expectiles model with variable τ is used for clustering each of 178 wine specimens belonging to three types of wine - Barolo, Grignolino and Barbera - on the basis of chemical measurements showing skewness, 200 banknotes on the basis of size measures and 202 athletes on the basis of body mass index and body fat.

4.1 Wine data

In section 4 the Fuzzy K-expectiles model with variable τ is used for clustering each of 178 wine specimens belonging to three types of wine.

Table 10 Skew Normal distributions of the clusters ($(j = 1$, left; $j = 2$, right)

	SN j=1	Mean	γ	SN j=2	Mean	γ
Cluster 1	$SN(\frac{0.99}{\sqrt{1-0.99^2}})$	0.7901	0.9209	$SN(\frac{-0.99}{\sqrt{1-0.99^2}})$	− 0.7901	− 0.9209
Cluster 2	$SN(\frac{0.99}{\sqrt{1-0.99^2}})$+2.0	2.7901	0.9209	$SN(\frac{-0.99}{\sqrt{1-0.99^2}})$− 2.0	− 2.7901	− 0.9209
Cluster 3	$SN(\frac{0.99}{\sqrt{1-0.99^2}})$+4.0	4.7901	0.9209	$SN(\frac{-0.99}{\sqrt{1-0.99^2}})$− 4.0	− 4.7901	− 0.9209

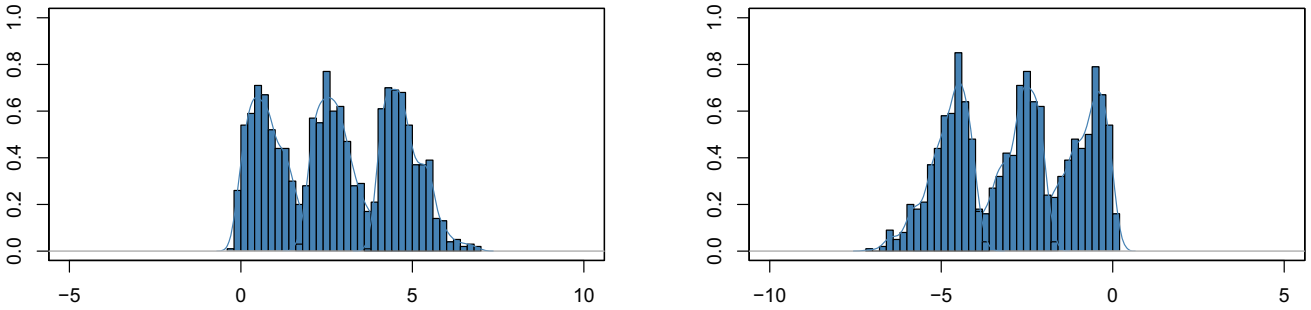

Fig. 5 *SkewNormal* distributions of the clusters

Table 11 Adjusted Rand Index and Fuzzy Silhouette index

	ARI	*FS*
Fuzzy variable τ	0.9378	0.8218
Crisp variable τ	0.9361	0.8011
PMSTFA clustering	0.9015	0.8025

Table 12 Estimated τ and centroids expectiles—Fuzzy K-expectiles

	τ_{jk} j=1	Expectile	τ_{jk} j=2	Expectile
Cluster 1	0.3688	0.6197	0.6297	− 0.6270
Cluster 2	0.3982	2.6342	0.5940	− 2.6526
Cluster 3	0.4075	4.6982	0.5964	− 4.6675

Table 13 Estimated τ and centroids expectiles—K-expectiles

	τ j=1	Expectile	τ j=2	Expectile
Cluster 1	0.4388	0.6742	0.6118	− 0.6684
Cluster 2	0.4372	2.6637	0.6124	− 2.6645
Cluster 3	0.4286	4.6629	0.6130	− 4.6929

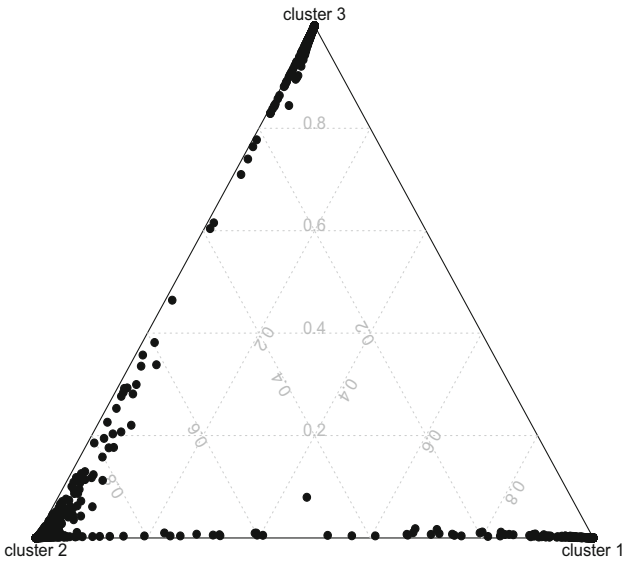

Fig. 6 Ternary plot - simulation 3

The data represent 28 chemical measurements on each of 178 wine specimens belonging to three types of wine - Barolo, Grignolino and Barbera - produced in the Piedmont region of Italy. The data have been presented and examined in Forina et al. (1986); Forina (2008). The data were downoaded by the package Azzalini (2023). Among the 28 variables, eight variables were selected for the analysis for their discriminant ability, according to previous studies (Forina et al. 1986 and related references). The selected variables are the following: Alcohol, Phenols, Flavanoids, Color intensity, Hue, OD280/OD315 of diluted wines (OD_{dw}), OD280/OD315 of flavanoids (OD_{fl}), Proline. All variables but the label class are continuous. The data was used also in Azzalini (2013).

