ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Wang, Bingling; Li, Yingxing; Härdle, Wolfgang

Working Paper

K-expectiles clustering

IRTG 1792 Discussion Paper, No. 2021-003

Provided in Cooperation with:
Humboldt University Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series"

Suggested Citation: Wang, Bingling; Li, Yingxing; Härdle, Wolfgang (2021) : K-expectiles clustering, IRTG 1792 Discussion Paper, No. 2021-003, Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Berlin

This Version is available at:
https://hdl.handle.net/10419/233511

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

IRTG 1792 Discussion Paper 2021-003

K-expectiles clustering

Bingling Wang [*]
Yingxing Li [* *2]
Wolfgang Karl Härdle [* *3 *4 *5]

[*] Humboldt-Universität zu Berlin, Germany
[*2] Xiamen University, China
[*3] Singapore Management University, Singapore
[*4] Charles University, Czech Republic
[*5] National Chiao Tung University, Taiwan

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

K-expectiles clustering *

Bingling Wang† Yingxing Li†‡ Wolfgang Karl Härdle †‡§¶‖**

February, 2021

Abstract

K-means clustering is one of the most widely-used partitioning algorithm in cluster analysis due to its simplicity and computational efficiency, but it may not provide ideal clustering results when applying to data with non-spherically shaped clusters. By considering the asymmetrically weighted distance, We propose the K-expectile clustering and search the clusters via a greedy algorithm that minimizes the within cluster τ-variance. We provide algorithms based on two schemes: the fixed τ clustering, and the adaptive τ clustering. Validated by simulation results, our method has enhanced performance on data with asymmetric shaped clusters or clusters with a complicated structure. Applications of our method show that the fixed τ clustering can bring some flexibility on segmentation with a decent accuracy, while the adaptive τ clustering may yield better performance. All calculation can be redone via quantlet.com.

Keywords: *clustering, expectiles, asymmetric quadratic loss, image segmentation*

*Financial support of the European Union's Horizon 2020 research and innovation program "FIN-TECH: A Financial supervision and Technology compliance training programme" under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA), the European Cooperation in Science & Technology COST Action grant CA19130 - Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry, the Deutsche Forschungsgemeinschaft's IRTG 1792 grant, the Yushan Scholar Program of Taiwan, the Czech Science Foundation's grant no. 19-28231X / CAS: XDA 23020303 are greatly acknowledged. All correspondence may be addressed to the authors by e-mail at yxli@xmu.edu.cn.

†Humboldt-Universität zu Berlin, IRTG 1792, Dorotheenstr.1, 10117 Berlin, Germany, E-mail: bingling.wang@hu-berlin.de

‡Wang Yanan Institute for Studies in Economics, Xiamen University, 422 Siming S Rd, 361005 Fujian, China

§Humboldt-Universität zu Berlin, Blockchain Research Center, Unter den Linden 6, 10099 Berlin, Germany

¶Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, Singapore 178899

‖Faculty of Mathematics and Physics, Charles University, Ke Karlovu 3, 121 16 Prague, Czech Republic

**Department of Information Management and Finance, National Chiao Tung University, Management Building 1, 1001 University Road, Hsinchu, Taiwan 30010, ROC.

1 Introduction

Clustering is a useful technique to discover and identify homogenous groups of data points in a given sample. As an unsupervised learning algorithm, it aims to extract information on the underlying characteristics via dividing the data into groups that maximize common information. Obviously the information about homogeniety of groups is key in such a sample dividing mechanism. Among the simplest choice is the K-means clustering method described by Steinhaus (1956) and Hartigan (1975), which adopt the Euclidean distance as neighbourhood measure, thus leading to spheres as silhouettes and means as centers of clusters. Indeed, while keeping a balance between group size and information gain, K-means is the most widely used partitioning algorithm due to its simplicity, efficiency in computing and easiness of interpretation. Successful applications include signal processing, image identification, customer segmentation.

The principle of a partitioning clustering algorithm is to assign data points to the nearest cluster by optimising some objective function. The objective function of K-means is the sum of within-group variance, and thus the correspondence cluster centers are the mean of each cluster. Minimizing the objective function is equivalent to maximizing the log-likelihood function with independent Gaussian density. Although K-means clustering is often viewed as a "distribution free" algorithm, it is actually partitioning using equal sized spherical contour lines which can be considered as assuming independent identically distributed (i.i.d.) Gaussian clusters. Therefore, K-means approach works better for cluster in the symmetric distribution than the skewed ones.

On the other hand, when applied on skewed or asymmetric distributed data whose characteristics may not be fully captured by the first two moments, new methods are required for non-spherical cluster. To account for within-cluster skewness, Hennig et al. (2019) introduce the K-quantile clustering algorithm based on the assymmetric absoluate discrepancy. Then they linked their approach to a fixed partition model of genralized asymmetric Laplace distributions. This quantile discrepancy based density relies on both the quantile level τ and some additional scale/penalty parameter λ. However, τ and λ are assumed the same across different clusters to reduce the computation complexity. An analogous work on quantile based clustering is proposed in Zhang et al. (2019), where they developed a model-based iterative algorithm to identify subgroups with heterogeneous slopes. In particularly, they consider clustering across multiple quantiles to capture the full picture of heterogeneity. For that accordance, how to specify the appropriate quantile level vector τ could be a problem for large dimensional data.

This motivates us to consider a novel method, K-expectile clustering. This method is based on a similar idea as K-means but with an expectile cluster center and aims at minimizing the so-called τ-variance, which is a weighted quadratic loss to take into account asymmetry. Besides being simple and fast, our algorithm can be applied on wider range of data compared with K-means. In particular, we consider two schemes, either with a pre-specify τ level or an adaptive τ that may vary across different dimensions or clusters, which accommodates either a fixed cluster shape or a data-driven cluster shape to capture heterogeniety.

To better understand the basic ideas of K-expectile clustering, we recall some basic knowledge about tail events. Quantile regression (Koenker and Bassett Jr, 1978) and expectile regression (Newey and Powell, 1987) have been suggested for displaying the whole picture of the conditional distribution of response variable on covariates, especially for data not sufficing the condition of homoskedasticity or conditional symmetry. For a

random variable $X \in \mathbb{R}$ drawn from distribution F, a location model of τ-th tail event measure with $\tau \in (0,1)$ could be defined as:

$$x_i = \theta_\tau + \varepsilon_i, \;\; i = 1, \ldots, n.$$

With an assumption on the τ-th quantile or expectile of the cdf of ε being zero, θ_τ is by definition the τ-th quantile or expectile of X accordingly. An estimator of the location model of quantiles and expectiles can be naturally formed:

$$\hat{\theta}_\tau = \arg\min_{\mu \in \mathbb{R}} \mathsf{E}\left[\rho_\tau(X - \mu)\right],$$

where the loss function $\rho_\tau(\cdot)$ is defined as:

$$\rho_\tau(u) = |u|^\alpha |\tau - \mathbf{I}_{\{u \le 0\}}|,$$

with $\alpha = 1$ and $\alpha = 2$ respectively.

Although the concept of expectiles is natural analogues of quantiles, expectiles enjoy the computation efficiency over quantiles (Schnabel, 2011). In finance, the expectile might be preferred as a favorable risk measures due to its desirable properties such as coherence and elicitability (Kuan et al., 2009, Ziegel, 2016). Recently, the use of expectiles attracts more and more attention, such as the nonparametric expectile regression by Sobotka and Kneib (2012) and Yang et al. (2018), the principle expectile analysis by Tran et al. (2019). Our proposed K-expectile clustering allows us to take into account tail characteristics and asymmetry when identifying homogenous groups of data, while simulation studies and applications justify its excellent performance.

The rest of the paper is organized as follows. In section 2, we will briefly review the classical K-means algorithm, and then propose our K-expectile clustering in two schemes. In section 3, we present the simulation study that includes data from different distribution and compare the performance of K-expectiles clustering with other methods. Section 4 applies our method to real crypto currency market analysis and image segmentation. Codes of all the functions, applications and data are uploaded to quantlet.com.

2 Methodology

2.1 K-means clustering

Suppose the data set $X = \{X_i\}_{i=1}^n$ comes from a random sample in $\mathbb{R}^p$. A clustering algorithm denoted by $Q(\cdot)$ generates K subsets $\{G_1, G_2, \ldots, G_K\}$ each with distribution $p_k(X)$. Any clustering algorithm maps X into a membership vector $C = (c(1), c(2), \ldots, c(n))$, i.e. $Q(x_i) = c(i)$, and $G_k = \{x_i : c(i) = k\}$, $c(i) \in \{1, 2, \ldots, K\}$.

A clustering criterion is defined via a cost function. In K-means clustering, the cost is defined as the sum of squared Euclidean distance between cluster members to the cluster centroids. Indeed the centroids can be considered as location parameters for clusters. In K-means clustering, cluster centroids are actually cluster means and the cost function is the sum of within-cluster variance. Let $G(\cdot)$ be the K-means objective function, $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$ be a set of cluster centroids with $\theta_k \in \mathbb{R}^p$,

$$G(\Theta, C, X) = \min_{\Theta} \sum_{k=1}^{K} \sum_{x_i \in G_k} \|x_i - \theta_k\|^2.$$

Clustering is now turned into an optimisation problem and is solved via iteration. For a fixed Θ, partition $c(i)$ is achieved by assigning each point to the nearest cluster centroid.

$$c(i) = \arg \min_{k \in \{1,2,\ldots,K\}} \|x_i - \theta_k\|^2.$$

For a fixed membership vector C, the centroid θ_k can be estimated by taking the within-cluster mean.

$$\hat{\theta}_k = \bar{x}_k = \frac{1}{|G_k|} \sum_{x_i \in G_k} x_i.$$

Suppose X are drawn i.i.d. from an equally-weighted Gaussian Mixture Model, with conditional unknown expectation μ_k and variance Σ_k. The estimation of this model requires the EM algorithm, and the details are explained in Deisenroth et al. (2020). If let μ_k be the k-th cluster centrioid, and covariance matrix simply equals to the identity matrix, then the K-means objective function is coincide with the expectation function in EM algorithm of a Gaussian Mixture Model with equal mixture weights. By a "hard" assignment of data points to nearest cluster centroid in K-means algorithm as described in MacKay and Mac Kay (2003), the computation of parameter Θ can be easily conducted by independently estimating cluster means. Simoultaneously K-means clustering is a distribution-free but distance based clustering technique.

