K-expectile clustering

Bingling Wang
Wolfgang Karl Härdle
Yingxing Li

Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
BRC Blockchain Research Center
lvb.wiwi.hu-berlin.de
Charles University, WISE XMU, NCTU 玉山学者

Clustering: a mega topic

- Marketing
  - Marketing research aims to discover distinct customer groups to develop targeting strategies

Clustering: a mega topic

- Computer vision
  - Image segmentation (pattern extraction)

Original

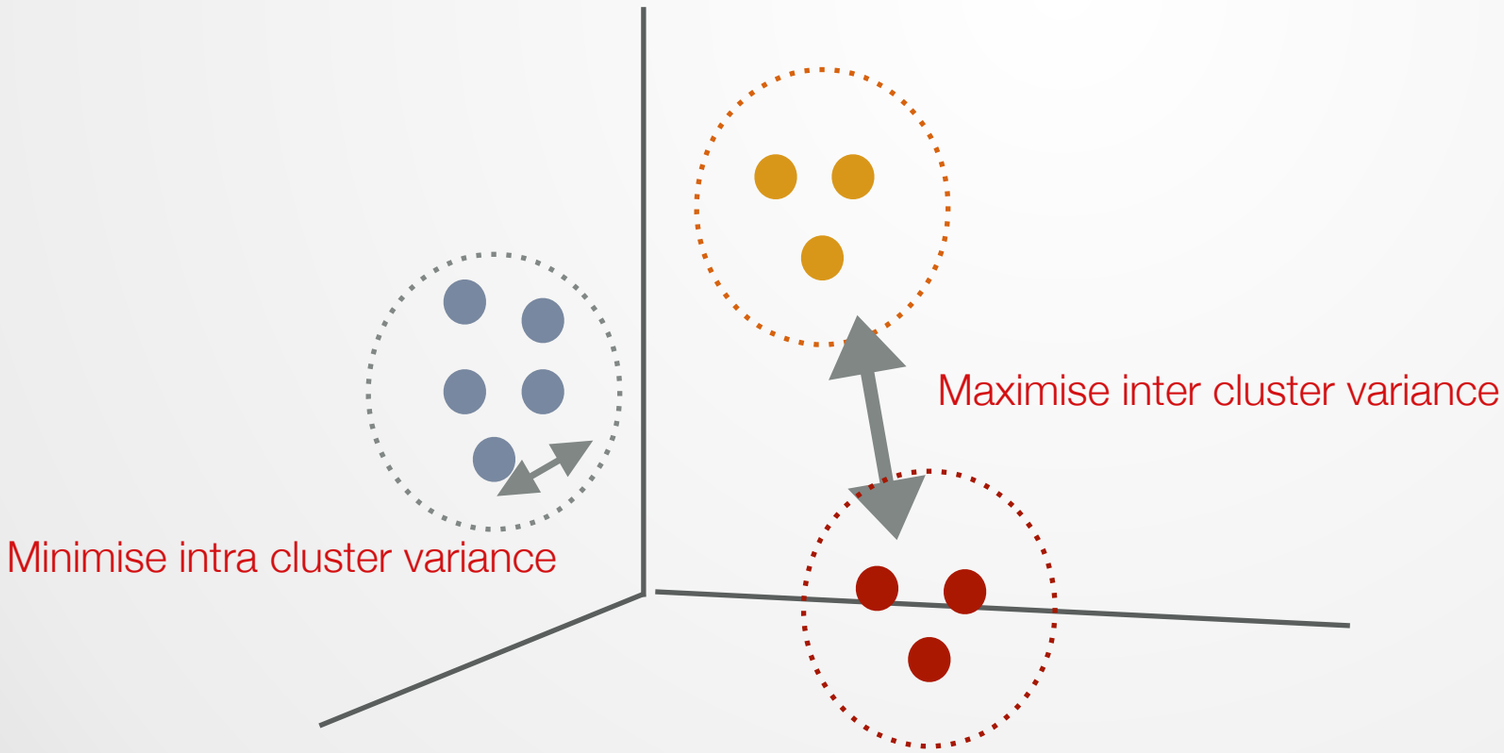
2 Clusters

4 Clusters

8 Clusters

K-means clustering

- Group observations into clusters
- Minimise intra cluster variance (around the cluster mean)
- Maximise inter cluster variance (around the overall mean)



K-means clustering

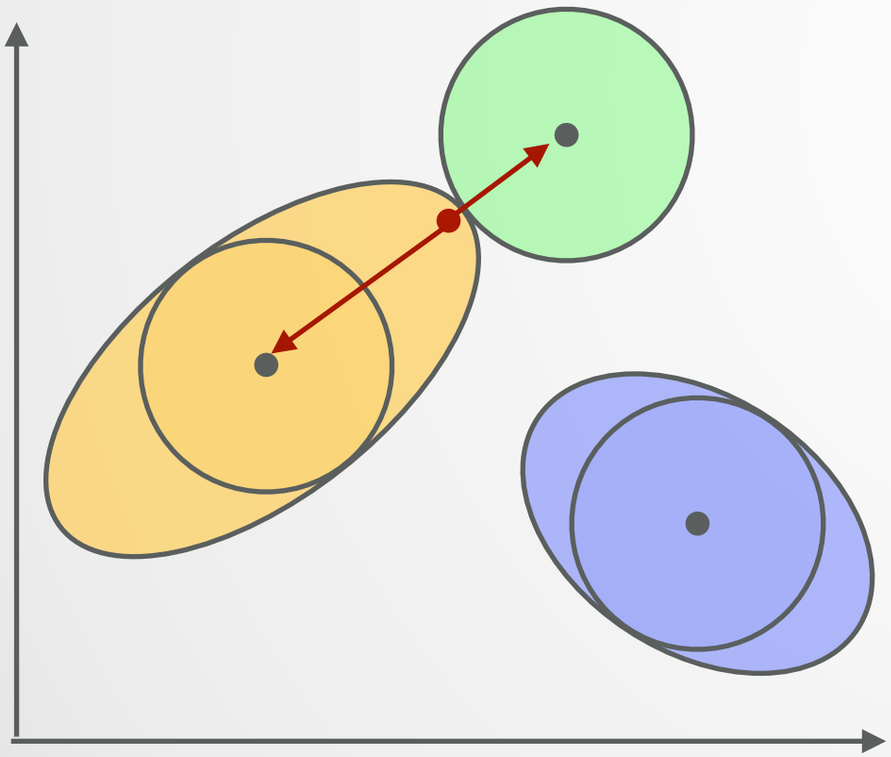
- Partition of dataset $X = \{X_i\}_{i=1}^{n}$, $X_i \in \mathbb{R}^p$ into K clusters, $\{G_1, G_2, \ldots, G_K\}$, recorded in a membership vector $C = (c(1), c(2), \ldots, c(n))$, $c(i) \in \{1,2,\ldots, K\}$
- Minimise the discrepancy over $\underset{K}{\Theta} = (\theta_1, \ldots, \theta_K)$:

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

- First step: for fixed centre, assign each point to the nearest cluster centre.
- Second step: for fixed clusters, estimate cluster centre $\hat{\theta}_k$. The centres are the within cluster means, $\hat{\theta}_k = \bar{x}_k = \frac{1}{|G_k|} \sum_{x_i \in G_k} x_i$.
- Iterate until convergence

Assumptions of K-means clustering

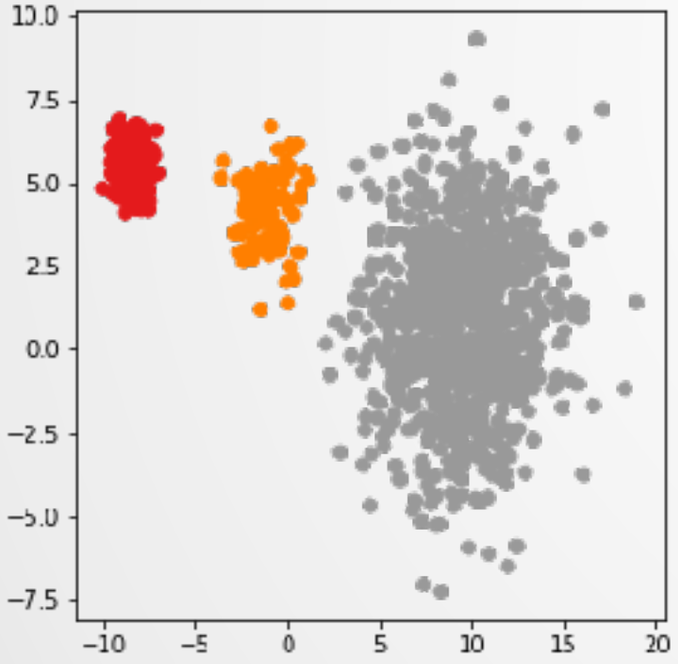
- Balanced cluster size within the dataset
- Spherical cluster shapes (joint distribution has equal variance, independent)
- Clusters have similar density



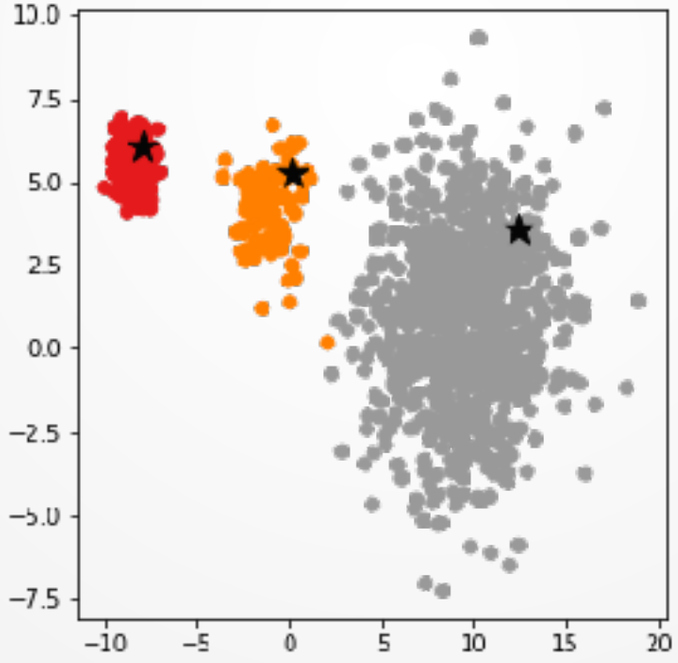
Advantages of K-means clustering

- Works ok even under weird data
- Easy to interpret in a “normal” world
- Fast and efficient in terms of computational cost

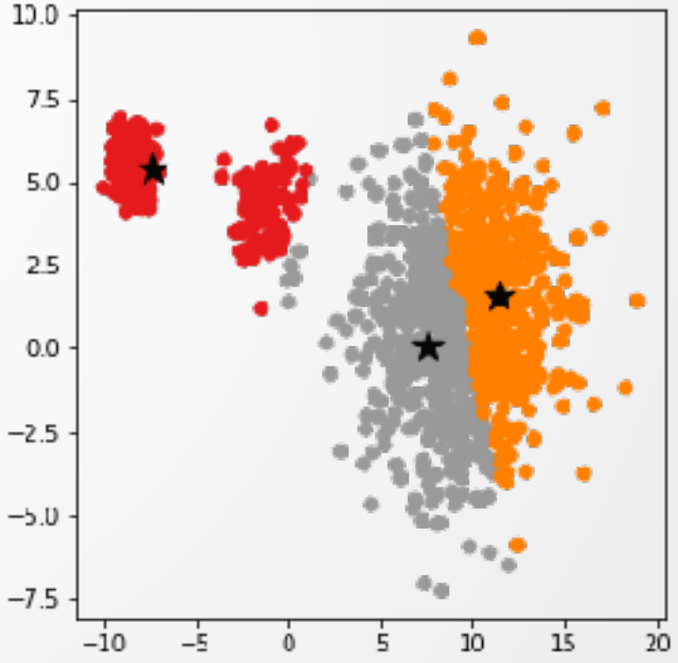
Cluster size: 900, 100, 500 std: 2.5, 1, 0.5 K=3



True cluster



K-expectile τ= 0.05
Accuracy = 0.99933



K-means
Accuracy = 0.64067

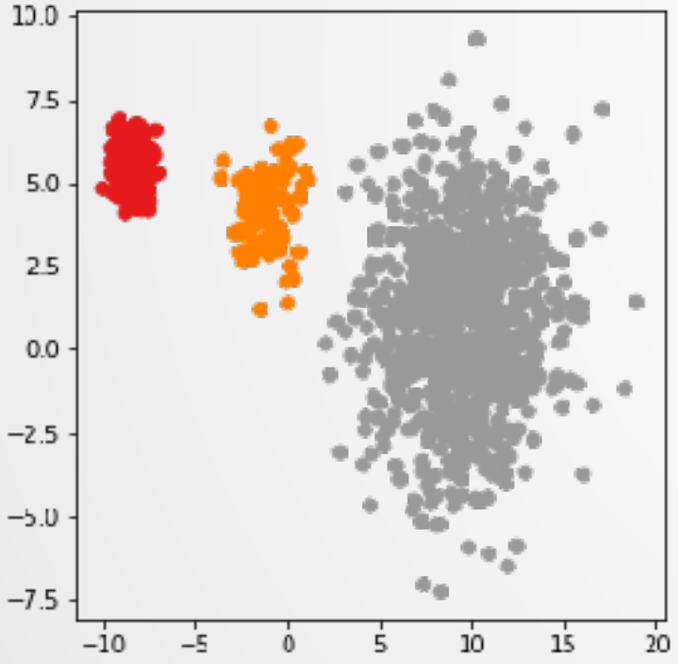
Research questions

Improve K-means clustering?