The descriptive statistics of the variables are presented in Table 14 for a total of $I = 178$ wines, 59 Barolo, 71 Grignolino, 48 Barbera, considering multiple quantiles for a comprehensive understanding of heterogeneity. The histograms of the variables are presented in Fig. 7. In Fig. 8 the histogram of the variable OD_{dw} is presented separately for

Table 14 Descriptive statistics of the variables of the 3 wines

	Alchol	Phenols	Flavanoids	Colour	Hue	OD_{dw}	OD_{fl}	Proline
Minimum	11.030	0.9800	0.3400	1.2800	0.4800	1.2700	1.1200	278.00
1st qu.	12.360	1.7420	1.2050	3.2200	0.7825	1.9380	2.2300	500.50
2st qu.	13.050	2.3550	2.1350	4.6900	0.9650	2.7000	3.1300	673.50
Mean	13.000	2.2950	2.0290	5.0580	0.9574	2.6120	2.9810	746.90
3st qu.	13.680	2.8000	2.8750	6.2000	1.1200	3.1700	3.6830	985.00
Maximum	14.830	3.8800	5.0800	13.000	1.7100	4.0000	4.7700	1680.00
Skewness (γ)	−0.0510	0.0859	0.0251	0.8612	0.0209	− 0.3047	− 0.3248	0.7613

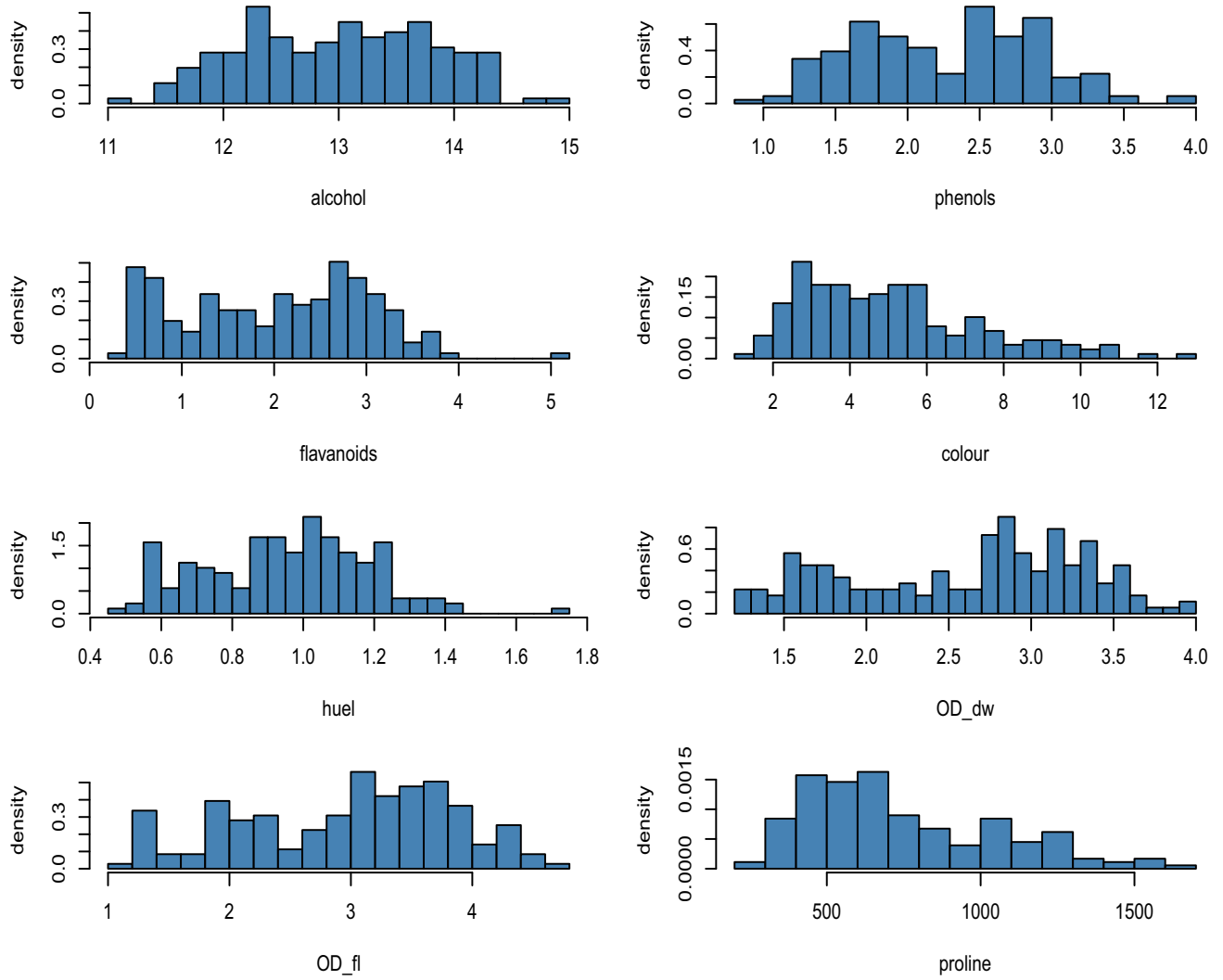

Fig. 7 Histograms of the variables

each type of wine, to show the heterogeneity among clusters. The pairs plot is shown in Fig. 9.

In Table 15 the mean and the skewness is presented separately by type of wine, to show the importance of the adaptive τ that may vary across variables and clusters.

The Fuzzy K-expectile model with variable τ and fixed τ were applied for $K = 2, 3, 4$ to the standardized variables. Three clusters were selected using the Fuzzy Silhouette index in formula (18) (Table 16).

In Table 17 the partition by the Fuzzy K-expectile model with variable τ is related to the three types of wine:

In Table 18 the values of the skewness in the three clusters are shown.

In Table 19 the values of τ_{jk} are shown.