2.2 K-expectiles clustering

Now consider a set of data with skewed or asymmetrically distributed clusters, i.e., cluster centroids Θ are not located on means and cluster variances are heterogeneous on different sides around centroids. As said before, it is the information measure of homogeneity that yields the clusters. A distance with cluster centroids offset from means and a distance metric which takes asymmetry into account is certainly a more flexible way of dividing groups.

For that purpose, assume each cluster is a group of data drawn independently from some distribution. Define the cluster centroid θ_k as the τ expectile of cluster distribution, and assign points according to expectile distances. Take the univariate random variable $X \in \mathbb{R}$ as an example. For a fixed $\tau \in (0,1)$, the τ-th expectile of X as proposed by Newey and Powell (1987) is identified as the minimizer of the asymmetric quadratic loss

$$e_\tau(X) = \arg \min_{\mu \in \mathbb{R}} \mathsf{E}\left[\rho_\tau(X - \mu)\right], \tag{1}$$

$$\rho_\tau(X - \mu) = \tau(X - \mu)_+^2 + (1 - \tau)(\mu - X)_+^2, \tag{2}$$

where $(x)_+ = \max(x, 0)$. It is worth noting here that the expectile location estimator can be interpreted as a maximum likelyhood estimator of a normal distributed sample with an unequal weight placed on positive and negative disterbances, showed in Aigner et al. (1976).

For $X \in \mathbb{R}^p$, define $(X)_+ = ((X_1)_+, \ldots, (X_p)_+)^\top$, then the multivariate expectile is:

$$e_\tau(X) = \arg \min_{\mu \in \mathbb{R}^p} \mathsf{E}\left[\tau \|(X - \mu)_+\|^2 + (1 - \tau)\|(\mu - X)_+\|^2\right]. \tag{3}$$

Here the dependence is taken into account by using the norm. Note that the construction of the multivariate expectiles are related to the dependence structure of each components. The choice of dependence modelling may differ according to the practical goal. For

simplification, we first elaborate the case when the dependence structure is ignored. Then the multivariate expectile $e_\tau \in \mathbb{R}^p$ can be defined via the marginal univariate expectiles, $e_\tau(X) = (e_\tau(X_1), \ldots, e_\tau(X_p))^\top$, where

$$e_\tau(X_j) = \arg\min_{\mu \in \mathbb{R}^p} \mathsf{E}\left[\tau\{(X_j - \mu_j)_+^2\} + (1-\tau)\{(\mu_j - X_j)_+^2\}\right]$$

We could gain further flexibility and more power of expectiles in p dimension by setting $\tau = (\tau_1, \tau_2, \ldots, \tau_p)^\top \in \mathbb{R}^p$, making the τ level variable over dimensions. Then we obtain $e_\tau(X) = (e_{\tau_1}(X_1), \ldots, e_{\tau_p}(X_p))^\top$, where

$$e_{\tau_j}(X_j) = \arg\min_{\mu \in \mathbb{R}^p} \mathsf{E}\left[\tau_j\{(X_j - \mu_j)_+^2\} + (1-\tau_j)\{(\mu_j - X_j)_+^2\}\right]$$

The empirical version reads as:

$$\hat{e}_{\tau,n}(X) = \arg\min_{\mu \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \{\tau_j (x_{ij} - \mu_j)_+^2 + (1-\tau_j)(\mu_j - x_{ij})_+^2\}, \tag{4}$$

whose idea is to consider an asymetrically weighted distance function with L_2 norm and fix the cluster centroids at the empirical expectile of the k-th cluster, i.e. $\hat{\theta}_k = \hat{e}_{\tau,n}(X \in G_k)$.

Instead of specifying an asymmetric form of distribution, we form a K-expectile objective function by defining an asymmetric τ-variance as described in Tran et al. (2019), which yields to an axis-aligned elipsoid unit ball. To include covariance or correlation, usually a matrix form of multivariate expctile will be considered. By introducing a $p \times p$ symmetric matrix Σ, one can form a score function as described in Maume-Deschamps et al. (2016),

$$e_\tau^\Sigma(X) \in \arg\min_{\mu \in \mathbb{R}^p} \mathsf{E}\left[\tau (x-\mu)_+^\top \Sigma (x-\mu)_+ + (1-\tau)(x-\mu)_+^\top \Sigma (x-\mu)_+\right]. \tag{5}$$

In Figure 2.1, the contour lines of unit circles of bivariate τ-variance with various τ levels on each axis are shown along with unit circles of a symmetric variance in the back. The covariance matrix is the inverse matrix of Σ in function (5). The last sub-plot shows the unit circles with different scales on two axis. These are equivalent to the contour lines of independent bivariate asymmetric normal distributions in comparison with the contour lines of independent normal distributions. Figure 2.2 shows the 3D contour surface of τ-variance unit ball with different τ-levels on each dimension, or the 3D cluster shapes.

2.3 Fixed τ clustering

Recall that for a univariate observation $x \in \mathbb{R}$, we could define τ-distance as

$$d(x, \tau, \theta) = \left\{\tau + (1-2\tau)\, \mathbf{I}_{\{x<\theta\}}\right\} (x-\theta)^2, \tag{6}$$

which coincides with loss function (2). Therefore, for a pre-specified τ vector, define the objective function

$$G^{Fixed}(\tau, \Theta, C, X) = \sum_{k=1}^{K} \sum_{x_i \in G_k} \sum_{j=1}^{p} d(x_{ij}, \tau_j, \theta_k) \tag{7}$$

$$= \sum_{k=1}^{K} \sum_{x_i \in G_k} \sum_{j=1}^{p} \left\{\tau_j + (1-2\tau_j)\, \mathbf{I}_{\{x_{ij}<\theta_{k,j}\}}\right\} (x_{ij} - \theta_{k,j})^2, \tag{8}$$

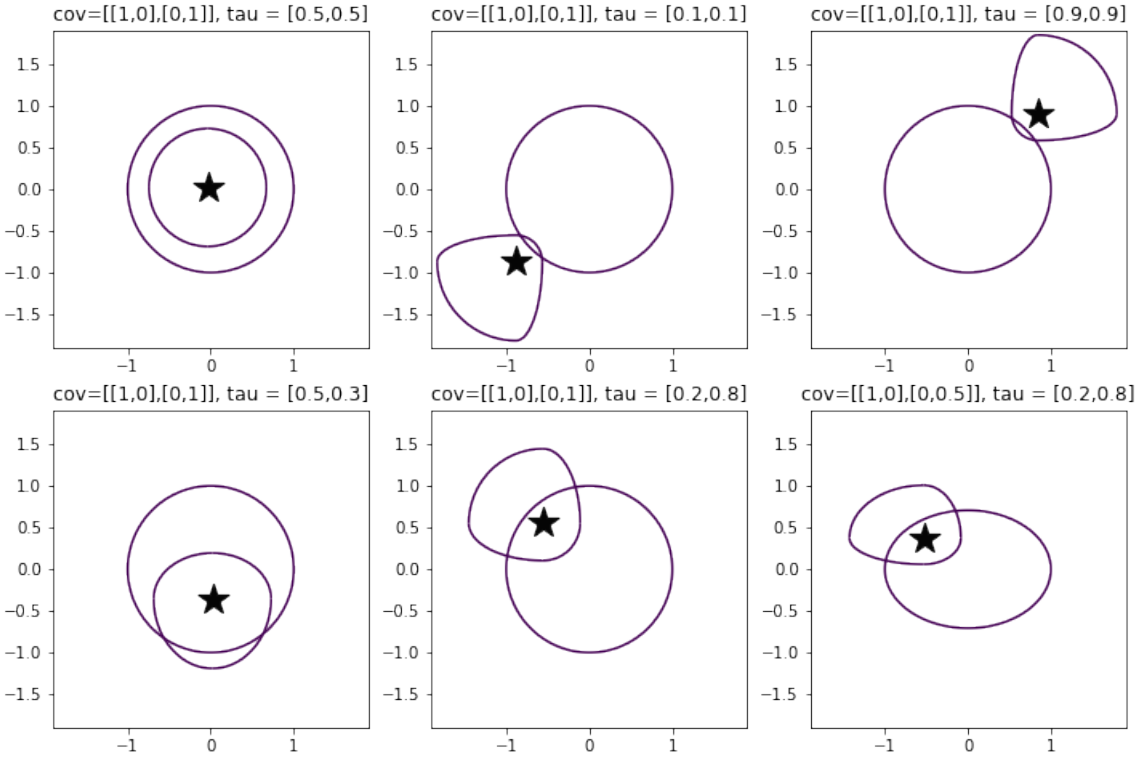

Figure 2.1: Contour lines of unit balls of various τ variances in comparison to unit balls of symmetric variance

KEC_cluster shapes

which aims to detect expectile-specified clusters by minimizing the sum of within-cluster τ-variance.