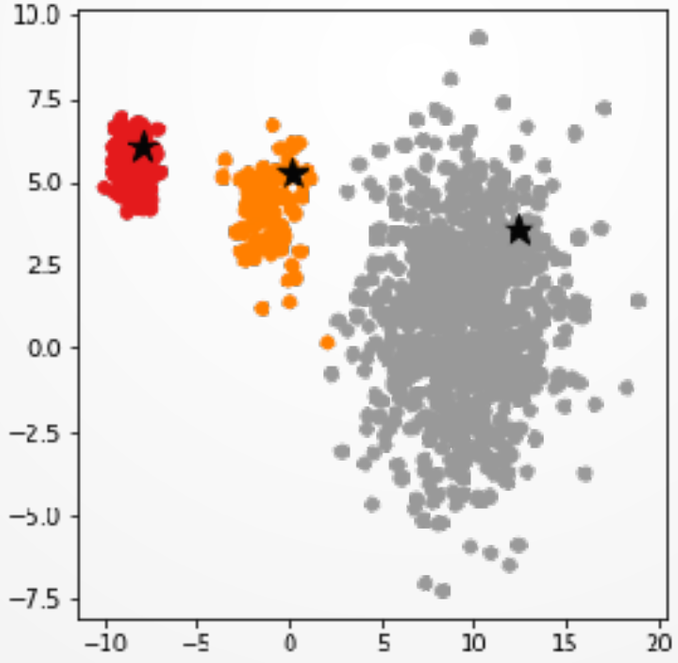
Define a proper measure for multi-dimensional expectiles?

Comparison with quantile based clustering?

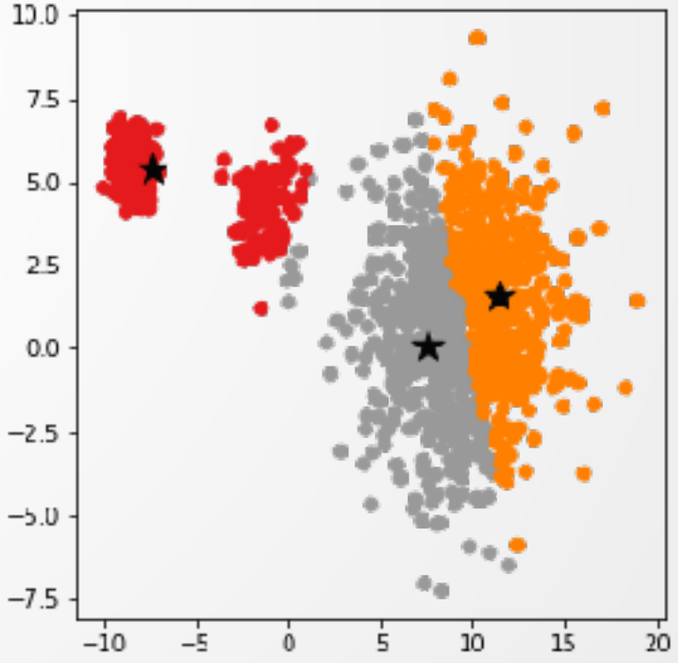
Cluster size: 900, 100, 500 std: 2.5, 1, 0.5



True cluster



K-expectile τ= 0.05
Accuracy = 0.99933



K-means
Accuracy = 0.64067

Outline

1. Motivation. ✔
2. Expectiles
3. K Expectile Clustering
4. Simulation
5. Real (nice) examples
6. What else we can cluster?

Quantiles and Expectiles

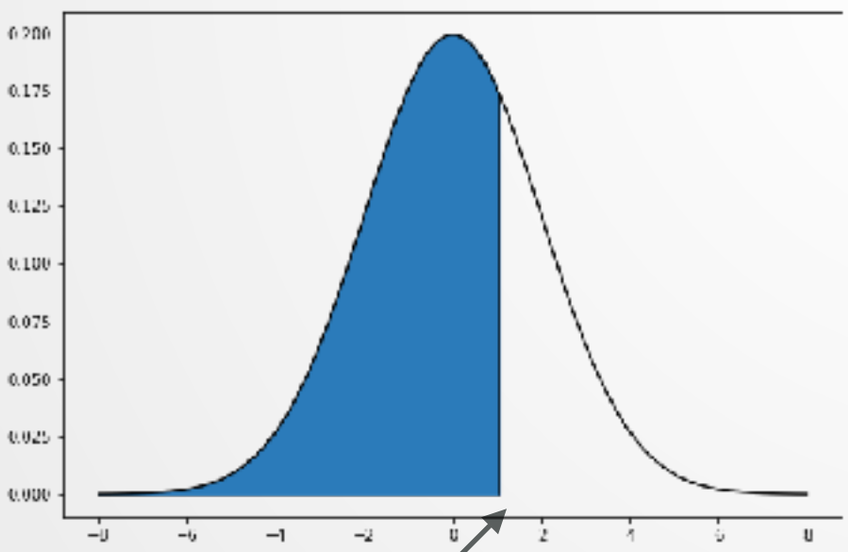
For rv X obtain tail event measure:

$$q_\tau = \arg\min_{\mu} \mathrm{E}\left[\rho_\tau(X-\mu)\right]$$

asymmetric loss function

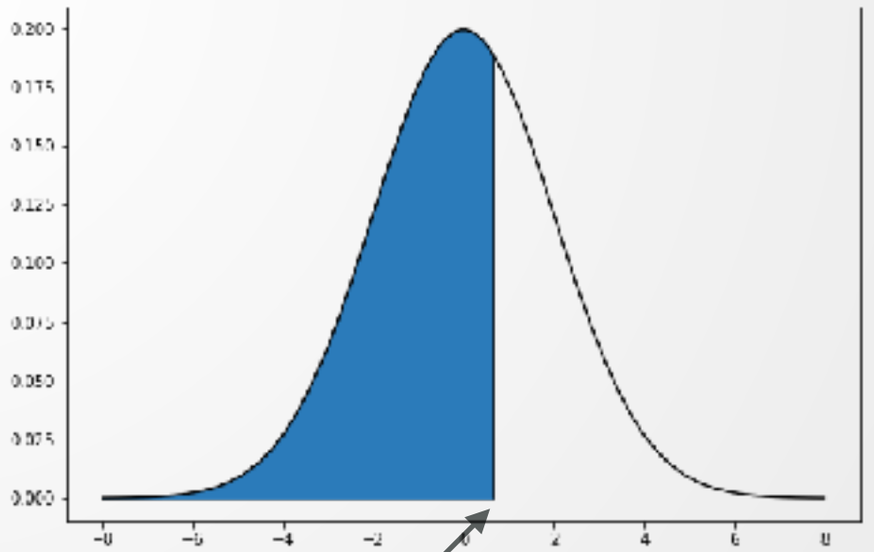
$$\rho_\tau(u) = |u|^\alpha \left|\tau - \mathbf{I}_{\{u<0\}}\right|$$

$\alpha = 1$ quantiles



Quantile = 1.05 $\tau = 0.7$

$\alpha = 2$ expectiles



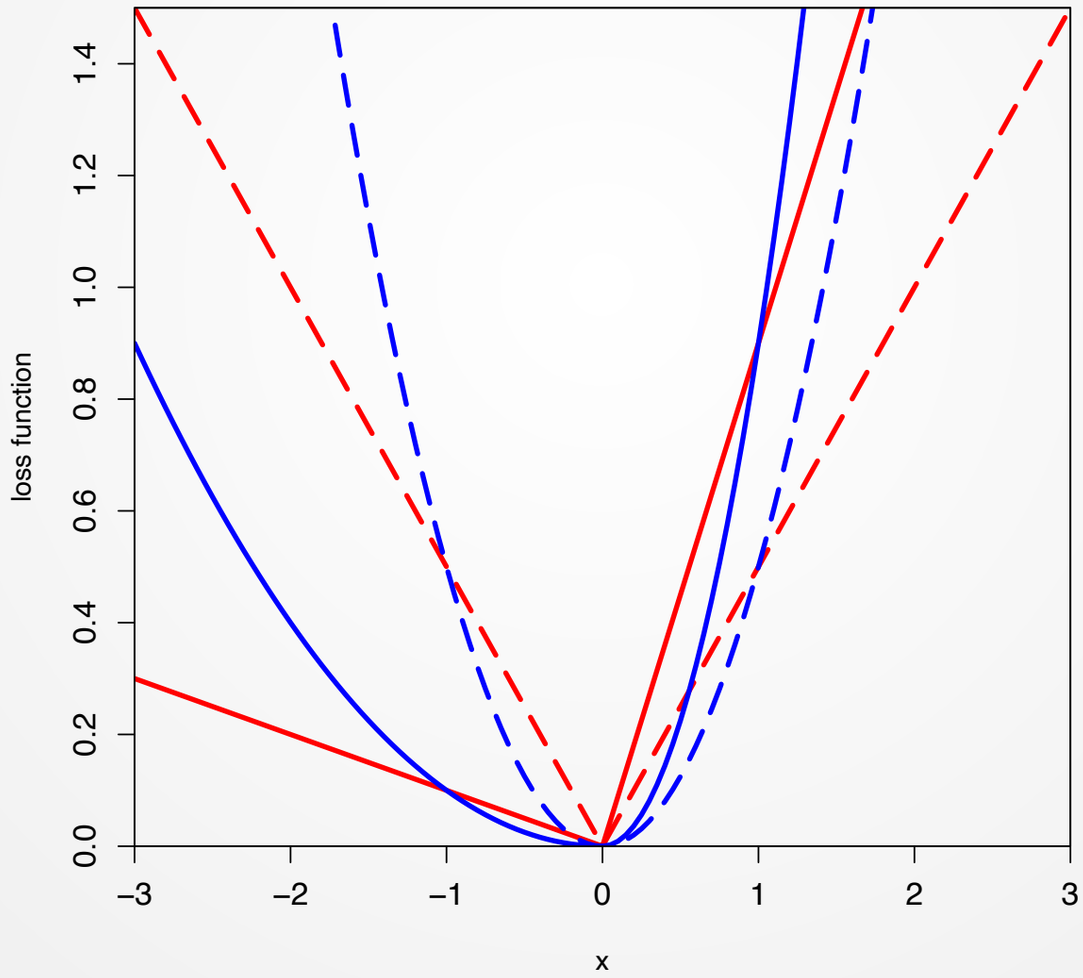
Expectile =0.66 $\tau = 0.7$

sample $N(0,4)$

motivated by PB

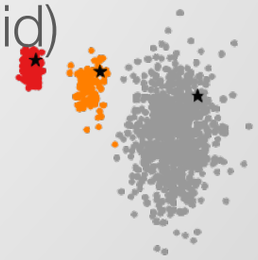
Quantiles and Expectiles

- Quantiles/Expectiles focus on tail events
- Expectiles loss function is differentiable
- LAWS algorithm fast and efficient



LQRcheck

Figure: Loss function of expectiles and quantiles for $\tau = 0.5$ (dashed) and $\tau = 0.9$ (solid)



Univariate expectiles

- Newey and Powell (1987): for univariate rv X and $\mu \in \mathbb{R}$

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

- Loss function (with $(x)_+ = \max(x,0)$)

$$\rho_\tau(X-\mu) = \tau(X-\mu)_+^2 + (1-\tau)(\mu-X)_+^2$$

- Unique Solution (via F.O.C.)

$$\tau\,\mathrm{E}[(X-\mu)_+] = (1-\tau)\mathrm{E}[(\mu-X)_+]$$

$$\frac{1-\tau}{\tau} = \frac{\mathrm{E}[(X-\mu)_+]}{\mathrm{E}[(\mu-X)_+]} = \frac{\int_\mu^{+\infty} |X-\mu| f(x)dx}{\int_{-\infty}^{\mu} |X-\mu| f(x)dx}$$

Multivariate expectiles

- Maume-Deschamps et al (2017): for rv X and $\mu \in \mathbb{R}^p$
- Define $(X)_+ = ((X_1)_+, \ldots, (X_p)_+)^\top$, and $\|\cdot\|$ a norm on $\mathbb{R}^p$
- Loss function

$$\rho_\tau(X-\mu) = \tau \|(X-\mu)_+\|^2 + (1-\tau)\|(\mu - X)_+\|^2$$

- Multivariate expectile

$$e_\tau(X) = \arg\min_{\mu \in \mathbb{R}^p} \mathrm{E}\left[\rho_\tau(X-\mu)\right]$$