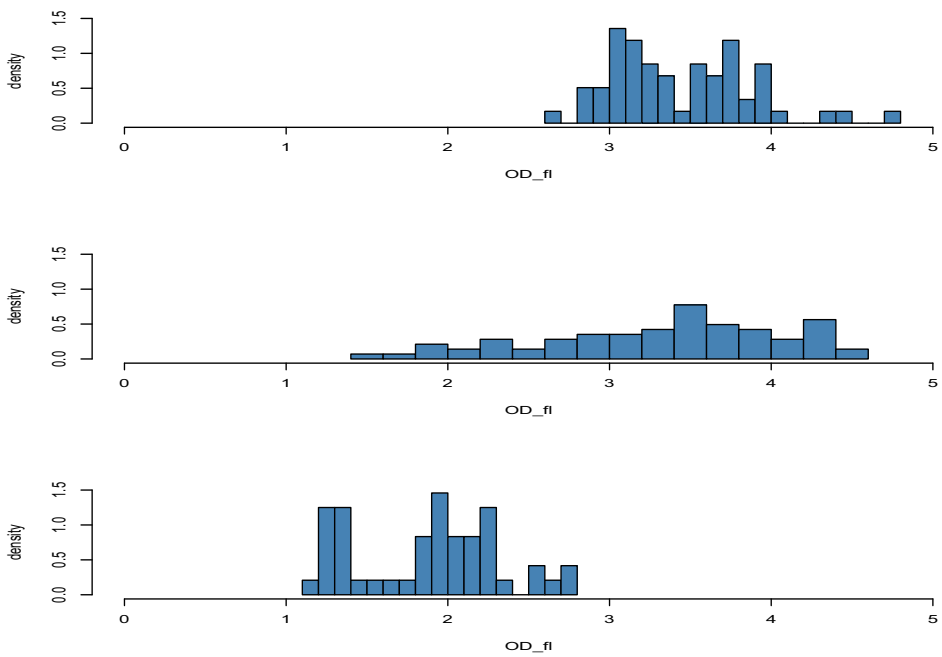

Fig. 8 Histograms of the variable OD_{dw} in the three groups of wines

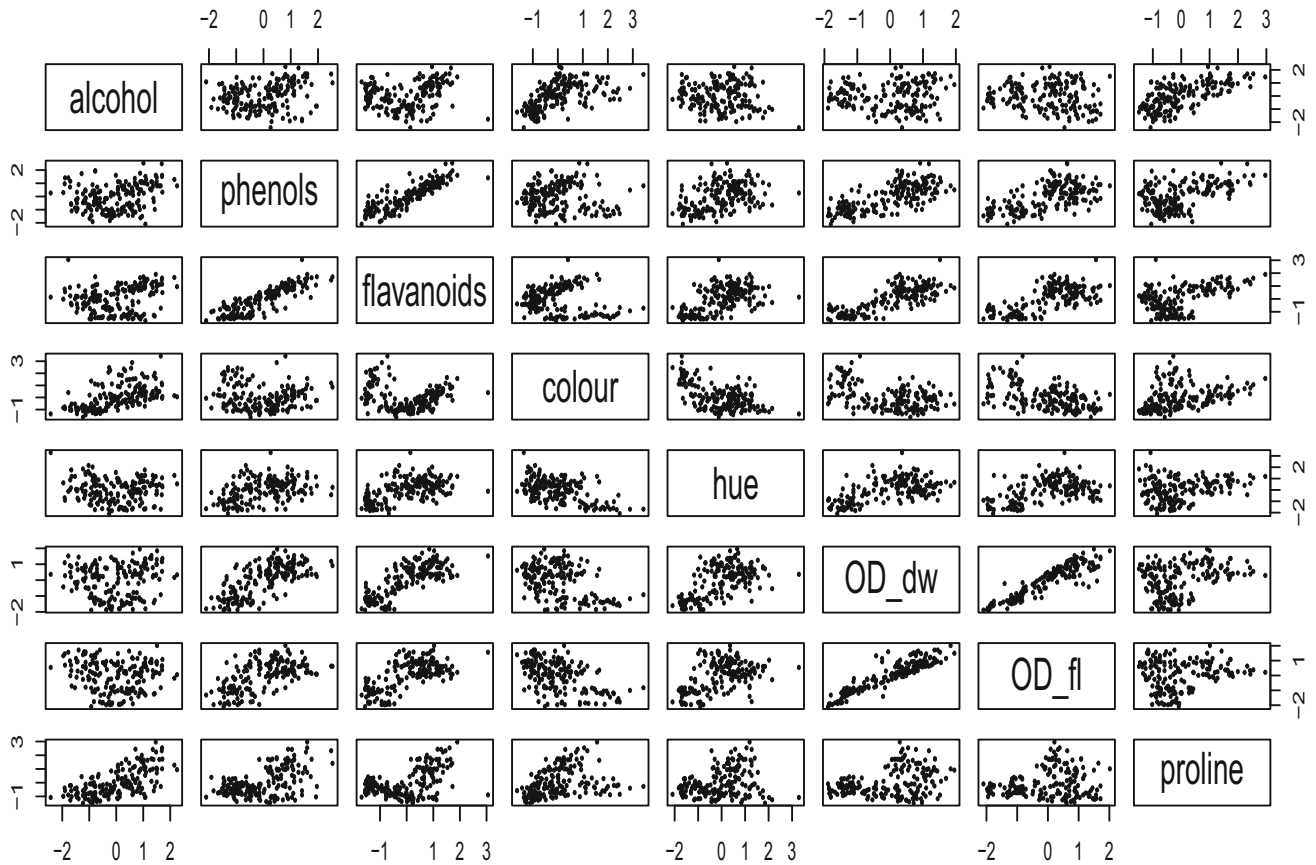

Fig. 9 Pairs plot of the wine variables

Table 15 Mean and skewness of the variables in the 3 groups of wines

	Alchol	Phenols	Flavanoids	Colour	Hue	OD_{dw}	OD_{fl}	Proline
	Mean							
Barolo	13.7447	2.8402	2.9824	5.5283	1.0620	3.1578	3.4485	1115.7119
Grignolino	12.2787	2.2589	2.0808	3.0866	1.0563	2.7854	3.3317	519.5070
Barbera	13.1538	1.6787	0.7815	7.3962	0.6827	1.6835	1.8860	629.8958
	Mean of the scaled variables							
Barolo	0.9166	0.8709	0.9542	0.2028	0.4575	0.7692	0.5272	1.1712
Grignolino	− 0.8892	− 0.0579	0.0516	− 0.8504	0.4324	0.2446	0.3956	− 0.7221
Barbera	0.1886	− 0.9848	− 1.2492	1.0085	− 1.2019	− 1.3072	− 1.2331	− 0.3715
	Skewness (γ)							
Barolo	0.0692	0.7954	0.2614	0.5749	− 0.0835	0.3075	0.6523	0.2238
Grignolino	0.5679	0.3526	1.1484	1.0177	0.4767	− 0.4344	− 0.5119	0.8730
Barbera	0.1469	0.9884	0.9774	0.2919	0.5719	0.6652	− 0.0041	0.3091