For known (τ, C), cluster centroids Θ are found by:

$$\hat{\theta}_k = \arg\min_{\mu \in \mathbb{R}^p} \sum_{x_i \in G_k} \sum_{j=1}^{p} \left\{ \tau_j + (1 - 2\tau_j)\, \mathbf{I}_{\{x_{ij} < \mu_j\}} \right\} (x_{ij} - \mu_j)^2 \tag{9}$$

$$= \arg\min_{\mu \in \mathbb{R}^p} \sum_{x_i \in G_k} \sum_{j=1}^{p} w_{ij}(\tau_j)(x_{ij} - \mu_j)^2 \tag{10}$$

where $w(\tau)$ is a weight function which is related to $\mu(\tau)$, the location parameter at the given τ level.

$$w_{ij}(\tau_j) = \begin{cases} \tau_j & \text{if } x_{ij} \le \mu_j(\tau_j) \\ 1 - \tau_j & \text{if } x_{ij} > \mu_j(\tau_j). \end{cases} \tag{11}$$

This implicit dependence of w on $\mu(\tau)$ leads to the application of the Least Absolute Square Estimator (LAWS), a version of the Stochastic Gradient Algorithm. For a fixed $\mu_j(\tau_j)$, the weight $w_{ij}(\tau_j)$ in equation (11) is calculated, therefore a closed form solution of $\mu_j(\tau_j)$ can be expressed as

$$\mu_j(\tau_j) = \frac{\tau_j \sum_{i \in \mathcal{I}^+_{\tau_j}} x_{ij} + (1 - \tau_j) \sum_{i \in \mathcal{I}^-_{\tau_j}} x_{ij}}{\tau_j n^+ + (1 - \tau_j) n^-} \tag{12}$$

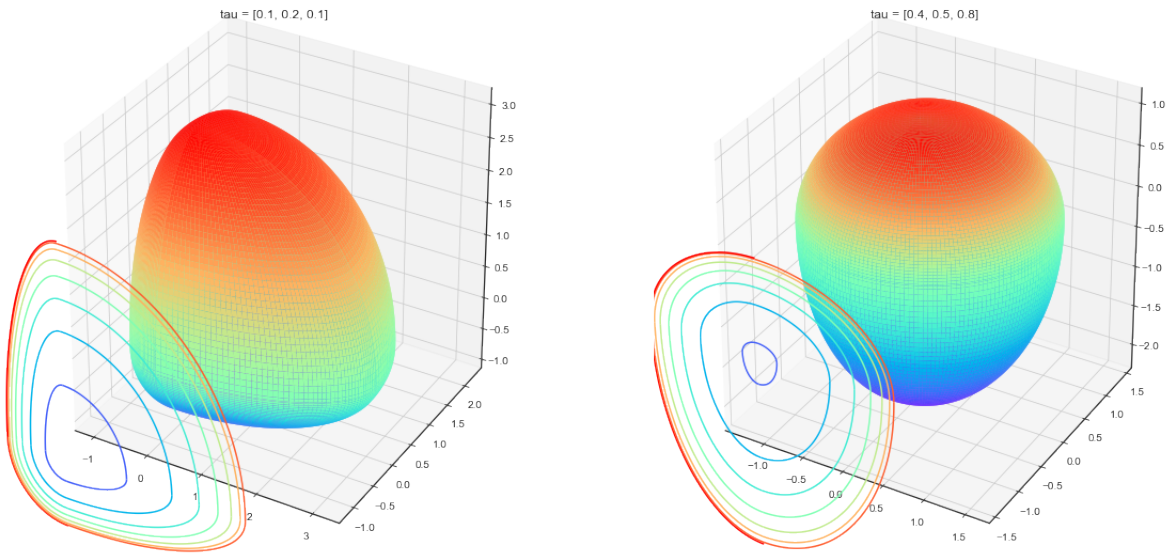

Figure 2.2: 3D contour surface of unit balls of various τ variances. Left: $\tau = [0.1, 0.2, 0.1]$. Right: $\tau = [0.4, 0.5, 0.8]$.

QKEC_cluster shapes

where

$$\mathcal{I}_\tau^+ = \{i \in \{1, \ldots, n\} : w_{ij} = \tau_j, c(i) = k\}$$
$$\mathcal{I}_\tau^- = \{i \in \{1, \ldots, n\} : w_{ij} = 1 - \tau_j, c(i) = k\}$$
$$n^+ = |\mathcal{I}_\tau^+| \qquad n^- = |\mathcal{I}_\tau^-|.$$

Cluter centroids can be estimated by iteratively repeating the two steps until the location of $\mu_j(\tau_j)$ does not change, see:

Algorithm 1 LAWS

Input: The set of points in cluster G_k; The vector of parameter, τ;
Output: Estimated cluster centroids, Θ

1: Initialize $\mu_j^0(\tau_j)$ as mean of j-th column of $x_i \in G_k$
2: **repeat**
3: Assign weight $w_{ij}^{t+1}(\tau_j)$ to each point x_{ij} based on $\mu^t(\tau)$
4: Update $\mu^{t+1}(\tau)$ according to equation (12) with input $w^{t+1}(\tau)$
5: **until** $\|\{\mu(\tau)^t, \mu(\tau)^{t-1}\| \leq threshold$

The K-expectiles clustering algorithm now read as follows:

Algorithm 2 Fixed τ clustering

Input: Data, X; Vector parameter, τ; # of clusters, K;
Output: Cluster membership vector, C; Estimated cluster centroids, Θ

1: Initialize centroids $\Theta^0 = \Theta_{k-means}$
2: **repeat**
3: Calculate cluster membership

$$c(i)^{t+1} = \arg \min_{k \in \{1,2,\ldots,K\}} \sum_{j=1}^{p} \sum_{i:c(i)^t=k} \left\{ \tau_j + (1-2\tau_j)\, \mathbf{I}_{\left\{x_{ij} < \theta^t_{k,j}\right\}} \right\} (x_{ij} - \theta^t_{k,j})^2$$

4: Update θ_k^{t+1} by applying Algorithm (1) with input $c(i)^{t+1}$
5: **until** $\|\theta^t - \theta^{t-1}\| \leq threshold$

2.4 Adaptive τ clustering

In the last section, clustering with fixed cluster shapes by pre-specifying the τ vector has been discussed, and this senario is shown in Sub-plot 1 of Figure 2.3. In comparison, regarding the issue of clusters with different shapes as shown in Sub-plot 2 of Figure 2.3, we present the result of a fully adaptive algorithm, both for different dimensions and for different clusters. Without pre-defined τ, we now assume τ is a $(K \times p)$ matrix. We optimize the following cluster objective function with respect to τ as well.

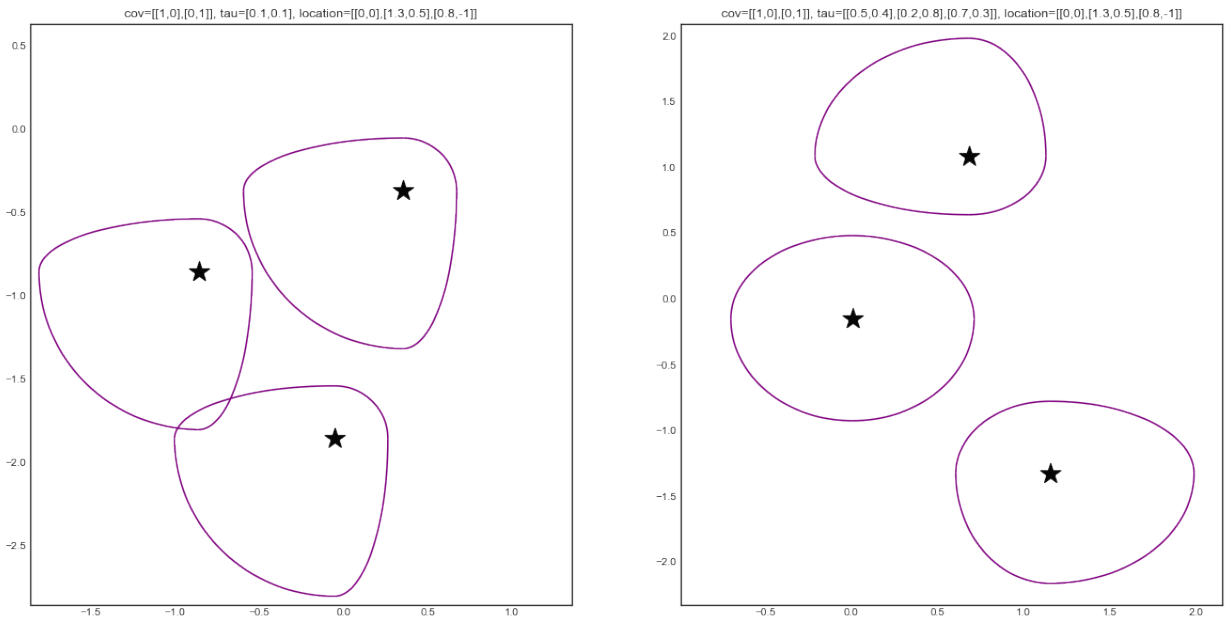

Figure 2.3: Different senarios regarding to cluster shapes

KEC_cluster shapes

$$G^{Adaptive}(\tau, \Theta, C, X) = \sum_{k=1}^{K} \sum_{x_i \in G_k} \sum_{j=1}^{p} d(x_{ij}, \tau_{k,j}, \theta_{k,j}) \tag{13}$$

$$= \sum_{k=1}^{K} \sum_{x_i \in G_k} \sum_{j=1}^{p} \left\{ \tau_{k,j} + (1 - 2\tau_{k,j}) \mathbf{I}_{\{x_{ij} < \theta_{k,j}\}} \right\} (x_{ij} - \theta_{k,j})^2, \tag{14}$$