- $\mathrm{E}\left[\rho_\tau(X-\mu)\right]$ is strictly convex in μ

Multivariate expectiles

- Maume-Deschamps et al (2017): multivariate expectile is constructed via marginal univariate expectiles

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

$$= \arg\min_{\mu\in\mathbb{R}^p} \mathrm{E}\left[\sum_{j=1}^p \{\tau(x_j-\mu_j)_+^2 + (1-\tau)(\mu_j-x_j)_+^2\}\right]$$

$$= (e_\tau(X_1), \ldots, e_\tau(X_p))^\top$$

- Now let $\tau = (\tau_1, \ldots, \tau_p)^\top \in \mathbb{R}^p$

$$e_\tau(X) = \arg\min_{\mu\in\mathbb{R}^p} \mathrm{E}\left[\sum_{j=1}^p \{\tau_j(x_j-\mu_j)_+^2 + (1-\tau_j)(\mu_j-x_j)_+^2\}\right]$$

$$= (e_{\tau_1}(X_1), \ldots, e_{\tau_p}(X_p))^\top$$

LAWS estimation (for marginal expectile)

Newey and Powell (1987): $(x_1, \ldots, x_n)^\top \in (\mathbb{R}^p)^n$; $(\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$;

$$(\hat{e}_{\tau_1,1}, \ldots, \hat{e}_{\tau_p,p})^\top \in \mathbb{R}^p \,;\, (\tau_1, \ldots, \tau_p)^\top \in \mathbb{R}^p$$

$$\hat{e}_{\tau_j,j} = \arg\min_{\mu_j \in \mathbb{R}} \sum_{i=1}^{n} w_{ij}(\tau_j)(x_{ij} - \mu_j)^2$$

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

fixed weights, closed form solution

recalculate weights until convergence

LAWS estimation

- Given the weights w_{ij}, and

$$\mathcal{I}_\tau^+ = \{i \in \{1,\dots,n\} : w_{ij} = \tau_j\},$$

$$\mathcal{I}_\tau^- = \{i \in \{1,\dots,n\} : w_{ij} = 1 - \tau_j\}$$

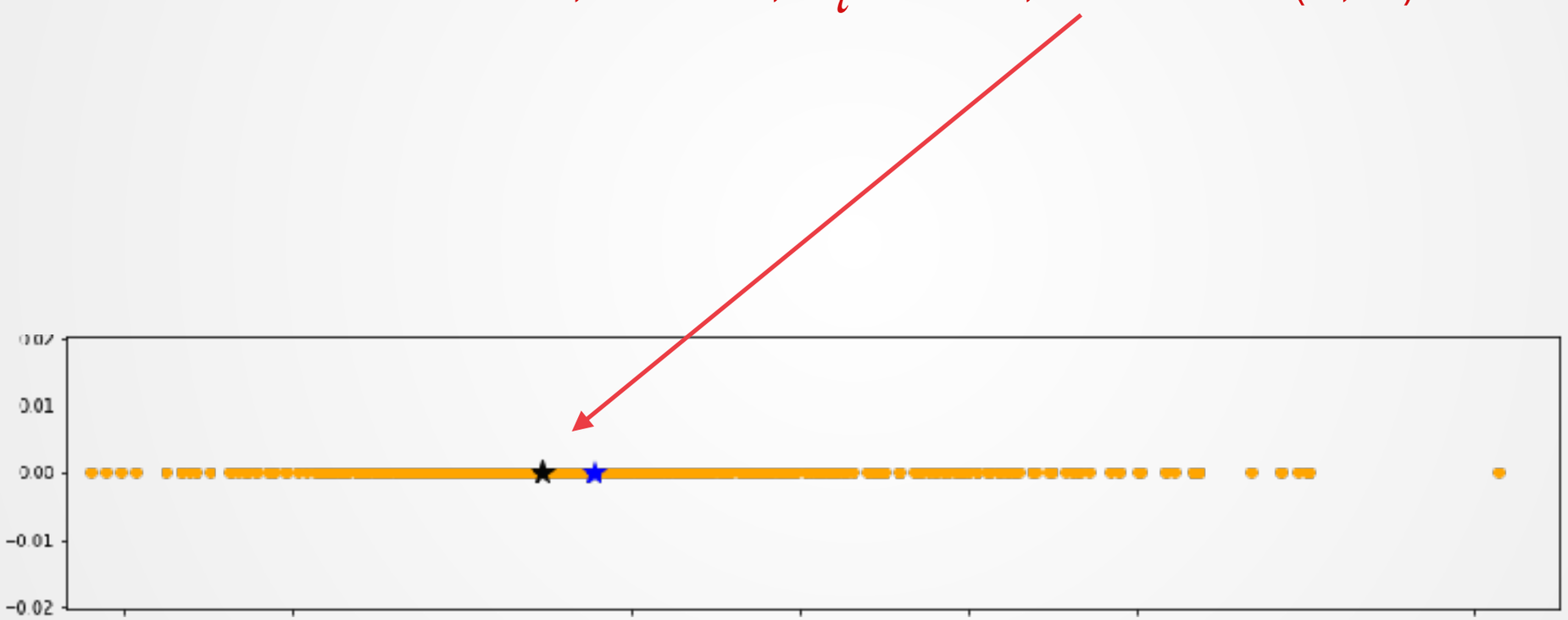
$$n^+ = |\mathcal{I}_\tau^+| \qquad n^- = |\mathcal{I}_\tau^-|$$

Estimators of τ-expectiles:

$$\hat{e}_{\tau_j,j} = \frac{\tau_j \sum_{i \in \mathcal{I}_{\tau_j}^+} x_{ij} + (1-\tau) \sum_{i \in \mathcal{I}_{\tau_j}^-} x_{ij}}{\tau_j n^+ + (1-\tau_j) n^-}$$

LAWS estimation

$n = 100$, $\tau = 0.3$, $e_\tau = 2.94$, X @ $AND(3, 4)$



3.557310 $\hat{e}_\tau$

τ- distance

Univariate τ- distance (Tran et al., 2019)

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

Coordinate-wise multivariate τ- distance where $\tau = (\tau_1, \ldots, \tau_p)^\top$

$$d(x_i, \tau_j, \theta) = \sum_{j=1}^{p} \left\{ \tau_j + (1 - 2\tau_j)\mathbf{I}_{\left\{x_{ij}<\theta_j\right\}} \right\} \|x_{ij} - \theta_j\|^2$$

which is accordance to the empirical version of expectile

Object: to minimise within-group τ- variance among K groups, where

$\tau = (\tau_1, \ldots, \tau_K)$

$$G^{K-expectiles}(\tau, \Theta, C, X) = \sum_{k=1}^{K} \sum_{j=1}^{p} d(x_{.j}, \tau_k, \theta_k)$$

Estimate τ

For fixed parameters (Θ, C), minimise objective function with respect to $\tilde{\tau}$

$$\hat{\tau}_k = \arg_{\tau \in (\mathbb{R}^p)^K} \min G(\tilde{\tau}, \Theta, C)$$

$$= \arg_{\tau} \min \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{j=1}^{p} \left\{ \tau_j + (1 - 2\tau_j) \mathbf{I}_{\left\{ x_{ij} < \theta_{C(i),j} \right\}} \right\} \| x_{ij} - \theta_{k,j} \|^2$$

$$\gamma_{k,j}$$

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

Unique Solution

$$\mathcal{I}_k^+ = \{ i \in \{1, \ldots, n\} : x_{ij} - \theta_{k,j} < 0,\ C(i) = k \}$$

$$\mathcal{I}_k^- = \{ i \in \{1, \ldots, n\} : x_{ij} - \theta_{k,j} \geq 0,\ C(i) = k \}$$

Where

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

$$n^- = |\mathcal{I}_k^-|$$

$$n^+ = |\mathcal{I}_k^+|$$

Cluster Centroids initialisation

Jain and Dubes, (1988): The K-means algorithm gave better results only when the initial partitions was close to the final solution

Cluster centroids Initialisation: K-means cluster centroids instead of randomly assigned initial points

Algorithm: Fixed τ clustering

Input Data, X; # of clusters, K; Vector parameter,τ;

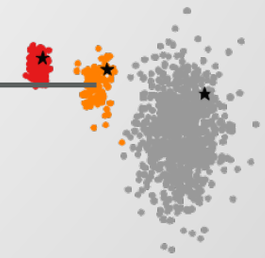
Output Cluster membership vector, C; Estimated cluster centroids, Θ

1. Initialize $\Theta_0 = \Theta_{\tau-expectile}$

Repeat

2. Assign points to the nearest cluster centre which minimises τ-distance, obtain C_0
3. Update Θ by minimise the Objective function $G(\tau, \Theta, C)$:
$$\Theta^{t+1} = \arg_{\Theta} \min G(\tau, \Theta, C^t)$$
4. Update membership vector C :
$$C^{t+1} = \arg_{C} \min G(\tau, \Theta^{t+1}, C)$$

until $\Theta^{t+1} - \Theta^t \leq \epsilon$



Algorithm: Adaptive τ clustering

Input Data, X; # of clusters, K

Output Cluster membership vector, C; Estimated cluster centroids, Θ

1. Initialize $\Theta_0 = \Theta_{k-means}$, $\tau_{0,kj} = 0.5$

Repeat

2. Assign points to the nearest cluster centre which minimises τ-distance, obtain C_0
3. Update Θ and τ by minimise the Objective function $G(\tau, \Theta, C)$:

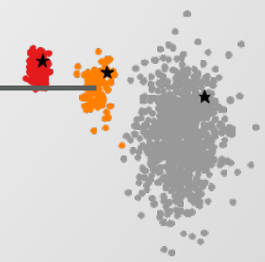
$$\hat{\tau}^{t+1} = \arg_{\tau} \min G(\tau, \Theta^t, C^t)$$

$$\Theta^{t+1} = \arg_{\Theta} \min G(\hat{\tau}^{t+1}, \Theta, C^t)$$

4. Update membership vector C :

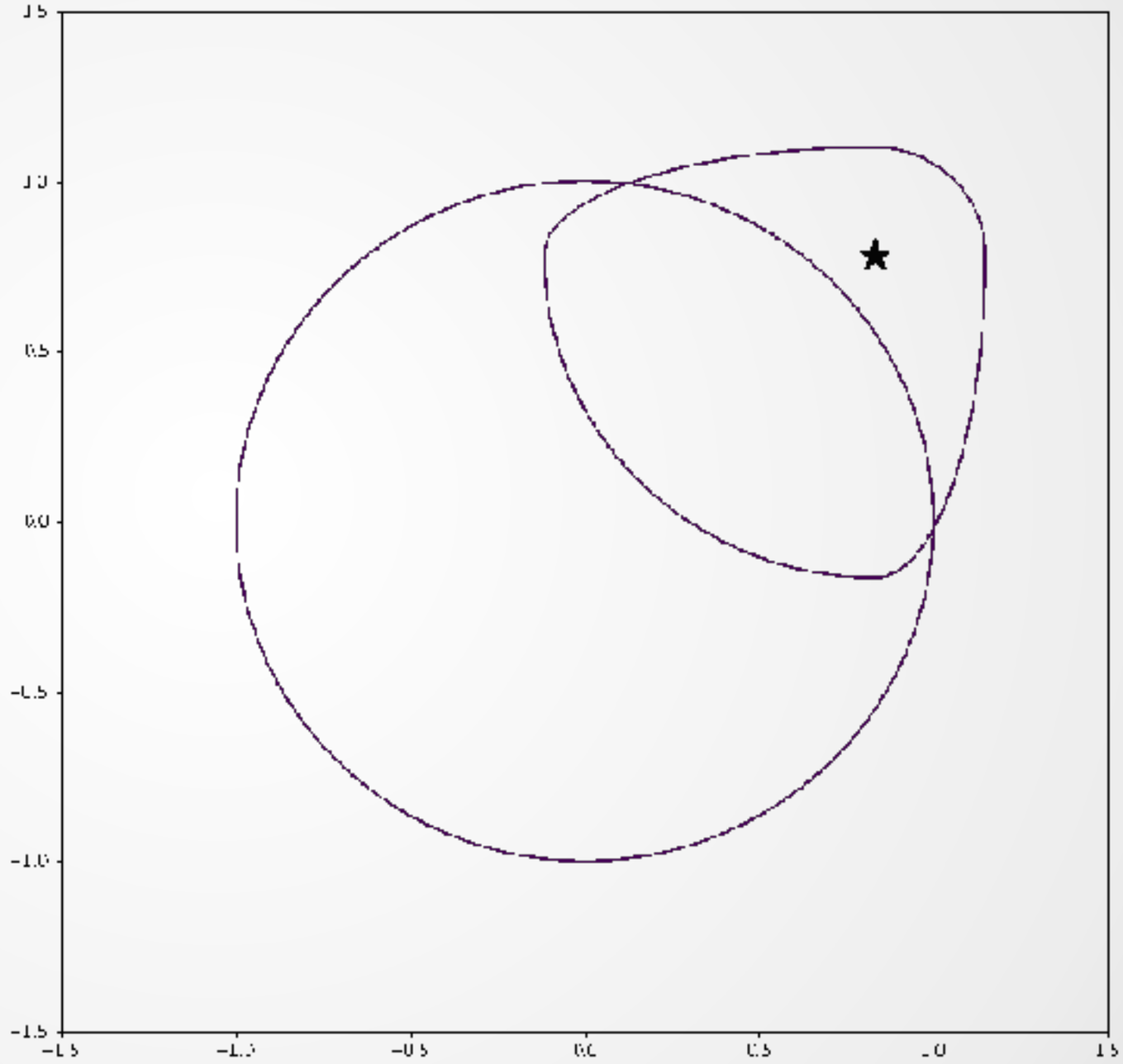
$$C^{t+1} = \arg_{C} \min G(\hat{\tau}^{t+1}, \Theta^{t+1}, C)$$

until $\Theta^{t+1} - \Theta^t \leq \epsilon$



Contour line of 2D expectiles

Sample: Normally-distributed
mean = [0,0]
Cov = [[1,0],[0,1]]
τ: from [0.9, 0.9] changed by [0.1, 0.1] to [0.1, 0.1]