Table 16 Fuzzy Silhouette index K=2, 3, 4

K	Fuzzy K-expectile variable τ	Fuzzy K-expectile fixed τ
2	0.6442	0.6030
3	0.6856	0.6360
4	0.5841	0.5420

Table 17 Partition into three clusters

	Cluster 1	Cluster 2	Cluster 3
Barolo	57	2	0
Grignolino	3	61	7
Barbera	0	0	48

Table 18 Skewness in the partition

	Alchol	Phenols	Flavanoids	Colour	Hue	OD_{dw}	OD_{fl}	Proline
Cluster 1	− 0.1702	0.6615	0.2180	0.5216	0.0004	0.3122	0.6755	0.0022
Cluster 2	0.7514	0.3601	1.3285	0.9364	0.5207	− 0.2594	− 0.4854	0.7084
Cluster 3	0.2501	0.8614	0.6745	0.2902	1.0669	0.4722	− 0.0094	0.2289

Table 19 $\tau_{jk}, k = 1, \ldots, 3$, $j = 1, \ldots 8$

	Alchol	Phenols	Flavanoids	Colour	Hue	OD_{dw}	OD_{fl}	Proline
Cluster 1	0.6051	0.5159	0.5187	0.5058	0.5300	0.5314	0.4561	0.5492
Cluster 2	0.5134	0.3874	0.1922	0.5681	0.2932	0.3037	0.5191	0.4840
Cluster 3	0.3958	0.4793	0.4857	0.2972	0.4514	0.6233	0.5998	0.2990

Table 20 Centroids expectiles $e_{jk}, k = 1, \ldots, 3, j = 1, \ldots 8$

	Alchol	Phenols	Flavanoids	Colour	Hue	OD_{dw}	OD_{fl}	Proline
Cluster 1	1.0059	0.9360	1.0025	0.2290	0.4786	0.7890	0.4867	1.1986
Cluster 2	−1.0442	−0.1320	0.0640	−1.0129	0.3714	0.4932	0.7097	−0.8912
Cluster 3	0.1270	−1.0819	−1.3433	0.9517	− 1.2581	− 1.3872	− 1.1906	− 0.4026

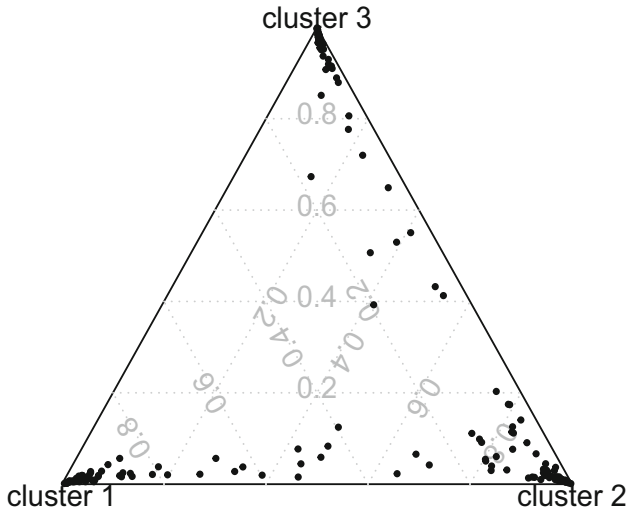

Fig. 10 Ternary plot

In Table 20 the centroids e_{jk} highlight the chemical differences among the three wines. Cluster 1 is characterized by values of all the variables above the mean, in particular alchol, phenols, flavanoids and proline; Cluster 3 by values below the mean with the exception of alchol in the mean and colour above the mean, Cluster 2 by values above and below the mean, with the lowest value of alchol and proline.

The ternary plot is shown in Fig. 10.

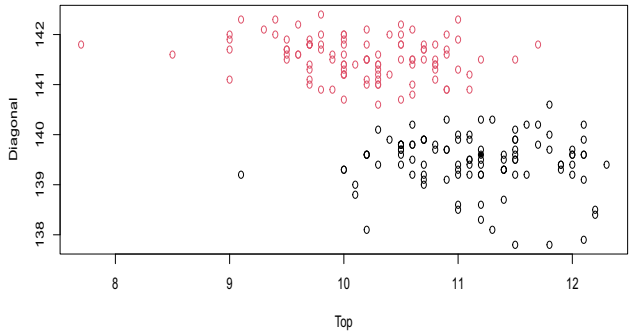

Fig. 12 Clustering of the banknotes (filled black dot banknote 70 in the group 101-200)

The membership degrees and the highest membership cluster are presented in Table 25. In the analysis a wine is considered a member of a cluster if its fuzzy membership to the cluster is a above 0.6 (Maharaj and D'Urso 2011 and related references). Wines that do not reach these thresholds for any cluster are considered as fuzzy units (in italic in Table 25).

The Fuzzy K-expectiles model with variable τ recovers the three types of wine, besides some overlapping between cluster 1 "Barolo" and cluster 2 "Grignolino" and between cluster 2 "Grignolino" and cluster 3 "Barbera". We recall that some wines in previous papers were excluded by the analysis (Forina et al. (1986).