For given (Θ, C), to optimize τ, require

$$\hat{\tau}_k = \arg \min_{\tau \in (\mathbb{R}^p)^K} G^{Adaptive}(\tau, \Theta, C, X). \tag{15}$$

By taking first order condition, we get the unique solution:

$$\tau_{k,j} = \frac{\gamma_{k,j}}{1 + \gamma_{k,j}}, \tag{16}$$

where

$$\gamma_{k,j} = \frac{n^- \sum_{i \in \mathcal{I}_\tau^+} \theta_{k,j} - x_{ij}}{n^+ \sum_{i \in \mathcal{I}_\tau^+} x_{ij} - \theta_{k,j}}.$$

Then the clustering algorithm for adaptive τ can be described as: The clustering algorithm can be implemented as follows:

Algorithm 3 Adaptive τ clustering

Input: Data, X; # of clusters, K;
Output: Cluster membership vector, C; Estimated cluster centroids, Θ

1: Initialize centroids $\Theta^0 = \Theta_{k-means}$; $\tau_{k,j}^0 = 0.5$
2: **repeat**
3: Calculate the cluster membership by

$$c(i)^{t+1} = \arg \min_{k \in \{1,2,\ldots,K\}} \sum_{j=1}^{p} \sum_{i: c(i)^t = k} \left\{ \tau_{k,j}^t + (1 - 2\tau_{k,j}^t) \mathbf{I}_{\left\{x_{ij} < \theta_{k,j}^t\right\}} \right\} (x_{ij} - \theta_{k,j}^t)^2$$

4: Update θ_k^{t+1} by applying Algorithm (1) with input τ_k^t
5: Update τ_k^{t+1} according to equation (16) with input θ_k^{t+1}
6: **until** $\|\theta^t, \theta^{t-1}\| \leq threshold$

Finally, we point out that our algorithm has the numerical convergence.

Proposition 2.1. *The sequence produced by the adaptive τ clustering algorithm could converge to its optimal solution.*

3 Simulation

To evaluate the performance of K-expectiles clustering, we design four simulated samples with K clusters. Let $k = 1, \ldots, K$, each cluster represented by G_k is an i.i.d. random sample drawn from a p-variate distribution in the size of $(n_1, n_2, \ldots, n_k)$. Each component of the multivariate distribution is assumed to be independent. Data set can be written as $X = (G_1, G_2, \ldots, G_k)$. Scale, location and skewness of the distribution can cause the overlapping of multiple clusters which in turn influece the cluster shapes and within-cluster data density, thus hinder the accuracy of grouping results. The simulated samples are designed to reserve some extend of overlap while ensure certain discrimination between clusters, in order to achieve the purpose of evaluating the robustness of the algorithms.

Sample 1: In the first sample we generate K multivariate Gaussian clusters with unit variance and different location parameters. $G_k \sim N(\mu_k, \mathcal{I}_p)$, where μ_1 is a p-dimensional integer vector whose elements are randomly generated in interval $(1, 10)$, and then shift the location of other clusters by $\mu_k = \mu_1 + 2k$. Clusters are in the same size of $n_k = n/k$.

Sample 2: To include some asymmetry on the basis of Sample 1, the second sample is designed as a mixture of K asymmetric normal distributions. Each cluster G_k is considered as a p-dimensional i.i.d. sub-sample, where $G_k = (W_1, W_2, \ldots, W_p)^\top$. The probability density function of W_j can be expressed as following, with $\mu_j (j = 1, 2, \ldots, p)$ as location parameters, and σ_l, σ_r as standard deviation of two sides around μ,

$$p(W_j \mid \theta) = \prod_{j=1}^{p} \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_{l_j} + \sigma_{r_j}} \begin{cases} \exp\left\{-\dfrac{(W_j - \mu_j)^2}{2\sigma_{l_j}^2}\right\} & W_j < \mu_j \\ \exp\left\{-\dfrac{(W_j - \mu_j)^2}{2\sigma_{l_j}^2}\right\} & W_j \geq \mu_j \end{cases},$$

Now let e_τ (τ-expectile of the variable) be the location parameter, and σ_j be the overall standard deviation of the variable, the density function of asymmetric normal distribution can be rewritten as:

$$p(Z_j \mid e_\tau, \sigma_j, \tau) = \begin{cases} \displaystyle\prod_{j=1}^{p} \sqrt{\frac{1}{2\pi(\tau^{-\frac{1}{2}}\sigma_j)}} \frac{2\sqrt{1-\tau}}{\sqrt{1-\tau}+\sqrt{\tau}} \exp\left\{-\frac{(Z_j - e_\tau)^2}{2(\tau^{-\frac{1}{2}}\sigma_j)^2}\right\} & Z_j < e_\tau \\ \displaystyle\prod_{j=1}^{p} \sqrt{\frac{1}{2\pi((1-\tau)^{-\frac{1}{2}}\sigma_j)}} \frac{2\sqrt{\tau}}{\sqrt{1-\tau}+\sqrt{\tau}} \exp\left\{-\frac{(Z_j - e_\tau)^2}{2((1-\tau)^{-\frac{1}{2}}\sigma_j)^2}\right\} & Z_j \geq e_\tau \end{cases},$$

which means the asymmetric normally distributed variable $W_j (j = 1, 2, \ldots, p)$ can be converted from univariate Gaussian distributed variables Z_j according to the formula:

$$W_j^k = \begin{cases} \dfrac{2\sqrt{\tau_j^k}}{\sqrt{1-\tau_j^k}+\sqrt{\tau_j^k}} \cdot \dfrac{1}{\sqrt{1-\tau_j^k}} \cdot Z_j + e_{\tau_j^k} & Z_j^k < 0 \\ \dfrac{2\sqrt{1-\tau_j^k}}{\sqrt{1-\tau_j^k}+\sqrt{\tau_j^k}} \cdot \dfrac{1}{\sqrt{\tau_j^k}} \cdot Z_j + e_{\tau_j^k} & Z_j^k \geq 0 \end{cases},$$

in our sample, each $Z_j^k \sim N(0,25)$. Parameter τ_j^k is given by using random generator with interval $[0.1, 0.9]$, and location parameter $e_{\tau_j^1}$ is randomly generated in $(0,10)k$-th cluster, then the location of k-th cluster can be shifted by $e_{\tau_j^k} = e_{\tau_j^1} + 7(-1)^j(j-1)$.

Sample 3: In the third sample we want to test the algorithm on skewed but not leptokurtic clusters, namely *Beta*-distributed clusters. For variables in cluster$\{k = 2c+1, c \in \mathbb{Z}\}$, $W_j^i \sim Beta(a_j, b_j)$, $(j = 1, 3, \ldots, p-1)$, and in cluster $\{k = 2c, c \in \mathbb{Z}\}$, $W_j^i \sim Beta(b_j, a_j)$, $(j = 2, 4, \ldots, p)$. We generate parameter a_j randomly from interval $(1, 10)$ and b_j from interval $(10, 20)$, again $K = 3$.

Sample 4: For the last sample, skewed and leptokurtic clusters are being considered. We set 2 different scenarios:

- K skewed generalized t-distributed samples. We first generated a random sample with dimension $p = 2$, parameters $df = [10, 10, 10], nc = [3, -1.5, 2.5]$, location randomly fluctuated with the difference of 0.5 around $[[0, 2], [1, 0], [0.5, 1]]$, $scale = 0.5$. And generate data repeatedly until p reaches 10 and 50.
- K multivariate F-distributed clusters. For variables in the first cluster, $W_j^1 \sim F(a_j, a_j) + 1$, and when $j = 1, 3, \ldots, p-1$; $W_j^1 \sim F(b_j, b_j) + 1$, when $j = 2, 4, \ldots, p$, where a_j and b_j are integers randomly selected from interval $(51, 60)$ and $(21, 30)$. In the second cluster, $W_j^2 \sim F(b_j, b_j)$, $j = 1, 3, \ldots, p-1$, $W_j^2 \sim F(a_j, a_j)$, $j = 2, 4, \ldots, p$, where a_j and b_j are integers randomly selected from interval $(5, 15)$ and $(25, 35)$. In the third cluster, $W_j^3 \sim F(a_j, b_j)$, $j = 1, 3, \ldots, p-1$, and $W_j^3 \sim F(b_j, a_j)$, $j = 2, 4, \ldots, p$, where a_j and b_j are integers randomly selected from interval $(15, 25)$ and $(60, 70)$.

For each of the first three samples, we evaluate combinations of $p = 50, 100, 500, n = 300, 1500$. For the last sample, $p = 2, 10, 50$. For each simulation setting, we re-generate the data 50 times and test the algorithms each round, and take the average of the Adjusted Rand Index (ARI) of the yielded classification compared with the true cluster membership. Rand Index measures the pair-wised agreement between data clustering. When it is djusted for the chance grouping of elements, this is the Adjusted Rand Index. Given two partitions $X = X_1, X_2, \ldots, X_r$, $Y = Y_1, Y_2, \ldots, Y_s$, and the contingency table,

$X \setminus Y$	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

,

the Ajusted Rand Index is defined as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}.$$

Results are shown in Apendix, where the demonstrated values are the 100 times of ARI. We have considered other distance based clustering algorithms such as K-means denoted by *K-means*, spectral clustering (Shi and Malik, 2000) denoted by *spectral*, Ward hierarchical clustering (Ward Jr, 1963) denoted by *h-ward*, and Quantile based clustering (Hennig

et al., 2019) as well as K-expectile clustering with adaptive τ. Note that Quantile based clustering algorithm allows quantile level (skewness parameter) τ to change variable-wisedly and introduces a scale/penalty parameter. *CU, CS, VU, VS* stands for the four modes of Quantile based clustering, corresponding to Common skewness parameter and Unscaled variables, Common skewness parameter and Scaled variable-wise, Variable-wise skewness parameter and Unscaled variables, Variable-wise skewness parameter and Scaled variable-wise. Note that spectral clustering sometimes does not work appropriate on data with outliers which lead to a not fully connected graph. This scenario can be easily occured in a highly skewed sample or sample with large dimensionality.