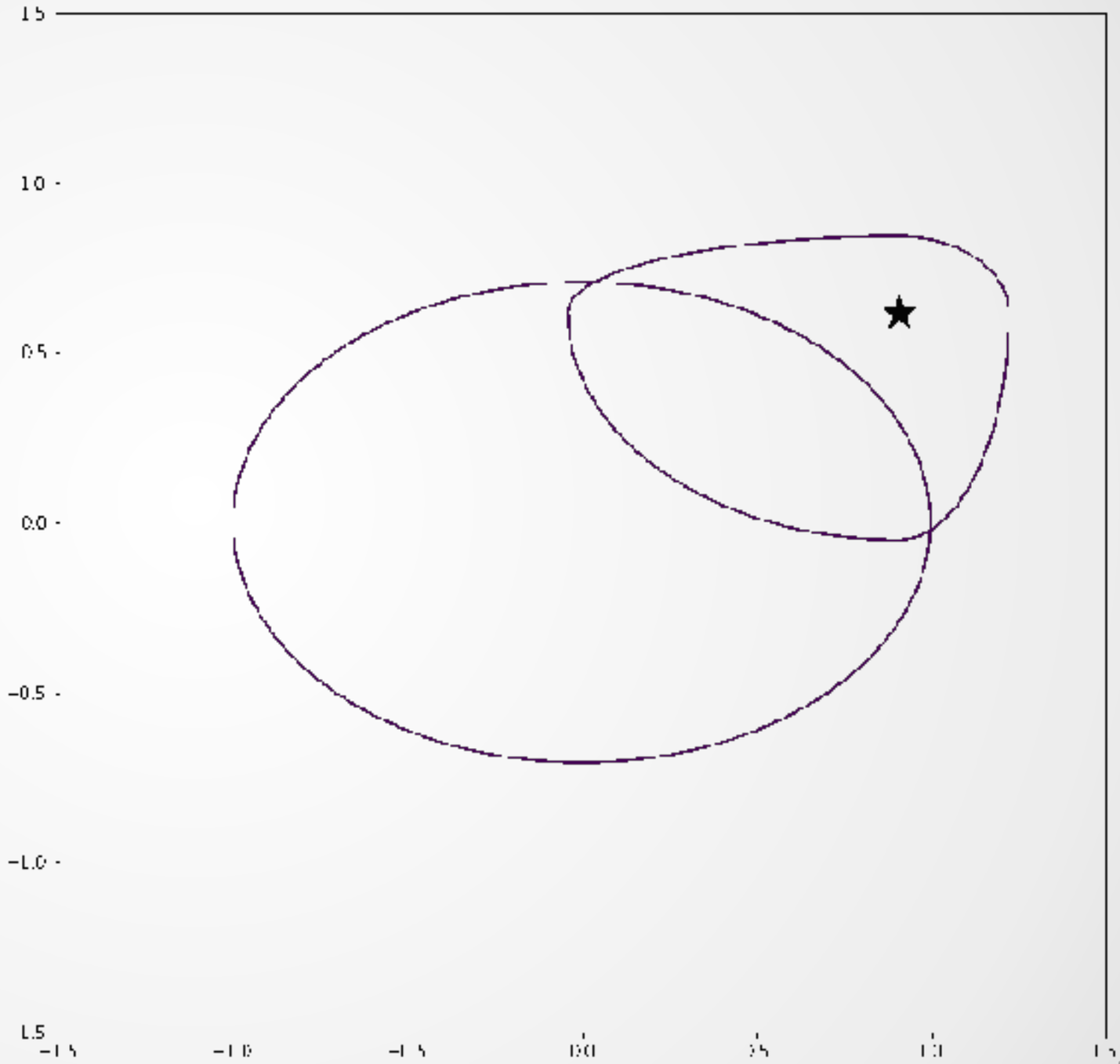
KEC_cluster shapes

Contour line of 2D expectiles

Sample: Normally-distributed
mean = [0,0]
Cov = [[1,0],[0,0.5]]

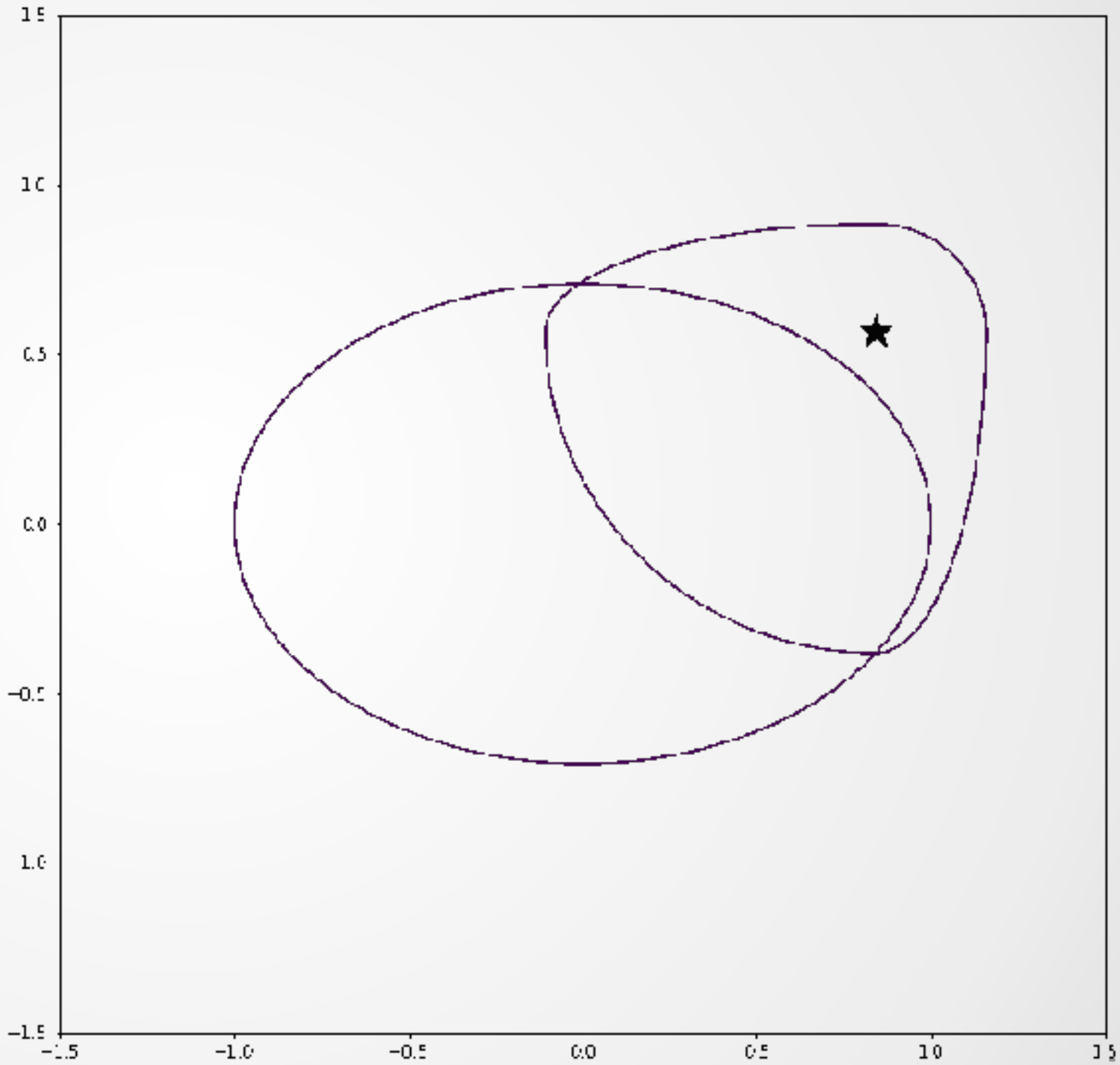
τ: from [0.9, 0.9] changed by
to [0.1, 0.1]

KEC_cluster shapes



Contour line of 2D expectiles

Sample: Normally-distributed
mean = [0,0]
Cov = [[1,0],[0,0.5]]
τ: from [0.9, 0.9] to [0.9,0.1]

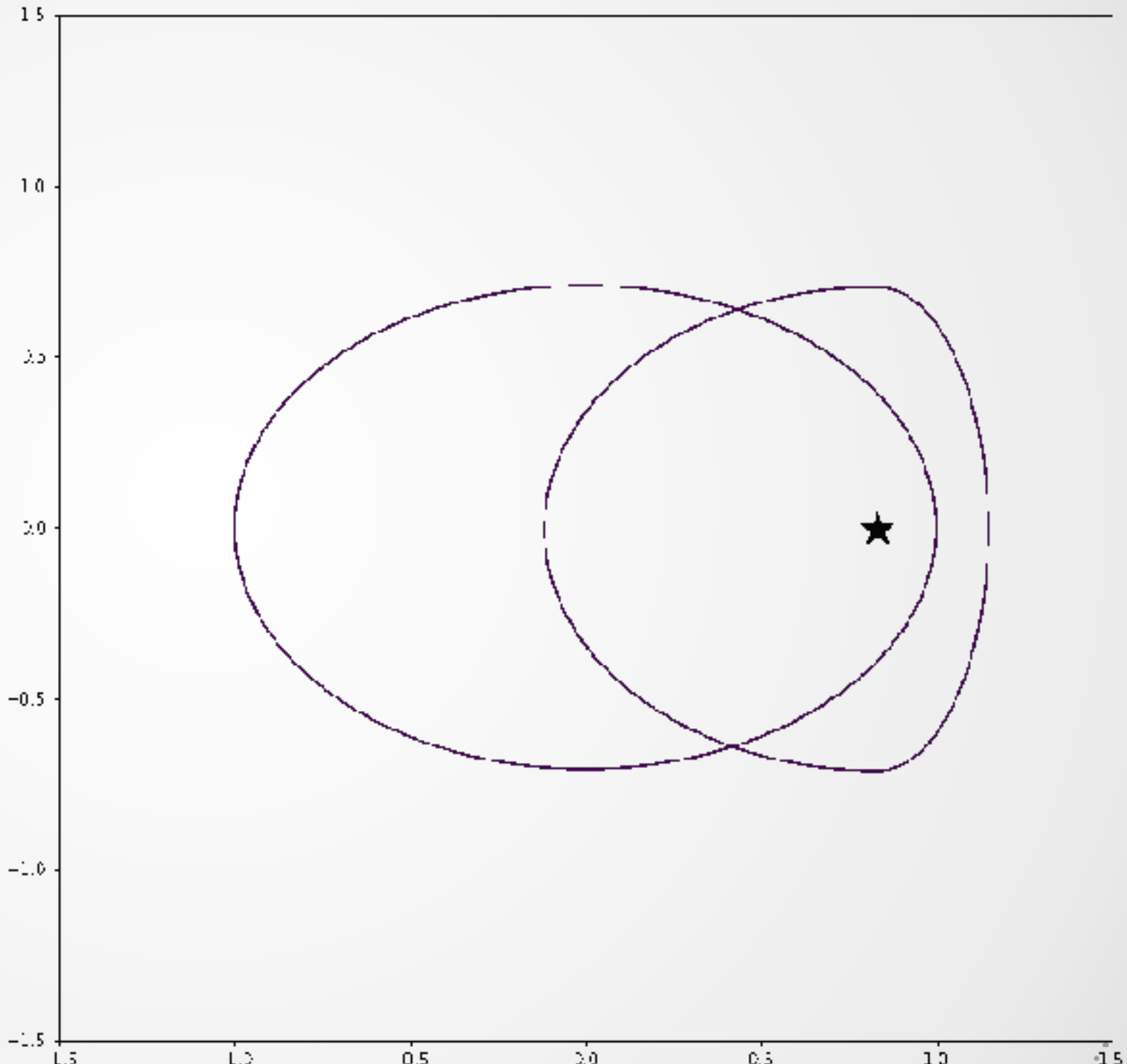


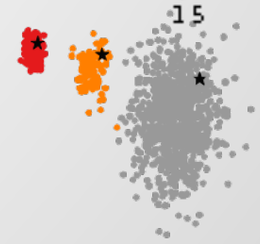
KEC_cluster shapes

Contour line of 2D expectiles

Sample: Normally-distributed
mean = [0,0]
Cov = [[1,0],[0,0.5]]
τ: from [0.9, 0.5]
to [0.1, 0.5]