Table 21 Descriptive statistics of the variables of the banknote data

	Length	Left	Right	Bottom	Top	Diagonal
Minimum	213.8	129.0	129.0	7.2	7.7	137.8
1st qu.	214.6	129.9	129.7	8.2	10.1	139.5
2st qu.	214.9	130.2	130.0	9.1	10.6	140.4
Mean	214.9	130.1	130.0	9.4	10.6	140.5
3st qu.	215.1	130.4	130.20	10.6	11.2	141.5
Maximum	216.3	131.0	131.1	12.7	12.3	142.4
Skewness (γ)	0.1877	− 0.1874	0.0386	0.3692	− 0.2266	− 0.1899

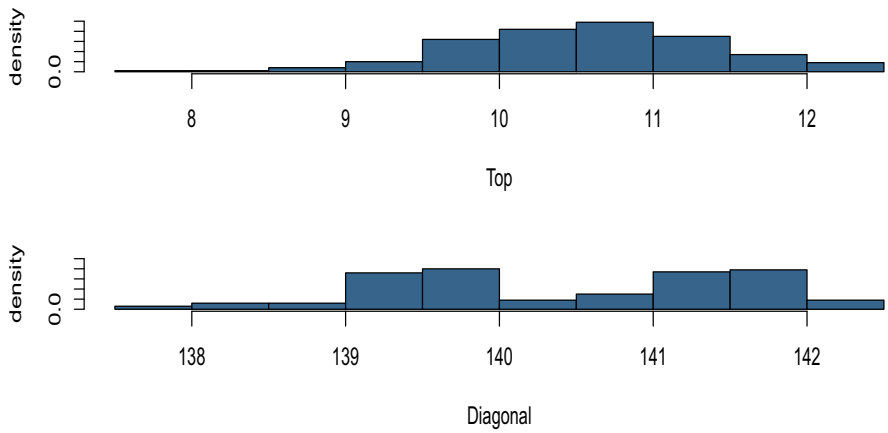

Fig. 11 Histograms of the variables

Table 22 Centroids expectiles (left) and $\tau_{jk}, k = 1, \ldots, 2, j = 1, \ldots 2$ (right)

	Top	Diagonal
Cluster 1	11.1492	139.520
Cluster 2	10.2210	141.609
	Top	Diagonal
Cluster 1	0.5392	0.5858
Cluster 2	0.5236	0.5874

4.2 Banknote data

This data set contains measurements on 200 Swiss banknotes: 100 genuine and 100 counterfeit. The variables are length of bill, width of left edge, width of right edge, bottom margin width and top margin width. All measurements are in millimetres (Flury and Riedwyl 1988). The descriptive statistics of the variables are presented in Table 21 for a total of I =200 banknotes. The selected variables are the top margin width and diagonal length. The histograms of the variables are presented in Fig. 11.

Two groups have been selected by the Fuzzy Silhouette index, of size 99 and 101). Banknote 70 (black filled circle Fig. 12 has highest membership to cluster 2.

In Table 22 the centroids expectile and the values of τ_{jk} are shown. As expected due to the negative asymmetry, the values of τ_{jk} are greater than 0.5.

4.3 Australian Institute of sport data

The data give measurements from high-performance athletes from the Australian Institute of Sport (AIS), for 202 athletes (102 males; 100 females) on 13 variables: Sex 100 Female, 102 Male; Sport, one of BBall basketball; Field; Gym gymnastics; Netball; Rowing; Swim swimming; T400m track, further than 400 m; Tennis; TPSprnt track sprint events; WPolo waterpolo; LBM lean body mass, in kg; Ht height, in cm; Wt, weight, in kg; BMI, body mass index, in kg per metre-squared; SSF, sum of skin folds; PBF, percentage body fat; RBC, red blood cell count and WBC, white blood cell count in 10^{12} per litre; HCT, hematocrit, in percent; HGB, hemoglobin concentration, in grams per decilitre; Ferr, plasma ferritins, in ng per decilitre. (Telford and Cunningham 1991) provide more information on how the data were collected. In Murray et al. (2014), body fat percentage (PBF) and body mass index (BMI) were considered among the 13 variables to compare the true and predicted classification of Sex. The descriptive statistics of the variables are presented in Table 23 for a total of I =202 athletes. The histograms of the variables are presented in Fig. 13.

Table 23 Descriptive statistics of the variables of the AIS data

	Sex (F=1)	PMI	PBF
Minimum		16.7	5.6
1st qu.		21.1	8.5
2st qu.		22.7	11.6
mean	0.98	22.9	13.5
3st qu.		24.5	18.1
Maximum		34.4	35.5
Skewness (γ)		0.9395	0.7539

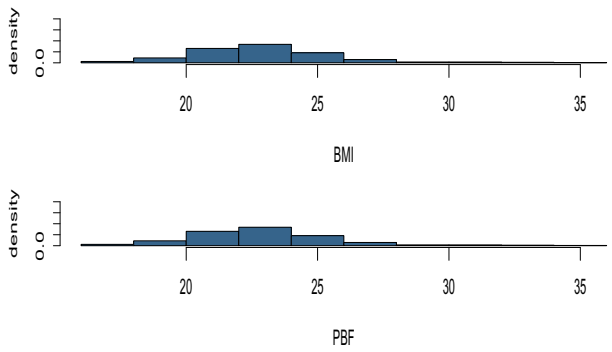

Fig. 13 Histograms of the variables

The two selected groups have size 75 and 127 (Fig. 14, where also the plot of the variables with respect to Sex is shown. The percentage of Females in cluster 1 is 67%, of males in cluster 2 is 92%. The value of the FS is 0.7838 for the Fuzzy K expectile model; 0.7583 for the PMSTFA model.

In Table 24 the centroids expectile and the values of τ_{jk} are shown. As expected due to the positive asymmetry, the values of τ_{jk} are smaller than 0.5.