From Table 7.1 we can conclude that K-expectile, as an algorithm that generalize K-means, works as good as but sometimes even better than K-means on spherical clusters. Meanwhile it is better than all the other clustering algorithms, incluing Ward hierarchical clustering (hierarchical clustering with the metric of Euclidean distance), spectral clustering, and quantile based clustering. For asymmetric normal distributed clusters, as Table 7.2 shows, K-expectile outperforms all the listed algorithms. Since the contour lines of the real distribution of the data correspond to the assumption of K-expectile cluster shapes, K-expectile yields a significantly better result than other algorithms. For more general skewed distributed clusters as demonstrated in Table 7.2, 7.3, 7.4 and 7.5, K-expectile still has a robust and outstanding performance.

4 Application

4.1 Application of adaptive τ clustering on CRIX and VCRIX data

An application based on the CRIX and VCRIX data is presented in this section. CRIX (CRyptocurrency IndeX) developed by Trimborn and Härdle (2018) provides a CC market price index weighted by market capitalization with a dynamic changing number of constituents of representative cryptos. The mechanism of selecting CRIX constituents with Akaike Informstion Creterion is introduced in the mentioned paper. VCRIX, developed by Kim et al. (2019) is a volatility index built on CRIX which offers a forecast for the mean annualized volatility of the next 30 days, re-estimated daily by using Heterogeneous Auto-Regressive (HAR) model.

The data are downloaded from thecrix.de, consists of two time series, CRIX and VCRIX, collected daily from 2017-01-02 to 2021-02-09, in total 1497 observations in two dimensions. Here we scaled the data by dividing each varaible by their standard deviations to ensure the data has equal variance. The descriptive statistics and density plots of the two variables are listed as following.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	Skewness	Kurtosis	JB statistic
CRIX	0.080	0.621	1.056	1.246	1.443	7.257	2.450	10.988	5481.283
VCRIX	0.801	1.814	2.225	2.458	2.884	6.565	1.370	5.360	816.166

Table 4.1: Descriptive statistics of location and dispersion for 1497 scaled data for the period from January 02, 2017 to February 09, 2021.

From Table 4.1, it is evident that neither of the two variables are normally distributed and both of them are skewed. This fact can be seen in Figure 4.1 as well, due to the longer right tail of the densities of both variables. From the plot of marginal distribution one might suspect several clusters exist.

Results of K-means clustering, K-expectile clustering with adaptive τ and Spectral clustering are shown in Figure 6.3. To select the appropriate K, we consider the clustering evaluation cretaria including silhouette score and Davies-Bouldin score (Figure 6.1 and 6.2), both of them showed that $K = 3$ might be an appropiate choice which balances the cluster efficiency and number of clusters from a maximising similarity perspective. Hence we fix $K = 3$ and let the algorithm find the optimal location of the cluster center based on the skewness nature of the data. Our adaptive τ clustering approach yields a form of a $(K \times p)$ matrix $[[0.515, 0.448]^T, [0.222, 0.301]^T, [0.299, 0.300]^T]$, corresponding to the blue, green and grey clusters in Figure 6.3 respectively. For comparison, we consider K-means, K-expectile and spectral clustering.

Figure 2.3 shows the shapes and distribution of the three clusters on the two dimensional space consisting of CRIX and VCRIX. From the plot we can observe that the three clusters of K-expectile represents different types of correlation between price and volatility index. The three clusters can be described as 'low-price-low-volatility cluster', 'low-price-high-volatility cluster', and 'positively correlated price and volatility cluster', corresponding to color blue, green and grey.

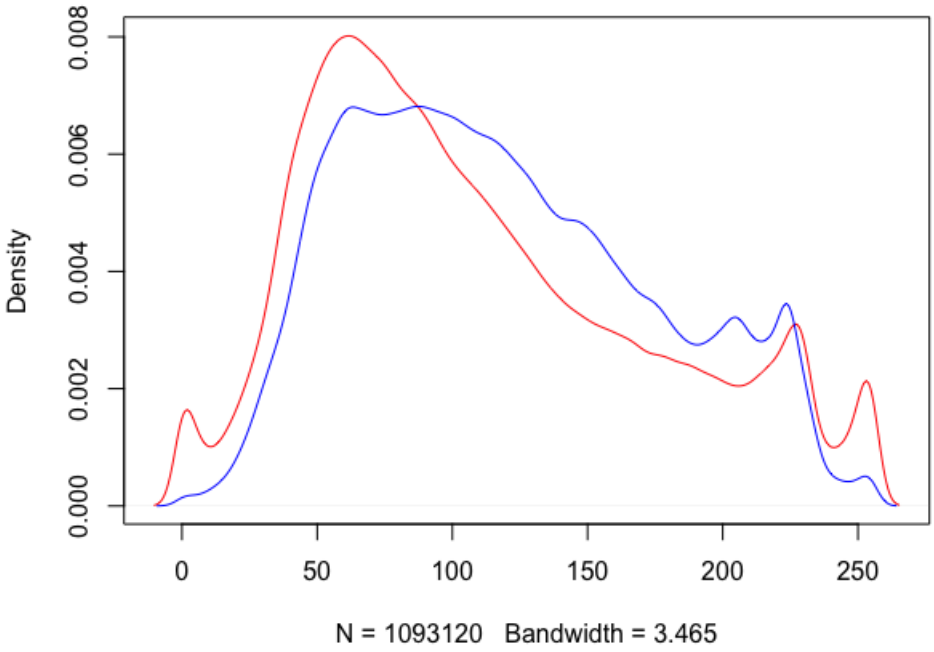

Figure 4.1: Variable densities. The red line and the blue line are the result of kernel density estimation of scaled CRIX and VCRIX.

QKEC_application

It is worth noting that the grey cluster only appears shortly in the end of 2017 and from the end of 2020 till now. Positive correlation between price and volatility of crypto markets means that the volatility and price drives each other in the same direction. Higher price and higher volatility shows an 'exciting' signal other than a 'panic' expectation, this phenomenon mostly occurs in the securities market dominated by individual investors, where increased volatility is a signal of market activation. On the other hand, low-volatility cluster appears in most period of the CC market, which means CC market is highly dominated by instituational investors most of the time. High volatility means unstable market sentiment and high trading volumn, and the green cluster often appears when the price start to change.

4.2 Application of fixed τ clustering on image segmentation

Image segmentation is a technique widely used in image processing, which partition an image into multiple parts sharing similar characteristics. Image segmentation includes separating foreground from background, or clustering regions of pixels based on color or shape. One of the commenly used methods in color-based segmentation is K-means clustering. In this case, pixel values are regarded as independent random samples in the 8 bits color space, and divided into K discrete regions which has minimal variances. The output of K- means segmentation can be visualized by converting all the pixels in a group to the color of the cluster centers.

By applying K-expectile clustering, we expect a more flexible choice of centers and thus a more flexible segmentation output of the image. Besides the adaptive τ clustering approach, we also consider two ways to prespecified τ, which put more weights on group-wised and dimension-wised tail behavior respectively. In particular, we set a group-specified τ as $[0.2, 0.7, 0.1, 0.9]$ to include groups emphasizes on both left tail and right tail for a choice of $K = 4$ groups. We also consider a dimension-specified τ as $[0.1, 0.8, 0.9]$ for these 3-dimensional RGB valued data to involve information on both tails. To evaluate the performance of K-expectile clustering, we take K-means clustering result as benchmark and bring Quantile based clustering results into comparison.