KEC_cluster shapes

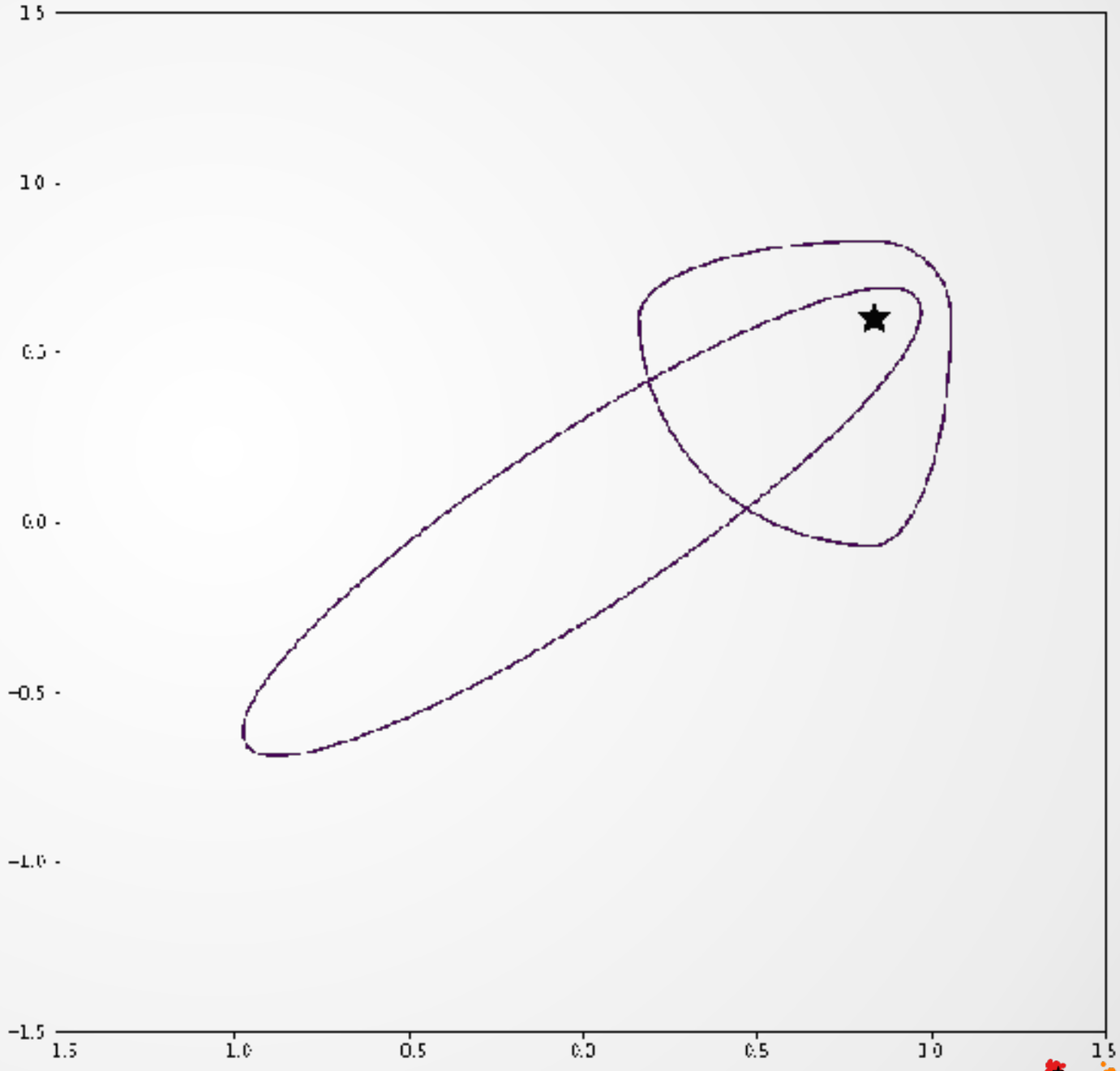


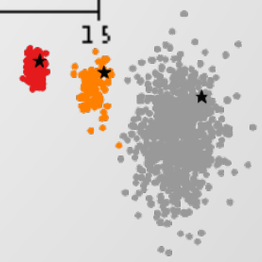


Contour line of 2D expertiles

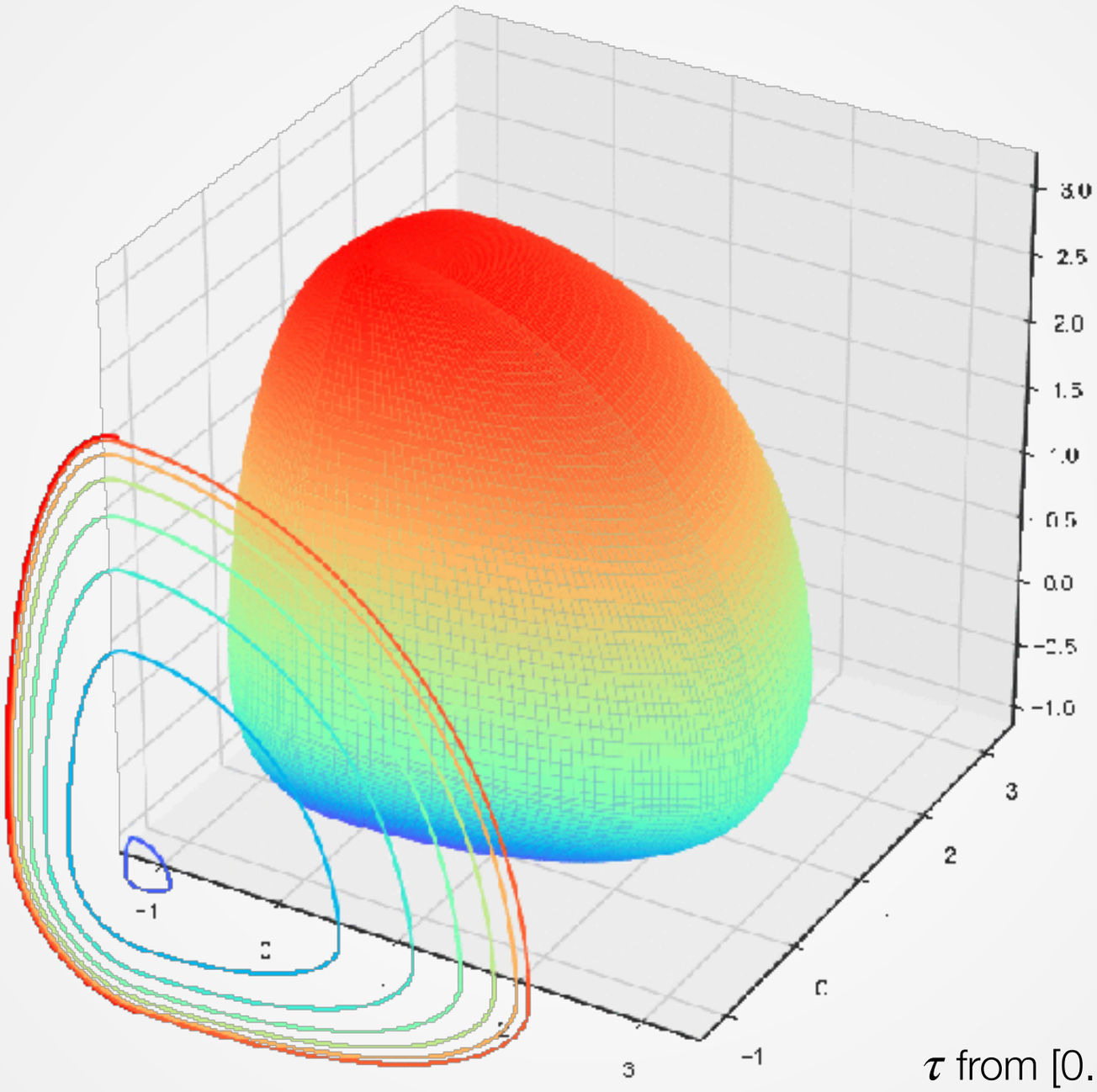
Sample: Normally-distributed
mean = [0,0]
Cov = [[1,1.27],[1.27,0.5]]
τ: from [0.9, 0.9]
to [0.1, 0.1]