5 Conclusions

In this paper, the Fuzzy K-expectile clustering model with an adaptive τ that may vary across variables and clusters is proposed. The model takes into account the asymmetry inherent in the data distribution, extending its applicability to a broader spectrum of data than the Fuzzy K-means.Three simulations are presented. The Fuzzy K-expectiles model with variable τ is compared with the model with fixed $\tau = 0.50$ in Simulation 1 in the case of asymmetry by variable and cluster and in Simulation 2 in the case of asymmetry by variable. The Fuzzy K-expectiles model with variable τ is compared with the (crisp) K-expectiles model in Simulation 3. In the three simulations a comparison with a clustering model based on mixtures of skewed factor analyzers is also

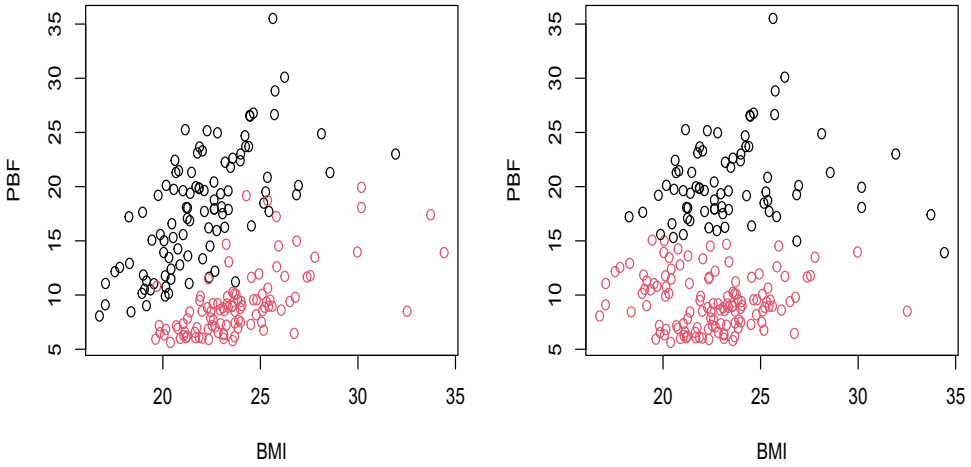

Fig. 14 Classification of the athletes (with respect to clustering in the left plot, with respect to Sex in the right plot)

Table 24 Centroids expectiles (left) and $\tau_{jk}, k = 1, \ldots, 2$, $j = 1, \ldots 2$ (right)

	BMI	PBF
Cluster 1	22.5180	8.8311
Cluster 2	23.0262	19.8987
	BMI	PDF
Cluster 1	0.3785	0.3947
Cluster 2	0.4954	0.3950

presented. In all the simulations the proposed model shows better performances.

In Sect. 4 the Fuzzy K-expectiles model with variable τ is used for clustering each of 178 wine specimens belonging to three types of wine—Barolo, Grignolino and Barbera—on the basis of chemical measurements showing skewness. The Fuzzy K-expectiles model with variable τ recovers the three types of wine. The Fuzzy K-expectiles model with variable τ is able to detect the (negative) asymmetry and to set the value of τ in clustering 200 banknotes on the basis of size measures and detect the (positive) asymmetry and to set the value of τ in clustering 202 athletes on the basis of body mass index and body fat.

Appendix

Table 25 Membership degrees—$K = 3$

	Wine	Cluster 1	Cluster 2	Cluster 3	Cluster		Wine	Cluster 1	Cluster 2	Cluster 3	Cluster
1	Barolo	0.9204	0.0628	0.0167	1	101	Grignolino	0.0242	0.9676	0.0082	2
2	Barolo	0.9348	0.0599	0.0053	1	102	Grignolino	0.0295	0.9258	0.0447	2
3	Barolo	0.9905	0.0079	0.0017	1	103	Grignolino	0.0439	0.9372	0.0189	2
4	Barolo	0.8882	0.0690	0.0428	1	104	Grignolino	0.0090	0.9821	0.0089	2
5	Barolo	0.6474	0.3249	0.0278	1	105	Grignolino	0.0701	0.9143	0.0156	2
6	Barolo	0.9655	0.0218	0.0127	1	106	Grignolino	0.0195	0.9505	0.0300	2
7	Barolo	0.9568	0.0329	0.0103	1	107	Grignolino	0.0148	0.9745	0.0107	2
8	Barolo	0.9722	0.0224	0.0055	1	108	*Grignolino*	0.0458	0.5414	0.4128	2
9	Barolo	0.9498	0.0316	0.0186	1	109	Grignolino	0.0070	0.9867	0.0063	2
10	Barolo	0.9863	0.0102	0.0035	1	110	Grignolino	0.0663	0.9187	0.0149	2
11	Barolo	0.9637	0.0259	0.0104	1	111	Grignolino	0.1183	0.8213	0.0604	2
12	Barolo	0.9079	0.0636	0.0285	1	112	Grignolino	0.0176	0.9664	0.0160	2
13	Barolo	0.9755	0.0174	0.0071	1	113	Grignolino	0.0538	0.8205	0.1256	2
14	Barolo	0.9298	0.0426	0.0276	1	114	Grignolino	0.0231	0.9607	0.0162	2
15	Barolo	0.9313	0.0439	0.0248	1	115	Grignolino	0.0199	0.9724	0.0077	2
16	Barolo	0.9544	0.0315	0.0141	1	116	Grignolino	0.1315	0.7770	0.0915	2
17	Barolo	0.9809	0.0118	0.0073	1	117	Grignolino	0.0073	0.9868	0.0059	2
18	Barolo	0.9578	0.0244	0.0177	1	118	Grignolino	0.0017	0.9969	0.0013	2
19	Barolo	0.8583	0.0852	0.0564	1	119	Grignolino	0.0358	0.3152	0.6490	3
20	Barolo	0.9310	0.0601	0.0089	1	120	Grignolino	0.0141	0.9764	0.0095	2
21	Barolo	0.9507	0.0396	0.0098	1	121	Grignolino	0.1379	0.8224	0.0397	2
22	Barolo	0.3315	0.6456	0.0229	2	122	*Grignolino*	0.3958	0.4796	0.1247	2
23	Barolo	0.8750	0.1088	0.0162	1	123	Grignolino	0.0128	0.9805	0.0068	2
24	*Barolo*	0.5977	0.3827	0.0197	1	124	Grignolino	0.1427	0.8028	0.0544	2
25	Barolo	0.7185	0.2597	0.0218	1	125	Grignolino	0.1439	0.8055	0.0506	2
26	*Barolo*	0.5290	0.4554	0.0156	1	126	Grignolino	0.0375	0.9483	0.0143	2
27	Barolo	0.9365	0.0545	0.0090	1	127	Grignolino	0.1409	0.7480	0.1112	2
28	Barolo	0.7918	0.1702	0.0380	1	128	Grignolino	0.0059	0.9890	0.0051	2
29	Barolo	0.9630	0.0317	0.0053	1	129	Grignolino	0.0068	0.9877	0.0055	2
30	Barolo	0.9635	0.0304	0.0061	1	130	Grignolino	0.0347	0.9347	0.0306	2
31	Barolo	0.9816	0.0129	0.0055	1	131	Barbera	0.0111	0.0368	0.9520	3
32	Barolo	0.9516	0.0326	0.0159	1	132	Barbera	0.0040	0.0130	0.9829	3
33	Barolo	0.8539	0.1225	0.0236	1	133	Barbera	0.0045	0.0128	0.9827	3
34	Barolo	0.9879	0.0094	0.0027	1	134	Barbera	0.0077	0.0260	0.9663	3
35	Barolo	0.8680	0.1071	0.0249	1	135	Barbera	0.0056	0.0184	0.9761	3
36	Barolo	0.9511	0.0436	0.0053	1	136	Barbera	0.0011	0.0031	0.9958	3
37	Barolo	0.8119	0.1668	0.0213	1	137	Barbera	0.0179	0.0627	0.9194	3
38	Barolo	0.6699	0.2736	0.0564	1	138	Barbera	0.0061	0.0304	0.9635	3
39	Barolo	0.6263	0.3359	0.0377	1	139	Barbera	0.0033	0.0092	0.9875	3
40	Barolo	0.9187	0.0622	0.0191	1	140	Barbera	0.0343	0.1594	0.8062	3
41	Barolo	0.9434	0.0447	0.0119	1	141	Barbera	0.0127	0.0574	0.9299	3
42	Barolo	0.8774	0.1040	0.0186	1	142	Barbera	0.0194	0.0435	0.9371	3
43	Barolo	0.9665	0.0253	0.0082	1	143	Barbera	0.0187	0.0685	0.9128	3