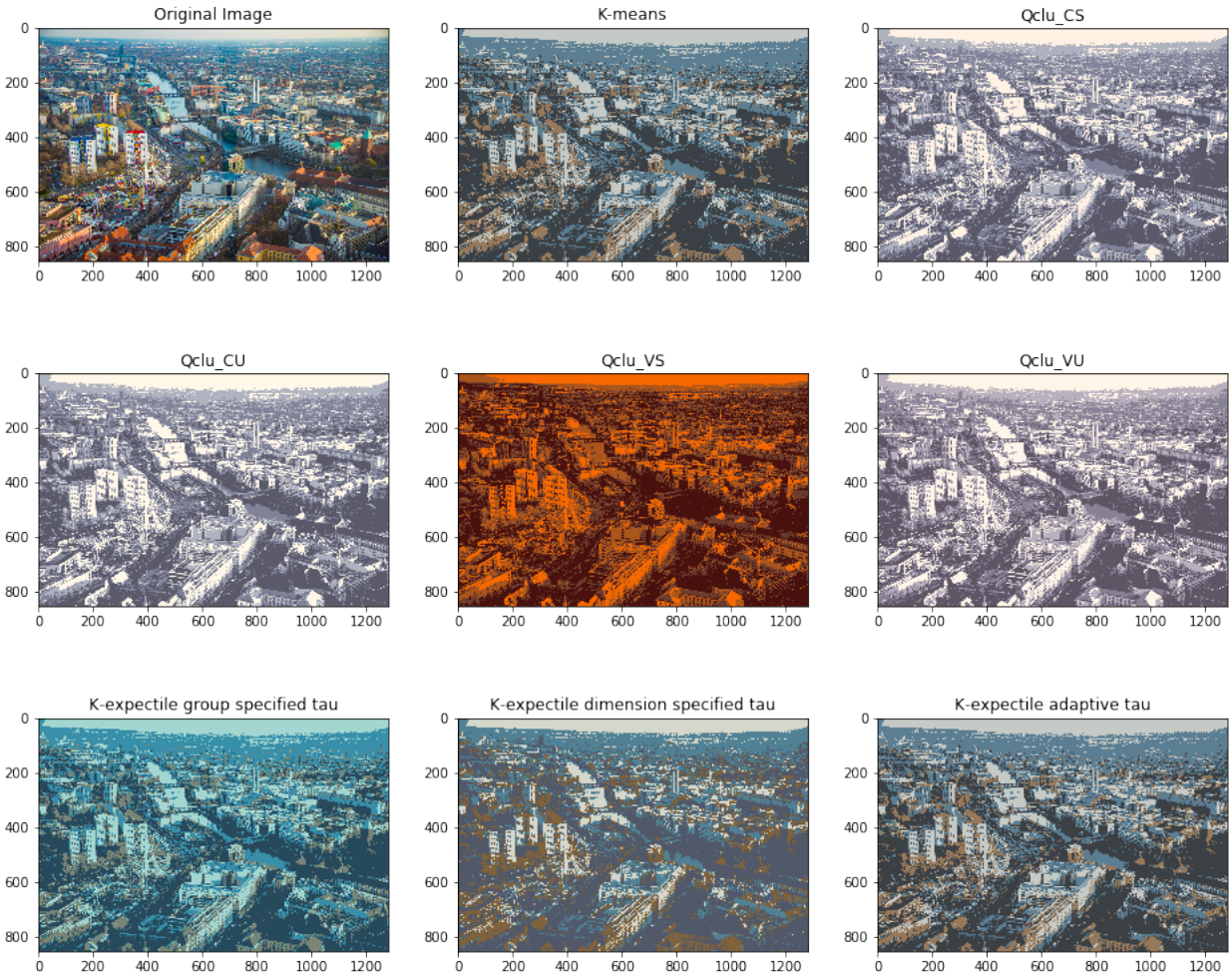

Figure 4.2: Image segmentation results using different clustering methods.

The original image is an aerial photo of Berlin, for pre-processing, we transform the image into pixel values in RGB color space, and flatten the data into a two-dimensional array. Figure 4.2 shows the original image and the segmented image, which can be considered as filtered image with 4 color clusters. Important information can be extracted from the image by displaying some clusters and mute others. The subplots showed in Figure 4.3 are image with only one cluster enabled.

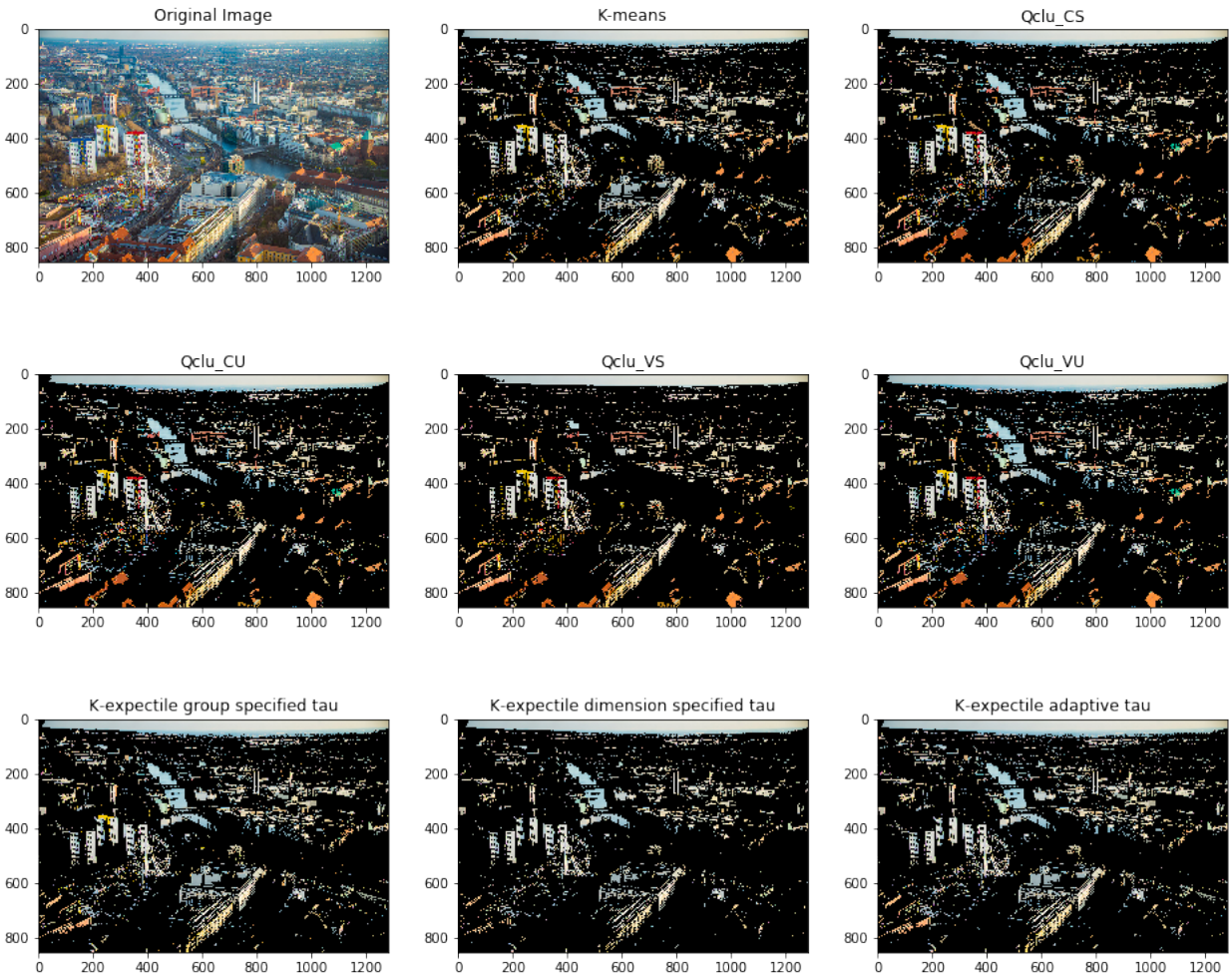

Figure 4.3: Segmented image with only one cluster displayed.

To evaluate the performance of segmentation methods, we use two indices, Mean Square Error (MSE) and Peak to Signal Noise Ration (PSNR). Given an $m \times n$ monochrome image I, Mean Square Error measures how much the approximation K differs from it. MSE is defined as:

$$MSE = \frac{1}{mn} \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \{I(i,j) - K(i,j)\}^2.$$

Peak to Signal Noise Ration is usually used to measure the quality of the compressed image. $PSNR$ is the proportion between maximum attainable powers and the corrupting noise that influence likeness of image. It is defined as following:

$$PSNR = 10 \log_{10}(\frac{MAX_I^2}{MSE}),$$

where MAX_I is the maximum possible pixel value of the image I, which equals to 225 when the sample is in 8 bits. The higher value of $PSNR$ and the lower value of MSE, the better the fitting of the approximated image.

Table 4.2 shows the MSE and $PSNR$ values of segmented image using multiple methods. Although MSE is usually calculated on monochrome data, here we take the average MSE on three RGB dimensions. Moreover, we convert both the original image and the segmented image from RGB data into GRAY and YCrCb color space. From the table, it can be concluded that the pre-specified parameter τ scheme, in both senario, gives us some flexibility to customize segmentation, but the adaptive τ scheme performs better.

	GREY		YCrCb		RGB	
	MSE	PSNR	MSE	PSNR	MSE	PSNR
K-means	509.18	21.06	839.12	18.89	742.53	28.96
K-expectiles_vtau	429.47	21.80	835.66	18.91	741.17	28.97
CS	2001.30	15.12	2886.34	13.52	2841.61	23.14
CU	2217.68	14.67	3546.76	12.63	3639.46	22.06
VS	5338.85	10.86	2799.57	13.66	2398.98	23.87
VU	2030.67	15.05	3332.15	12.90	3507.04	22.22
K-expectiles_gp_spec_tau	519.81	20.97	1449.73	16.51	1203.65	26.87
K-expectiles_dim_spec_tau	876.31	18.70	1304.86	16.97	1214.55	26.83

Table 4.2: Performance of different clustering methods on image segmentation. Data is transformed into RGB, GREY and YCrCb space. Algorithms listed from top to bottom are K-means, K-expectile with adaptive τ, four modes of Quantile based clustering: Common skewness parameter and Scaled variable-wise, Common skewness parameter and Unscaled variables, Variable-wise skewness parameter and Scaled variable-wise, Variable-wise skewness parameter and Unscaled variables, K-expectile with group-specified τ and K-expectile with dimension-specified τ.

References

Aigner, D. J., T. Amemiya, and D. J. Poirier (1976): "On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function," *International Economic Review*, 377–396.

Deisenroth, M. P., A. A. Faisal, and C. S. Ong (2020): *Mathematics for machine learning*, Cambridge University Press.

Hartigan, J. A. (1975): *Clustering algorithms*, John Wiley & Sons, Inc.

Hennig, C., C. Viroli, L. Anderlucci, et al. (2019): "Quantile-based clustering," *Electronic Journal of Statistics*, 13, 4849–4883.

Kim, A., S. Trimborn, and W. K. Härdle (2019): "VCRIX-A Volatility Index for Crypto-Currencies," *Available at SSRN 3480348*.

Koenker, R. and G. Bassett Jr (1978): "Regression quantiles," *Econometrica: journal of the Econometric Society*, 33–50.

Kuan, C.-M., J.-H. Yeh, and Y.-C. Hsu (2009): "Assessing value at risk with CARE, the conditional autoregressive expectile models," *Journal of Econometrics*, 150, 261–270.