KEC_cluster shapes





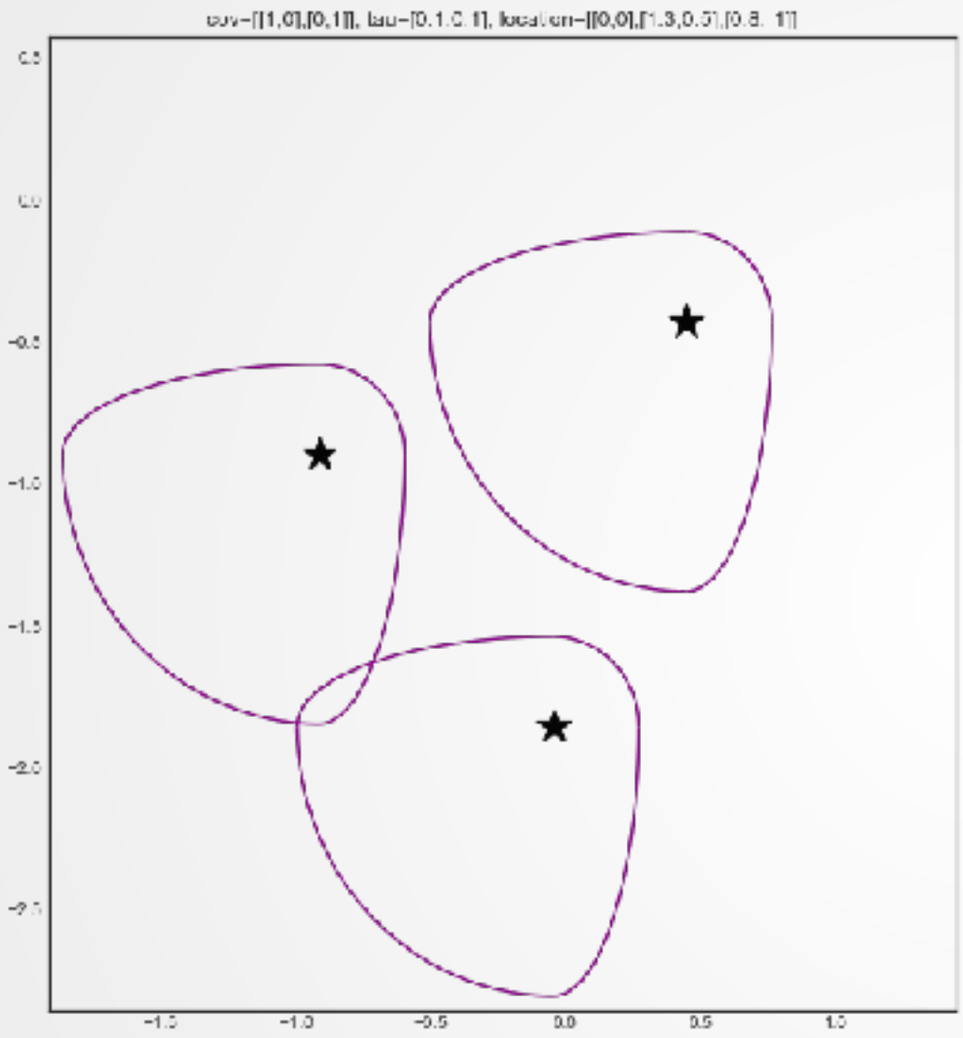
Cluster shape 3D



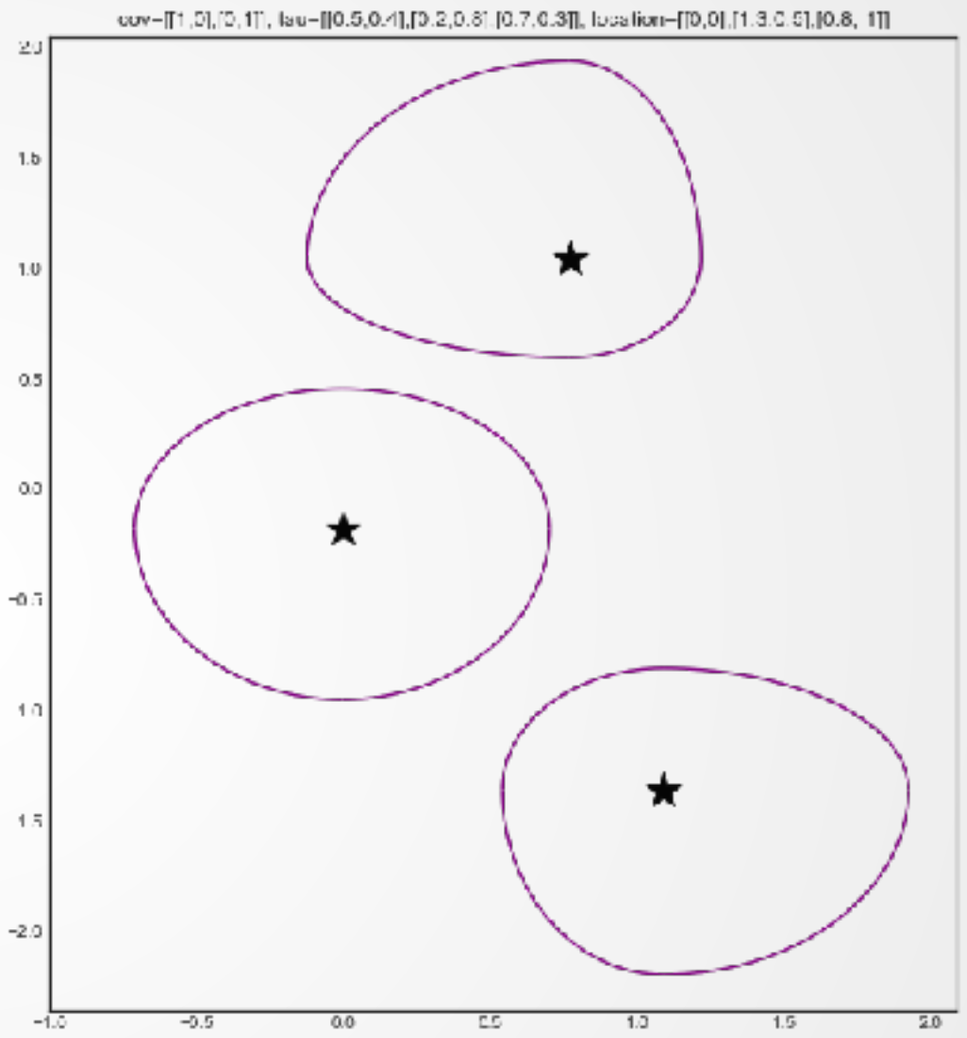
τ from [0.1, 0.1] to [0.9, 0.9]

KEC_cluster shapes

Why adaptive τ?



Univariate τ

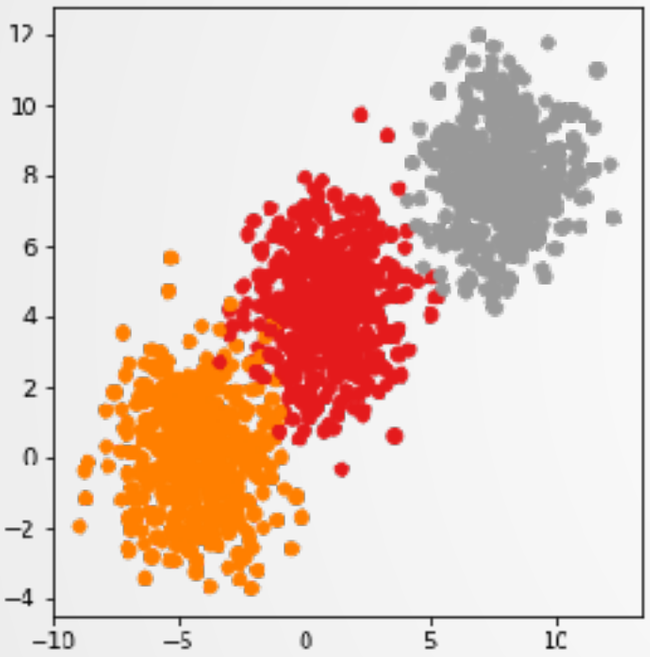


Adaptive τ

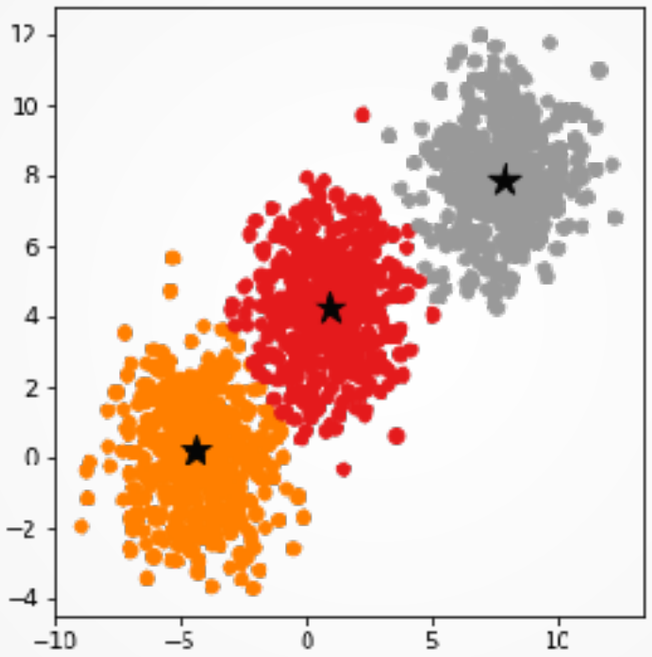
KEC_cluster shapes

Simulation results (normal clusters)

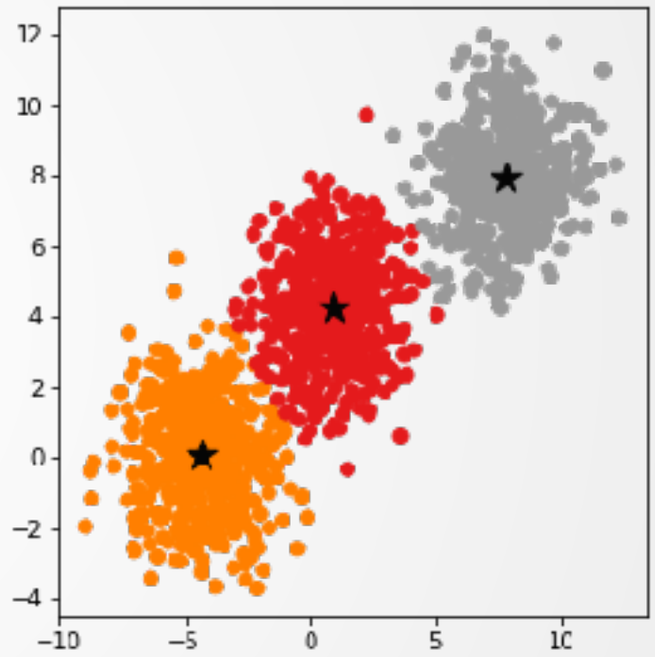
Cluster size: 500, 500, 500 std:1.5, 1.5, 1.5



True cluster



K-expectile $\hat{\tau}$=
[[0.47, 0.53]
[0.51, 0.48]
[0.51, 0.51]]
Accuracy =0.9893

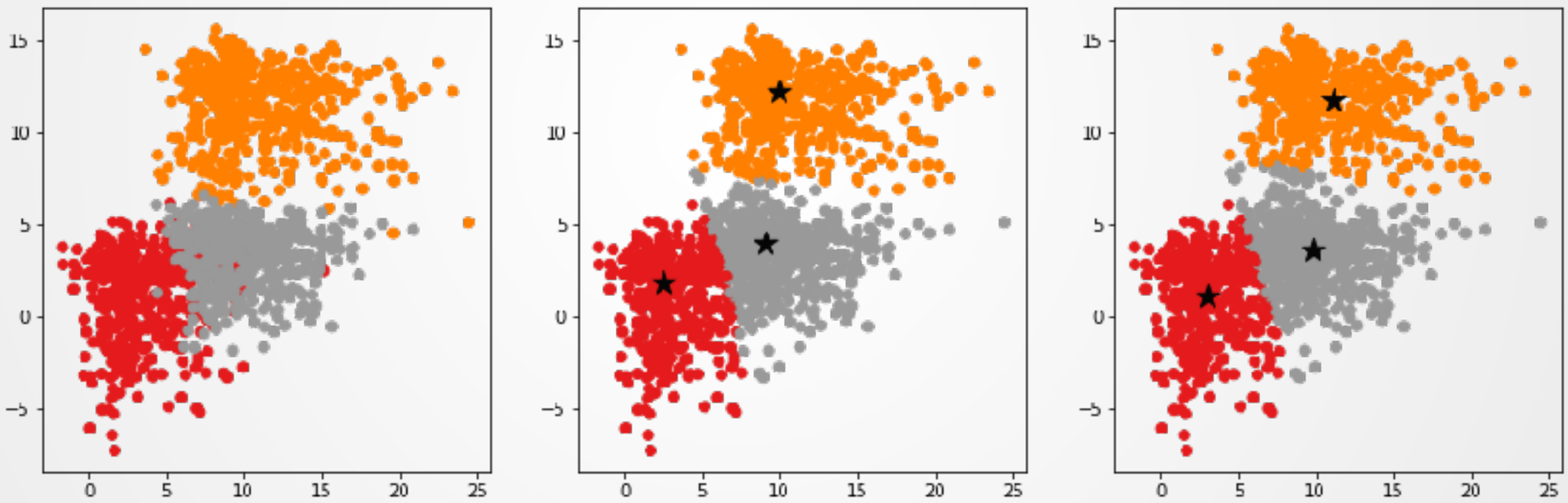


K-means
Accuracy =0.9893

KEC_simulations

Simulation results (AND clusters)

Cluster size: 500, 500, 500 location = [[3, 2], [10,12], [9,4]]
cov1 = [[2, 0], [0, 2]], cov2 = [[5, 0], [0, 2]], cov3 = [[3, 0], [0, 1]]
τ = [0.3, 0.7]



True cluster

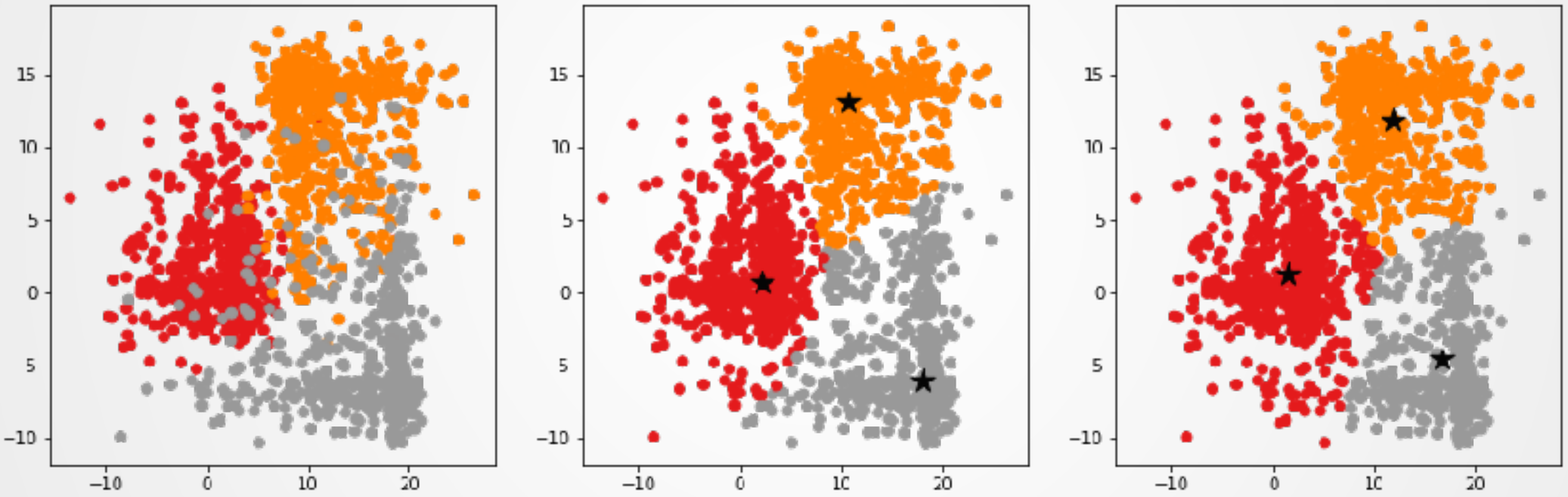
K-expectile $\hat{\tau}$=
[[0.330, 0.673],
 [0.300, 0.626],
 [0.353, 0.681]]
Accuracy =0.922

K-means
Accuracy =0.91

KEC_simulations

Simulation results (AND clusters)

Cluster size: 500, 500, 500 location = [[2,1], [10,13], [18,-6]]
cov1 = [[5, 0], [0, 5]], cov2 = [[5, 0], [0, 5]], cov3 = [[5, 0], [0, 5]]
τ = [[0.7, 0.3],[0.23, 0.78],[0.88, 0.2]]



True cluster

K-expectile $\hat{\tau}$=
[[0.70, 0.36],
 [0.34, 0.73],
 [0.84, 0.20]]
Accuracy =0.923

K-means
Accuracy =0.899

KEC_simulations

Asymmetric normal distribution (AND)