Table 25 continued

	Wine	Cluster 1	Cluster 2	Cluster 3	Cluster		Wine	Cluster 1	Cluster 2	Cluster 3	Cluster
44	*Barolo*	0.4368	0.4799	0.0832	2	144	Barbera	0.0501	0.1732	0.7767	3
45	Barolo	0.8538	0.1293	0.0169	1	145	Barbera	0.0095	0.0240	0.9665	3
46	Barolo	0.9752	0.0185	0.0063	1	146	Barbera	0.0118	0.0284	0.9598	3
47	Barolo	0.9867	0.0100	0.0033	1	147	Barbera	0.0141	0.0312	0.9547	3
48	Barolo	0.9867	0.0096	0.0037	1	148	Barbera	0.0008	0.0018	0.9974	3
49	Barolo	0.9834	0.0102	0.0063	1	149	Barbera	0.0020	0.0035	0.9945	3
50	Barolo	0.9513	0.0302	0.0185	1	150	Barbera	0.0054	0.0109	0.9837	3
51	Barolo	0.9530	0.0351	0.0119	1	151	Barbera	0.0051	0.0095	0.9854	3
52	Barolo	0.9755	0.0196	0.0049	1	152	Barbera	0.0160	0.0330	0.9510	3
53	Barolo	0.9357	0.0436	0.0206	1	153	Barbera	0.0062	0.0143	0.9796	3
54	Barolo	0.9868	0.0095	0.0038	1	154	Barbera	0.0100	0.0166	0.9735	3
55	Barolo	0.9889	0.0082	0.0029	1	155	Barbera	0.0014	0.0034	0.9953	3
56	Barolo	0.9943	0.0041	0.0015	1	156	Barbera	0.0010	0.0019	0.9971	3
57	Barolo	0.9793	0.0151	0.0056	1	157	Barbera	0.0061	0.0101	0.9838	3
58	Barolo	0.9864	0.0102	0.0034	1	158	Barbera	0.0059	0.0137	0.9804	3
59	Barolo	0.9509	0.0298	0.0193	1	159	Barbera	0.1749	0.1518	0.6732	3
60	*Grignolino*	0.0520	0.5156	0.4324	2	160	Barbera	0.0661	0.0827	0.8512	3
61	*Grignolino*	0.0791	0.3912	0.5297	3	161	Barbera	0.0081	0.0223	0.9696	3
62	Grignolino	0.0185	0.1021	0.8794	3	162	Barbera	0.0288	0.0635	0.9077	3
63	*Grignolino*	0.1925	0.4148	0.3927	2	163	Barbera	0.0169	0.0937	0.8894	3
64	Grignolino	0.2731	0.6613	0.0656	2	164	Barbera	0.0015	0.0045	0.9939	3
65	Grignolino	0.0590	0.8291	0.1119	2	165	Barbera	0.0065	0.0109	0.9826	3
66	Grignolino	0.1322	0.7692	0.0986	2	166	Barbera	0.0037	0.0086	0.9877	3
67	*Grignolino*	0.5091	0.4463	0.0445	1	167	Barbera	0.0166	0.0281	0.9553	3
68	Grignolino	0.0281	0.9620	0.0100	2	168	Barbera	0.0065	0.0142	0.9793	3
69	*Grignolino*	0.1421	0.3512	0.5067	3	169	Barbera	0.0026	0.0041	0.9933	3
70	Grignolino	0.0399	0.9274	0.0327	2	170	Barbera	0.0025	0.0045	0.9930	3
71	Grignolino	0.0504	0.2296	0.7200	3	171	Barbera	0.0152	0.0748	0.9100	3
72	*Grignolino*	0.4628	0.4783	0.0589	2	172	Barbera	0.0049	0.0119	0.9831	3
73	Grignolino	0.0782	0.8455	0.0764	2	173	Barbera	0.0147	0.0189	0.9664	3
74	Grignolino	0.7827	0.1969	0.0205	1	174	Barbera	0.0014	0.0021	0.9965	3
75	Grignolino	0.2589	0.6991	0.0420	2	175	Barbera	0.0004	0.0007	0.9989	3
76	Grignolino	0.0471	0.7500	0.2030	2	176	Barbera	0.0085	0.0130	0.9785	3
77	Grignolino	0.0610	0.8247	0.1143	2	177	Barbera	0.0038	0.0061	0.9901	3
78	Grignolino	0.0139	0.9493	0.0368	2	178	Barbera	0.0145	0.0191	0.9664	3
79	Grignolino	0.0431	0.8662	0.0908	2						
80	Grignolino	0.0563	0.9338	0.0100	2						
81	Grignolino	0.0321	0.9506	0.0174	2						
82	Grignolino	0.0788	0.9078	0.0135	2						
83	Grignolino	0.0214	0.9631	0.0155	2						
84	Grignolino	0.0141	0.0707	0.9152	3						
85	Grignolino	0.0167	0.9705	0.0128	2						
86	Grignolino	0.0107	0.9847	0.0046	2						
87	Grignolino	0.0373	0.8962	0.0665	2						
88	Grignolino	0.0341	0.9393	0.0265	2						
89	Grignolino	0.0085	0.9836	0.0079	2						