MacKay, D. J. and D. J. Mac Kay (2003): *Information theory, inference and learning algorithms*, Cambridge university press.

Maume-Deschamps, V., D. Rullière, and K. Said (2016): "Multivariate extensions of expectiles risk measures," *arXiv preprint arXiv:1609.07637*.

Newey, W. K. and J. L. Powell (1987): "Asymmetric least squares estimation and testing," *Econometrica: Journal of the Econometric Society*, 819–847.

Schnabel, S. (2011): *Expectile smoothing: new perspectives on asymmetric least squares. An application to life expectancy*, Utrecht University.

Shi, J. and J. Malik (2000): "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, 22, 888–905.

Sobotka, F. and T. Kneib (2012): "Geoadditive expectile regression," *Computational Statistics & Data Analysis*, 56, 755–767.

Steinhaus, H. (1956): "Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV: 801-804," .

Tran, N. M., P. Burdejová, M. Ospienko, and W. K. Härdle (2019): "Principal component analysis in an asymmetric norm," *Journal of Multivariate Analysis*, 171, 1–21.

Trimborn, S. and W. K. Härdle (2018): "CRIX an Index for cryptocurrencies," *Journal of Empirical Finance*, 49, 107–122.

Ward Jr, J. H. (1963): "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, 58, 236–244.

Yang, Y., T. Zhang, and H. Zou (2018): "Flexible expectile regression in reproducing kernel Hilbert spaces," *Technometrics*, 60, 26–35.

Zhang, Y., H. J. Wang, and Z. Zhu (2019): "Quantile-regression-based clustering for panel data," *Journal of Econometrics*, 213, 54–67.

Ziegel, J. F. (2016): "Coherence and elicitability," *Mathematical Finance*, 26, 901–918.

5 APPENDIX

Proof of Proposition 1: In order to proof the convergence of the algorithm, we first recall the objective function:

$$\begin{aligned} G^{Adaptive}(\tau, \Theta, C, X) &= \sum_{k=1}^{K} \sum_{x_i \in G_k} \sum_{j=1}^{p} d(x_{ij}, \tau_{k,j}, \theta_{k,j}) \\ &= \sum_{k=1}^{K} \sum_{x_i \in G_k} \sum_{j=1}^{p} \left\{ \tau_{k,j} + (1 - 2\tau_{k,j}) \, \mathbf{I}_{\{x_{ij} < \theta_{k,j}\}} \right\} (x_{ij} - \theta_{k,j})^2. \end{aligned}$$

Then define

$$\hat{\theta}_{C(i),j} = \arg\min_{\theta_j} \sum_{C(i)=k} \sum_{j=1}^{p} \left\{ \tau_{C(i),j} + (1 - 2\tau_{C(i),j}) \, \mathbf{I}_{\{x_{ij} < \theta_j\}} \right\} (x_{ij} - \theta_j)^2,$$

$$\hat{\tau}_{C(i),j} = \arg\min_{\tau_j} \sum_{C(i)=k} \sum_{j=1}^{p} \left\{ \tau_j + (1 - 2\tau_j) \, \mathbf{I}_{\{x_{ij} < \theta_{C(i),j}\}} \right\} (x_{ij} - \theta_{C(i),j})^2.$$

Let $C_{(i)}^{(t-1)}$ be the previous partition, $\hat{\theta}_{k,j}^{(t-1)}$ and $\hat{\tau}_{k,j}^{(t-1)}$ be previous estimated centroid and τ parameters, $C_{(i)}^{(t)}$ be the new partition,

$$G(C_{(i)}^{(t)}) \leq \sum_{k=1}^{K} \sum_{C_{(i)}^{(t)}=k} \sum_{j=1}^{p} \left\{ \hat{\tau}_{k,j}^{(t-1)} + (1 - 2\hat{\tau}_{k,j}^{(t-1)}) \, \mathbf{I}_{\{x_{ij} < \theta_j\}} \right\} (x_{ij} - \hat{\theta}_{k,j}^{(t-1)})^2.$$

New partition $C_{(i)}^{(t)}$ minimises $\sum_{k=1}^{K} \sum_{C(i)=k} \sum_{j=1}^{p} \left\{ \hat{\tau}_{k,j}^{(t-1)} + (1 - 2\hat{\tau}_{k,j}^{(t-1)}) \, \mathbf{I}_{\{x_{ij} < \theta_j\}} \right\} (x_{ij} - \hat{\theta}_{k,j}^{(t-1)})^2$ over all possible partitions:

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{C_{(i)}^{(t)}=k} \sum_{j=1}^{p} \left\{ \hat{\tau}_{k,j}^{(t-1)} + (1 - 2\hat{\tau}_{k,j}^{(t-1)}) \, \mathbf{I}_{\{x_{ij} < \theta_j\}} \right\} (x_{ij} - \hat{\theta}_{k,j}^{(t-1)})^2 \\ \leq \underbrace{\sum_{k=1}^{K} \sum_{C_{(i)}^{(t-1)}=k} \sum_{j=1}^{p} \left\{ \hat{\tau}_{k,j}^{(t-1)} + (1 - 2\hat{\tau}_{k,j}^{(t-1)}) \, \mathbf{I}_{\{x_{ij} < \theta_j\}} \right\} (x_{ij} - \hat{\theta}_{k,j}^{(t-1)})^2}_{G(C_{(i)}^{(t-1)})}. \end{aligned}$$

Therefore, $G(C_{(i)}^{(t)}) \leq G(C_{(i)}^{(t-1)})$. As $G(C_{(i)}^{(t)}) - G(C_{(i)}^{(t-1)})$ is a monotonically nonincreasing sequence that converges to 0, we conclude that the limit point obtained from the adaptive τ clustering algorithm is the optimal solution.

6 Figures

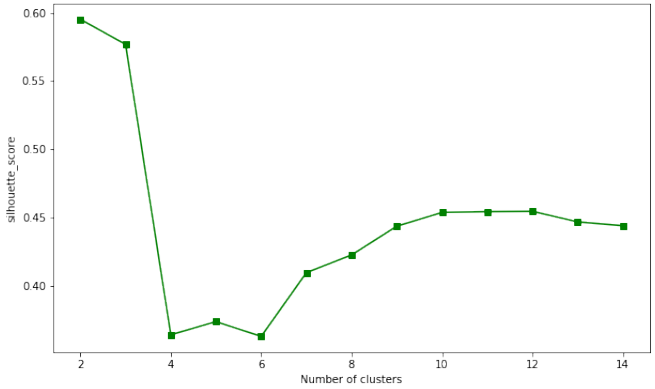

Figure 6.1: Sillouette score of K-expectiles clustering results with different number of clusters
QKEC_application

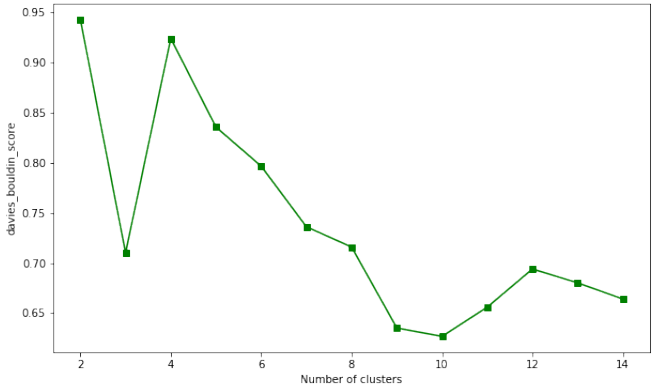

Figure 6.2: Davies-Bouldin scoreof K-expectiles clustering results with different number of clusters
QKEC_application

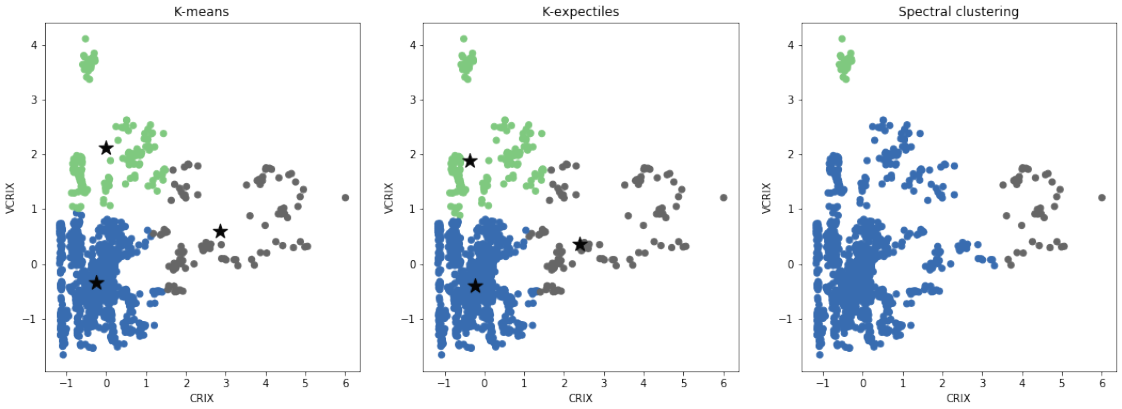

Figure 6.3: Clustering results of K-means, K-expectiles and Spectral clustering. The two variables are plot along x-axis and y-axis. Clusters are shown in different colours, whereas cluster centroids are shown by stars.
QKEC_application