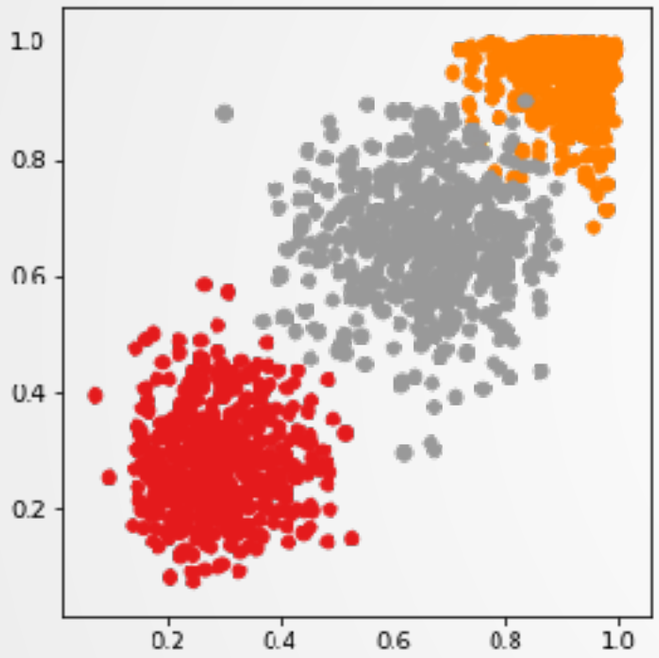
$X = (X_1, \ldots, X_p) \in (\mathbb{R}^n)^p$, θ is the set of parameters

Density function of AGD, where $\mu = (\mu_1, \ldots, \mu_j)^\top$ is parameters

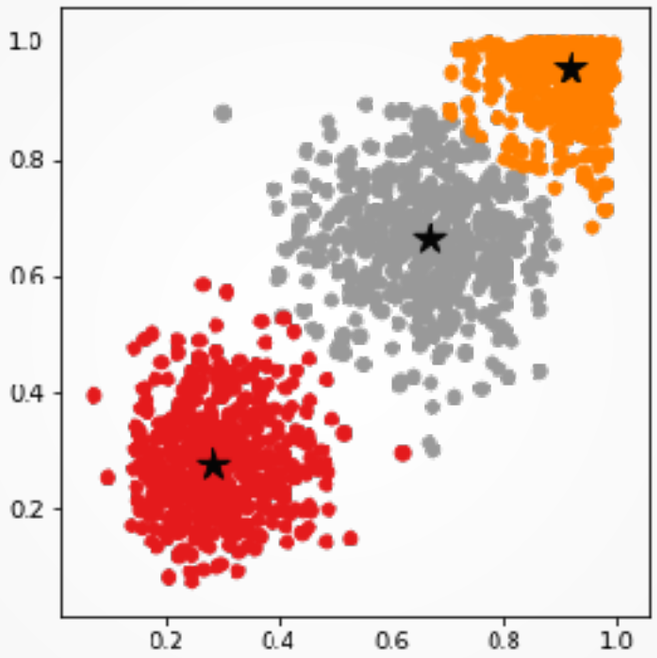
$$p(X_j \mid \theta) = \prod_{j=1}^{p} \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_l + \sigma_r} \begin{cases} \exp\left(-\dfrac{(X_j - \mu_j)^2}{2\sigma_{l_j}^2} \right) & X_j < \mu_j \\ \exp\left(-\dfrac{(X_j - \mu_j)^2}{2\sigma_{l_j}^2} \right) & X_j \geq \mu_j \end{cases}$$

Simulation results (skewed-distributed non-leptokurtic clusters)

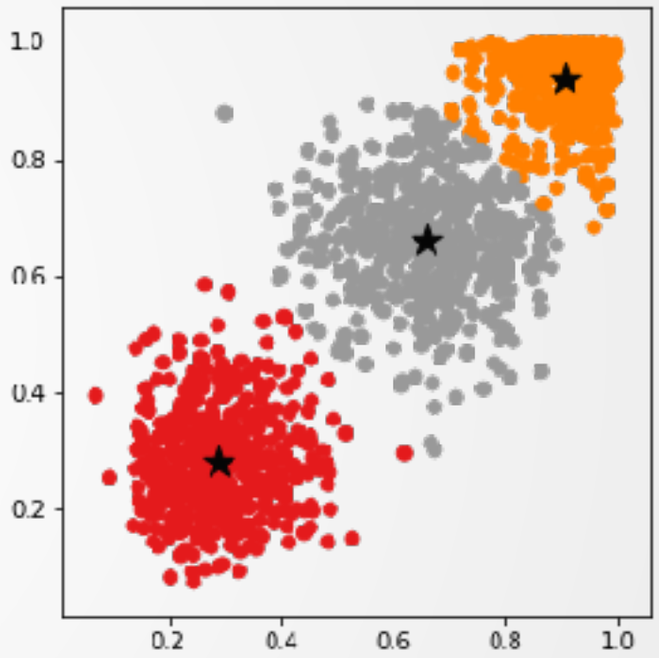
Cluster size: 500, 500, 500 Beta distribution parameter: α_1, β_1= [8, 20], [20,2], [12,6] α_2, β_2= [8, 20], [15,1], [12,6]



True cluster



K-expectile τ=
[[0.536, 0.516],
[0.591, 0.706],
[0.443, 0.480]]
Accuracy = 0.9833



K-means
Accuracy = 0.9826

KEC_simulations

Simulation results (50 rounds)

K multivariate Gaussian samples with unit variance and different location parameters. $G_k \sim N(\mu_k, \mathcal{I}_p)$, where μ_1 is a p dimensional vector randomly generated in the interval (1,10), and then shift the location of other clusters by $\mu_k = \mu_1 + 2k$. Clusters are in the same size.

	n =1500			n = 300		
	p = 10	p = 50 (c = 5)	p = 100(c = 20)	p = 10	p = 50 (c = 5)	p = 100(c = 20)
	ARI	ARI	ARI	ARI	ARI	ARI
K-expectiles_vtau	99,36	99,60	99,87	97,00	97,99	97,99
K-means	99,36	99,60	99,87	97,00	97,99	97,99
Spectral	No convergence	31,22	86,74	No convergence	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

Best

OK

Not so OK

Simulation results (50 rounds)

A mixture of asymmetric normal distributions. For each cluster $G_k = (W_1, W_2, \ldots, W_p)^\top$, we first generate p standard normally distributed samples, each $Z_j^k \sim N(0,25)$. Given parameter τ_j^k , which is randomly chosen in $[0.1,0.9]$, and the j-th element of location parameter $e_{\tau_j^1}$ is generated in $(0,10)$, for k-th cluster, we shift the location by $e_{\tau_j^k} = e_{\tau_j^1} + 7(-1)^j(j-1)$, W_j^k can be converted as following:

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

	n = 1500			n = 300		
	p = 10	p = 50 (c = 5)	p = 100 (c = 20)	p = 10	p = 50 (c = 5)	p = 100 (c = 20)
	ARI	ARI	ARI	ARI	ARI	ARI
K-expectiles_vtau	93,22	99,60	99,60	92,20	97,99	97,99
K-means	91,19	99,60	99,60	81,70	97,99	97,99
Spectral	No convergence	No convergence	No convergence	No convergence	No convergence	No convergence
h-ward	77,19	99,60	99,60	76,98	97,01	97,99
CS	86,61	99,60	99,60	88,98	97,99	71,74
CU	80,73	99,60	99,60	72,76	97,50	76,27
VS	88,86	99,60	99,60	93,16	97,99	97,99
VU	85,74	99,60	99,60	80,41	97,99	97,99

Simulation results (50 rounds)

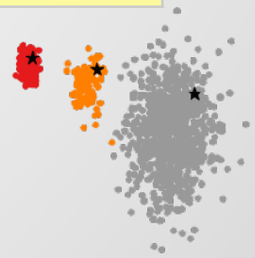
K multivariate skewed generalized t-distribution samples. Dimension $p = 2$, parameters $df = [10,10,10], nc = [3, -1.5,2.5], loc = [[0,2], [1,0], [0.5,1]], scale = 0.5$.

	n = 1500	n = 300
	p = 2	p = 2
	ARI	ARI
K-expectiles_vtau	93,78	94,09
K-means	93,49	93,15
Spectral	93,03	92,24
h-ward	94,35	91,79
CS	93,78	93,62
CU	93,21	92,67
VS	93,78	93,13
VU	93,21	92,67

Simulation results (50 rounds)

K multivariate Beta-distributed samples. 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 choose a_j randomly from interval $(1,10)$ and b_j randomly from interval $(10,20)$.

	n = 1500			n = 300		
	p = 10	p = 50 (c = 5)	p = 100 (c = 20)	p = 10	p = 50 (c = 5)	p = 100 (c = 20)
	ARI	ARI	ARI	ARI	ARI	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	No convergence	99,60	93,17	No convergence	97,99
h-ward	94,80	99,60	99,60	88,88	97,99	97,99
CS	68,92	99,60	78,27	92,20	97,99	55,70
CU	68,14	99,60	78,22	92,17	77,06	56,12
VS	93,28	94,30	82,12	91,98	97,99	23,81
VU	94,03	99,60	69,70	92,56	97,99	10,23



Simulation results (skewed-distributed leptokurtic clusters)(50 rounds)

K multivariate F-distributed samples. For variables in the first cluster, $W_j^1 \sim F(a_j, a_j) + 1$,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 randomly chosen within 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 generated from $(15,25)$ and $(60,70)$.

	n = 1500		n = 300	
	p = 2	p = 400 (c=10)	p = 2	p = 400 (c=10)
	ARI	ARI	ARI	ARI
K-expectiles_vtau	95,8	97,99	94,58	97,99
K-means	95,19	97,99	94,01	97,99
Spectral	94,89	No convergence	93,82	No convergence
h-ward	96,82	97,99	95,25	97,99
CS	97,96	97,99	96,03	97,99
CU	95,42	97,99	94,19	97,99
VS	97,72	97,99	95,64	97,99
VU	95,44	97,99	94,19	97,99

Simulation results (skewed-distributed leptokurtic clusters)(50 rounds)

K multivariate skewed generalized t-distribution samples. Dimension $p = 2$, parameters $df = [4,5,6], nc = [3, -1.5, -2.5], loc = [[0,7], [6,0], [6,4]], scale = 0.5$.

	n = 1500	n = 300
	p = 2	p = 2
	ARI	ARI
K-expectiles_vtau	91,08	87,75
K-means	90,78	87,37
Spectral	83,55	78,55
h-ward	89,6	84,78
CS	91,49	88,31
CU	91,59	88,7
VS	91,37	88,03
VU	91,55	88,51

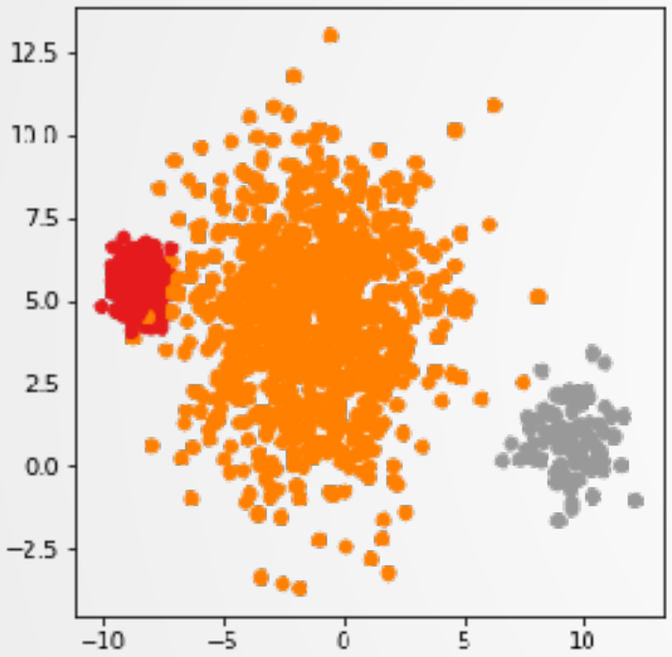
Simulation results (skewed-distributed leptokurtic clusters)(50 rounds)

K multivariate χ^2 distributed samples. Dimension $p = 2$, parameters $df = [4,5,6], loc = [0,4,3, -1.5], scale = 0.5$.