Table 25 continued

	Wine	Cluster 1	Cluster 2	Cluster 3	Cluster	Wine	Cluster 1	Cluster 2	Cluster 3	Cluster
90	Grignolino	0.0053	0.9921	0.0026	2					
91	Grignolino	0.0352	0.7908	0.1740	2					
92	Grignolino	0.0301	0.8294	0.1405	2					
93	*Grignolino*	0.0410	0.4084	0.5505	3					
94	Grignolino	0.0095	0.9864	0.0042	2					
95	Grignolino	0.0439	0.9312	0.0248	2					
96	Grignolino	0.1437	0.8263	0.0299	2					
97	Grignolino	0.0376	0.7875	0.1748	2					
98	Grignolino	0.0195	0.9712	0.0094	2					
99	*Grignolino*	0.4987	0.4244	0.0770	1					
100	Grignolino	0.1123	0.8566	0.0311	2					

Funding Open access funding provided by Luiss University within the CRUI-CARE Agreement.

Data Availability Data will be made available on request.

Declarations

Conflict of interest None of the authors has any financial or personal relationships that influenced the development of this work.

References

Anderson, D.T., Bezdek, J.C., Popescu, M., et al.: Comparing fuzzy, probabilistic, and possibilistic partitions. IEEE Trans. Fuzzy Syst. **18**, 906–918 (2010)

Azzalini, A.: The Skew-Normal and Related Families. Cambridge University Press, Institute of Mathematical Statistics Monographs (2013)

Azzalini, A.: The R package `sn`: the skew-normal and related distributions such as the skew-*t* and the SUN (version 2.1.1). Università degli Studi di Padova, Italia (2023). https://cran.r-project.org/package=sn, home page: http://azzalini.stat.unipd.it/SN/

Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Springer Science & Business Media, Berlin (2013)

Bezdek, J.C., Keller, J., Krisnapuram, R., et al.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer Science & Business Media, Berlin (1999)

Campello, R.J.G.B.: A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. Pattern Recogn. Lett. **28**(7), 833–841 (2007)

Campello, R.J.G.B., Hruschka, E.: A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets Syst. **157**, 2858–2875 (2006)

Coppi, R., D'Urso, P., Giordani, P.: A fuzzy clustering model for multivariate spatial time series. J. Classif. **27**(1), 54–88 (2010)

D'Urso, P.: Fuzzy clustering. In: Hennig, C., Meila, M., Murtagh, F., et al. (eds.) Handbook of Cluster Analysis, pp. 545–573. Chapman and Hall, London (2015)

Flury, B., Riedwyl, H.: Multivariate Statistics: A Practical Approach. Chapman & Hall, Springer (1988)

Forina, M., Armanino, C., Castino, M., et al.: Multivariate data analysis as a discriminating method of the origin of wines. Vitis **25**, 189–201 (1986)

Forina, M., Armanino, C., Castino, M., et al.: V-parvus 2008: an extendible package of programs for exploratory data analysis, classification and regression analysis. Dip Chimica e Tecnologie Farmaceutiche ed Alimentari, Università di Genova, Italia (2008)

Hathaway, R.J., Bezdek, J.C.: Local convergence of the fuzzy c-means algorithms. Pattern Recogn. **19**(6), 477–480 (1986)

Hathaway, R.J., Bezdek, J.C.: Recent convergence results for the fuzzy c-means clustering algorithms. J. Classif. **5**(2), 237–247 (1988)

Hennig, C., Viroli, C., Anderlucci, L.: Quantile-based clustering. Electron. J. Stat. **13**(2), 4849–4883 (2019)

Holzmann, H., Klar, B.: Expectile asymptotics. Electron. J. Stat. **10**(2), 2355–2371 (2016)

Hüllermeier, E., Rifqi, M., Henzgen, S., et al.: Comparing fuzzy partitions: a generalization of the Rand index and related measures. Fuzzy Syst. IEEE Trans. **20**(3), 546–556 (2012)

Maharaj, E., D'Urso, P.: Fuzzy clustering of time series in the frequency domain. Inform. Sci. **181**(7), 1187–1211 (2011)

Murray, P.M., Browne, R.P., McNicholas, P.D.: Mixtures of skew-t factor analyzers. Comput. Stat. Data Anal. **77**, 326–335 (2014)

Newey, W.K., Powell, J.L.: Asymmetric least squares estimation and testing. Econometrica **55**(4), 819–847 (1987)

Pal, N., Bezdek, J.: On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst. **3**, 370–379 (1995)

Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)

Telford, R.D., Cunningham, R.B.: Sex, sport, and body-size dependency of hematology in highly trained athletes. Med. Sci. Sports Exerc. **23**(7), 788–94 (1991)

Wang, B., Li, Y., Härdle, W.K.: K-expectiles clustering. J. Multivar. Anal. **189**, 104869 (2022)
Zhang, Y., Wang, H.J., Zhu, Z.: Quantile-regression-based clustering for panel data. J. Econom. **213**(1), 54–67 (2019)