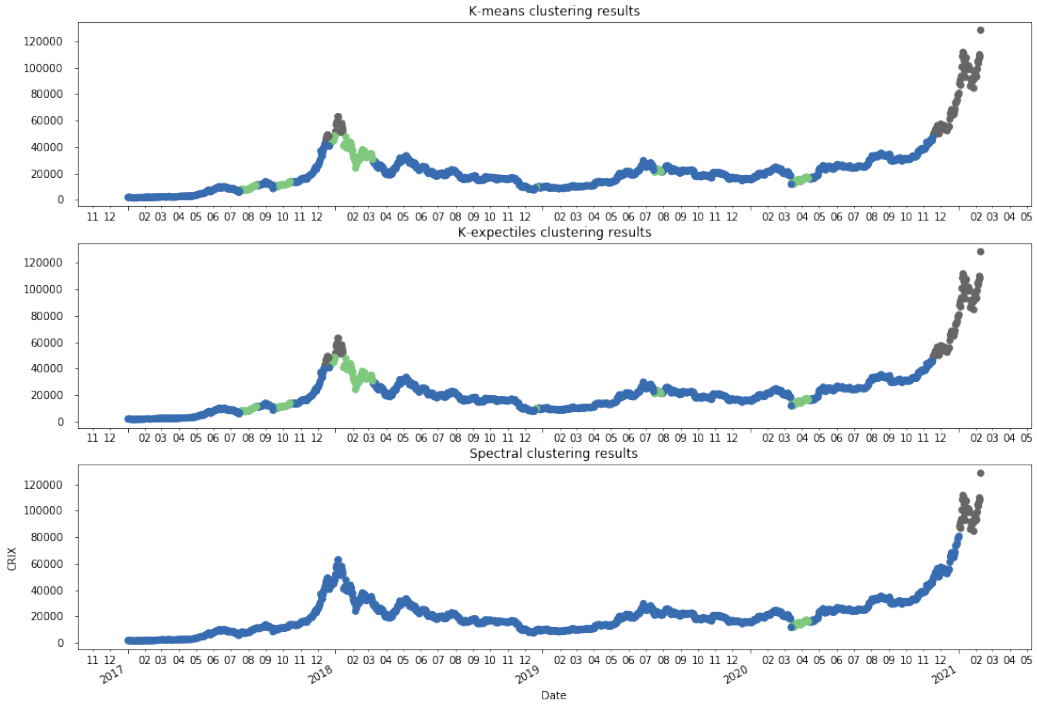

Figure 6.4: Clustering results of K-means, K-expectiles and Spectral clustering on CRIX .Clusters are shown in different colours.
QKEC_application

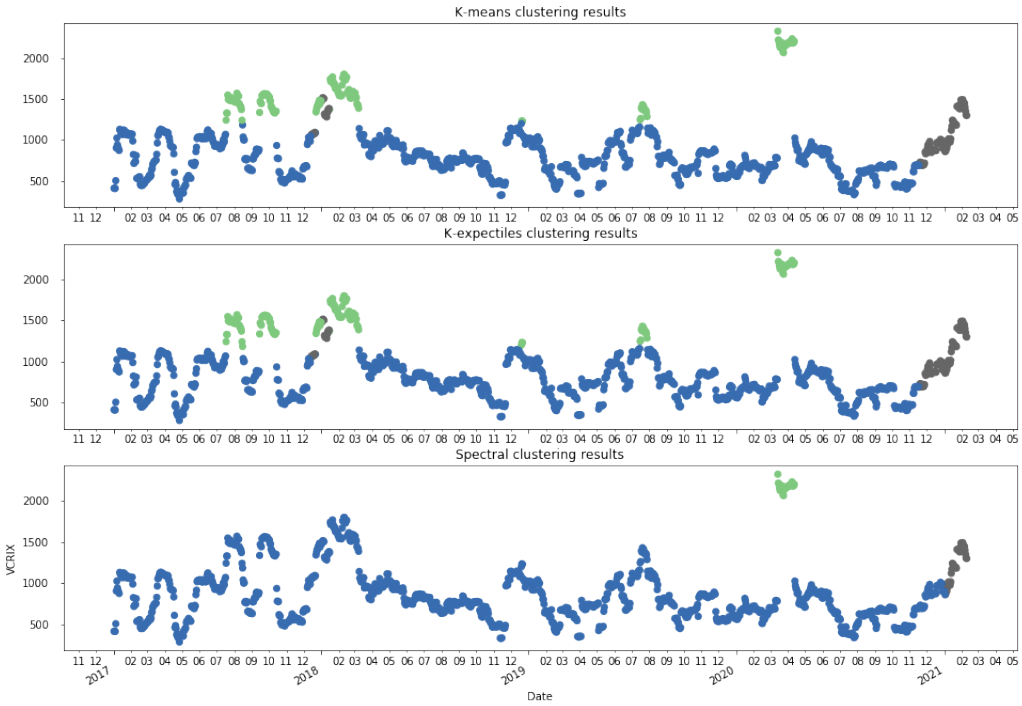

Figure 6.5: Clustering results of K-means, K-expectiles and Spectral clustering on VCRIX .Clusters are shown in different colours.

QKEC_application

7 Tables

For all tables below, algorithms listed from top to bottom are K-expectile with adaptive τ, K-means, Spectral clustering, Ward hierarchical clustering and four modes of Quantile based clustering: Common skewness parameter and Scaled variable-wise, Common skewness parameter and Unscaled variables, Variable-wise skewness parameter and Scaled variable-wise, Variable-wise skewness parameter and Unscaled variables, K-expectile with group-specified τ and K-expectile with dimension-specified τ. We report the adjusted rand index (ARI).

Table 7.1: Sample 1: Simulation results of Gaussian clusters

QKEC_simulations

	n =1500			n = 300		
	p=10 ARI	p=50 ARI	p=100 ARI	p=10 ARI	p=50 ARI	p=100 ARI
K-expectiles_vtau	99.36	99.60	99.87	97.00	97.99	97.99
K-means	99.36	99.60	99.60	97.00	97.99	97.99
Spectral		31.22	86.74		27.48	85.03
h-ward	99.20	99.60	99.87	93.54	97.99	97.99
CS	99.24	99.60	99.87	96.61	97.99	97.99
CU	99.24	99.60	99.87	96.61	97.99	97.99
VS	99.28	99.60	99.87	96.03	97.99	97.99
VU	99.20	99.60	99.87	96.61	97.99	97.99

Table 7.2: Sample 2: Simulation results of Asymmetric normal clusters

QKEC_simulations

	n =1500			n = 300		
	p=10 ARI	p=50 ARI	p=100 ARI	p=10 ARI	p=50 ARI	p=100 ARI
K-expectiles_vtau	93.22	99.60	99.60	92.20	97.99	97.99
K-means	91.19	99.59	99.60	81.70	97.99	97.99
Spectral		-0.02				
h-ward	77.19	99.52	99.60	76.98	97.01	97.99
CS	86.61	99.58	99.60	88.98	97.99	71.74
CU	80.73	99.28	94.70	72.76	93.36	76.27
VS	88.86	99.59	99.57	93.16	91.71	97.99
VU	85.74	99.55	99.60	80.41	93.73	97.99

Table 7.3: Sample 3: Simulation results of *Beta*-distributed clusters

KEC_simulations

	n =1500			n = 300		
	p=10 ARI	p=50 ARI	p=100 ARI	p=10 ARI	p=50 ARI	p=100 ARI
K-expectiles_vtau	94.04	99.60	99.60	93.17	97.99	97.99
K-means	94.79	99.60	99.60	93.16	97.99	97.99
Spectral	93.63	99.60	99.60	93.17		
h-ward	94.80	99.60	99.60	88.88	97.99	97.99
CS	68.92	96.89	82.52	92.20	97.99	65.54
CU	68.14	94.08	73.26	92.17	93.36	65.62
VS	93.28	97.47	79.84	91.98	91.71	62.46
VU	94.03	94.69	73.28	92.56	93.73	46.90

Table 7.4: Sample 4-1: Simulation results of generalized t-distributed clusters

KEC_simulations

	n =1500			n = 300		
	p=2 ARI	p=10 ARI	p=50 ARI	p=2 ARI	p=10 ARI	p=50 ARI
K-expectiles_vtau	96.50	97.99	97.99	95.10	97.99	97.99
K-means	96.26	97.99	97.99	94.80	97.99	97.99
Spectral	96.21	26.31		94.81	93.10	92.43
h-ward	96.24	97.99	97.99	94.29	97.99	97.99
CS	96.48	97.99	97.99	95.01	97.99	97.99
CU	96.08	97.99	97.99	94.57	97.99	97.99
VS	96.48	97.99	97.99	95.10	97.99	97.99
VU	96.07	97.99	97.99	94.57	97.99	97.99

Table 7.5: Sample 4-2: Simulation results of F-distributed clusters

KEC_simulations

	n =1500			n = 300		
	p=2 ARI	p=10 ARI	p=50 ARI	p=2 ARI	p=10 ARI	p=50 ARI
K-expectiles_vtau	95.80	99.60	99.60	94.58	99.60	99.60
K-means	95.19	99.60	99.60	94.01	99.60	99.60
Spectral	94.89	26.31		93.82		
h-ward	96.82	99.60	99.60	95.25	99.60	99.60
CS	97.96	99.60	99.60	96.03	99.60	99.60
CU	95.42	99.60	99.60	94.19	99.60	99.60
VS	97.72	99.60	99.60	95.64	99.60	99.60
VU	95.44	99.60	99.60	94.19	99.60	99.60

IRTG 1792 Discussion Paper Series 2021

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 "Surrogate Models for Optimization of Dynamical Systems" by Kainat Khowaja, Mykhaylo Shcherbatyy, Wolfgang Karl Härdle, January 2021.

002 "FRM Financial Risk Meter for Emerging Markets" by Souhir Ben Amor, Michael Althof, Wolfgang Karl Härdle, February 2021.

003 "K-expectiles clustering" by Bingling Wang, Yingxing Li, Wolfgang Karl Härdle, March 2021.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.