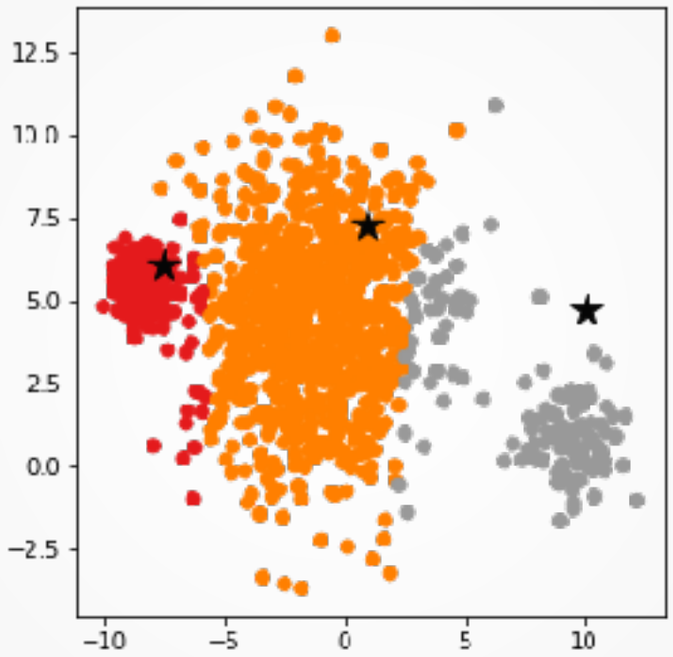
	n = 1500	n = 300
	p = 2	p = 2
	ARI	ARI
K-expectiles_vtau	76,49	75,70
K-means	75,15	73,91
Spectral	62,49	64,40
h-ward	71,04	67,12
CS	84,26	76,90
CU	77,38	66,84
VS	84,03	81,27
VU	73,17	63,54

Simulation results (unevenly-sized Gaussian clusters)

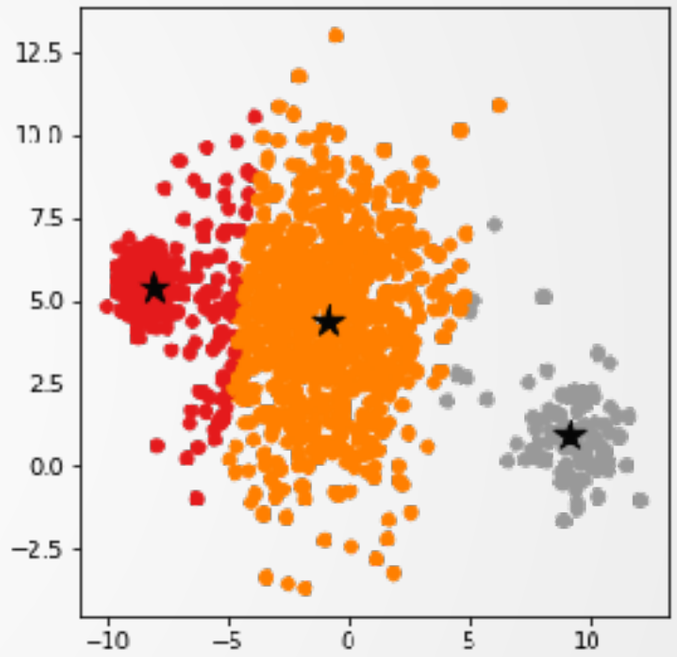
Cluster size: 100, 900, 500 std:1, 2.5, 0.5



True cluster



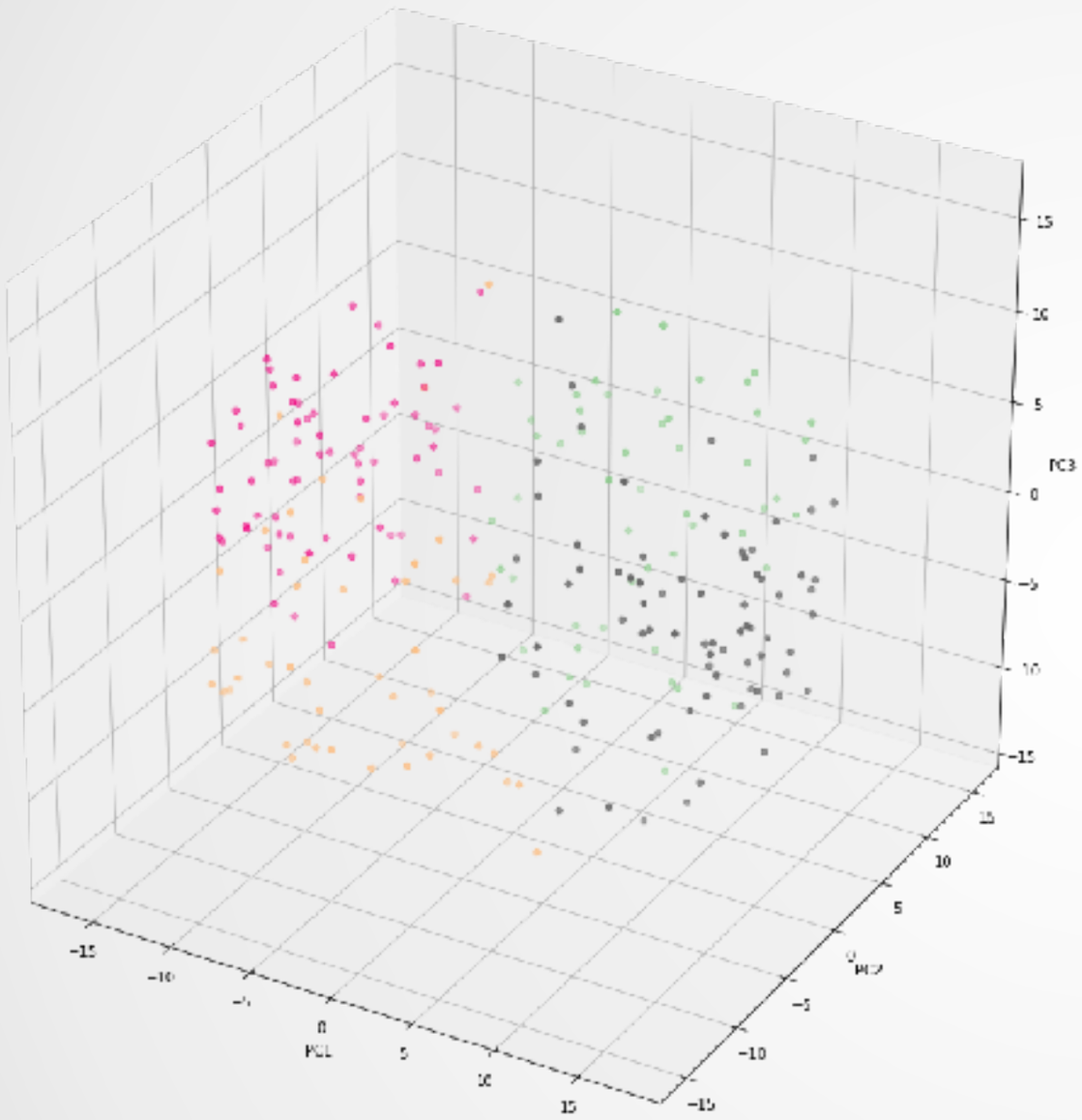
K-expectile τ= [0.05, 0.05]
Accuracy = 0.9506



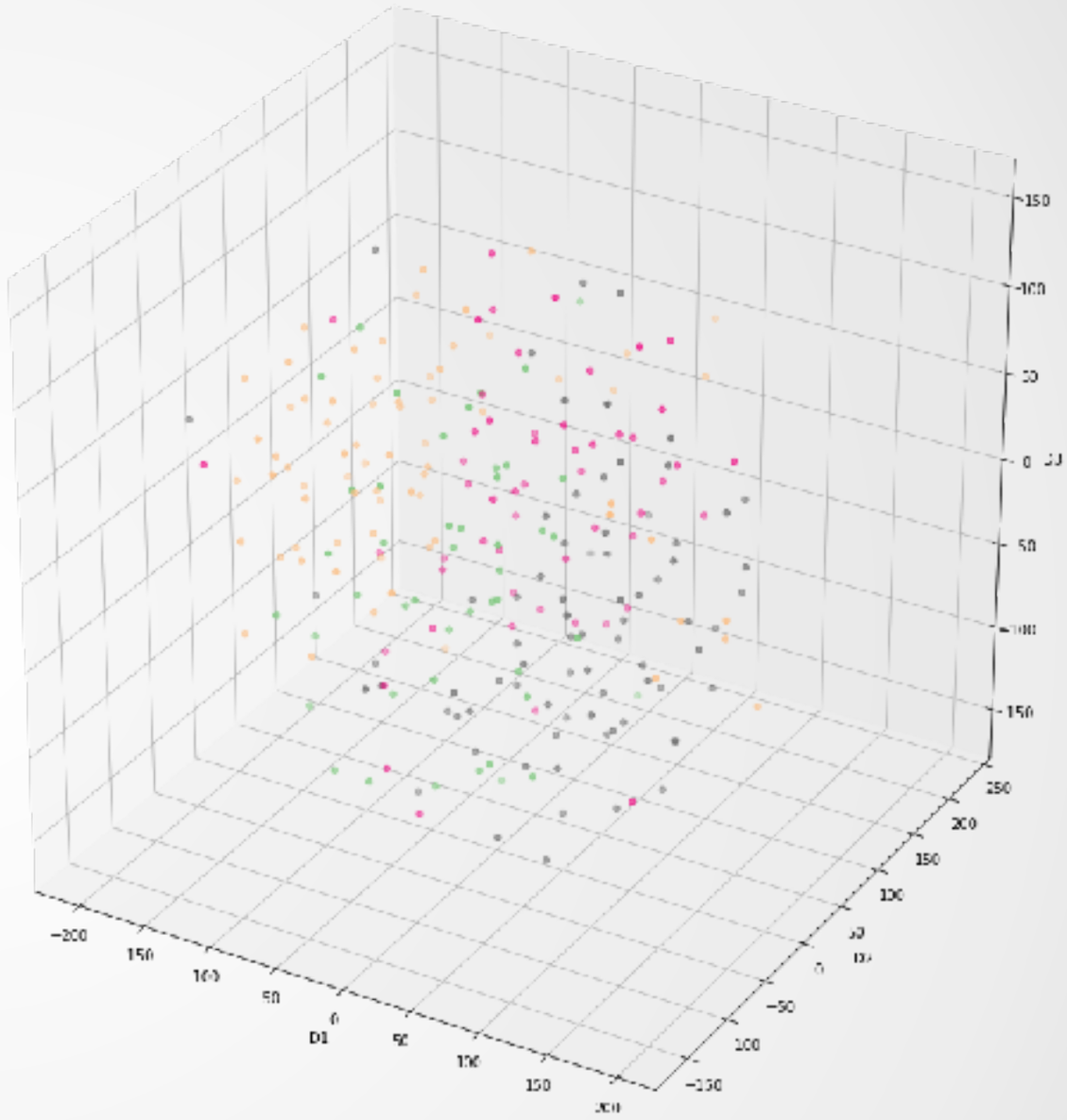
K-means
Accuracy = 0.9373

KEC_simulations

4 clusters of 248 Cryptos: log returns 20170730-20200425



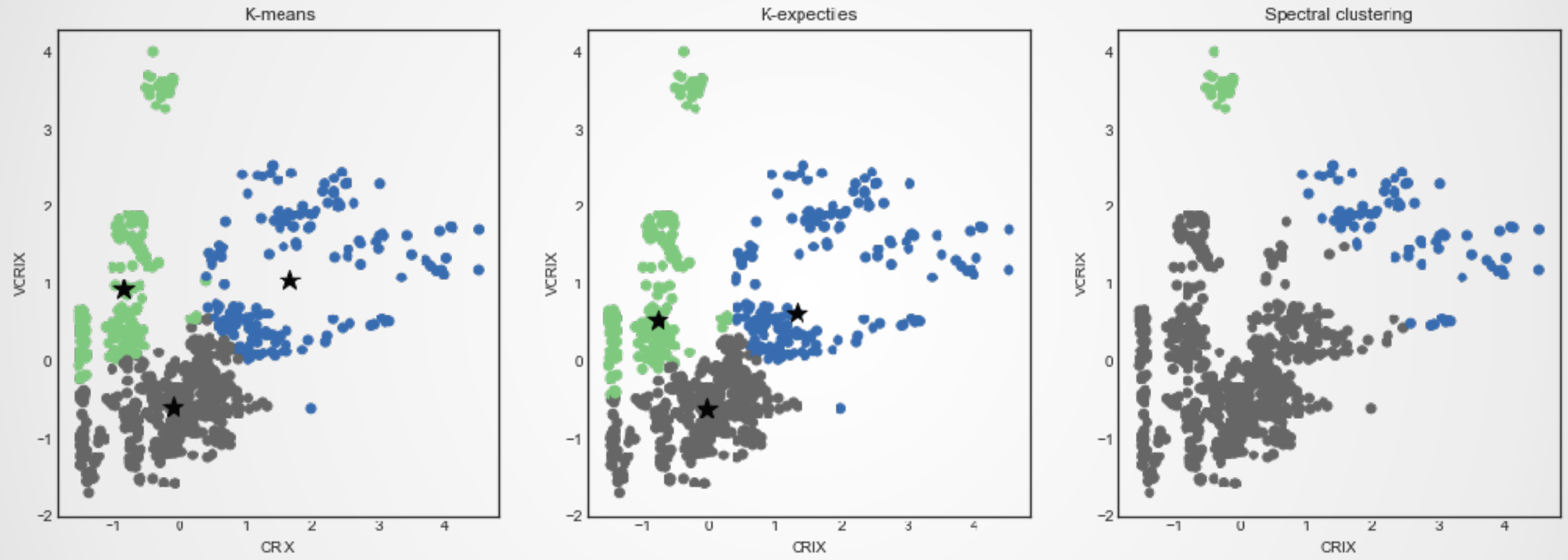
PCA



TSNE

KEC_applications

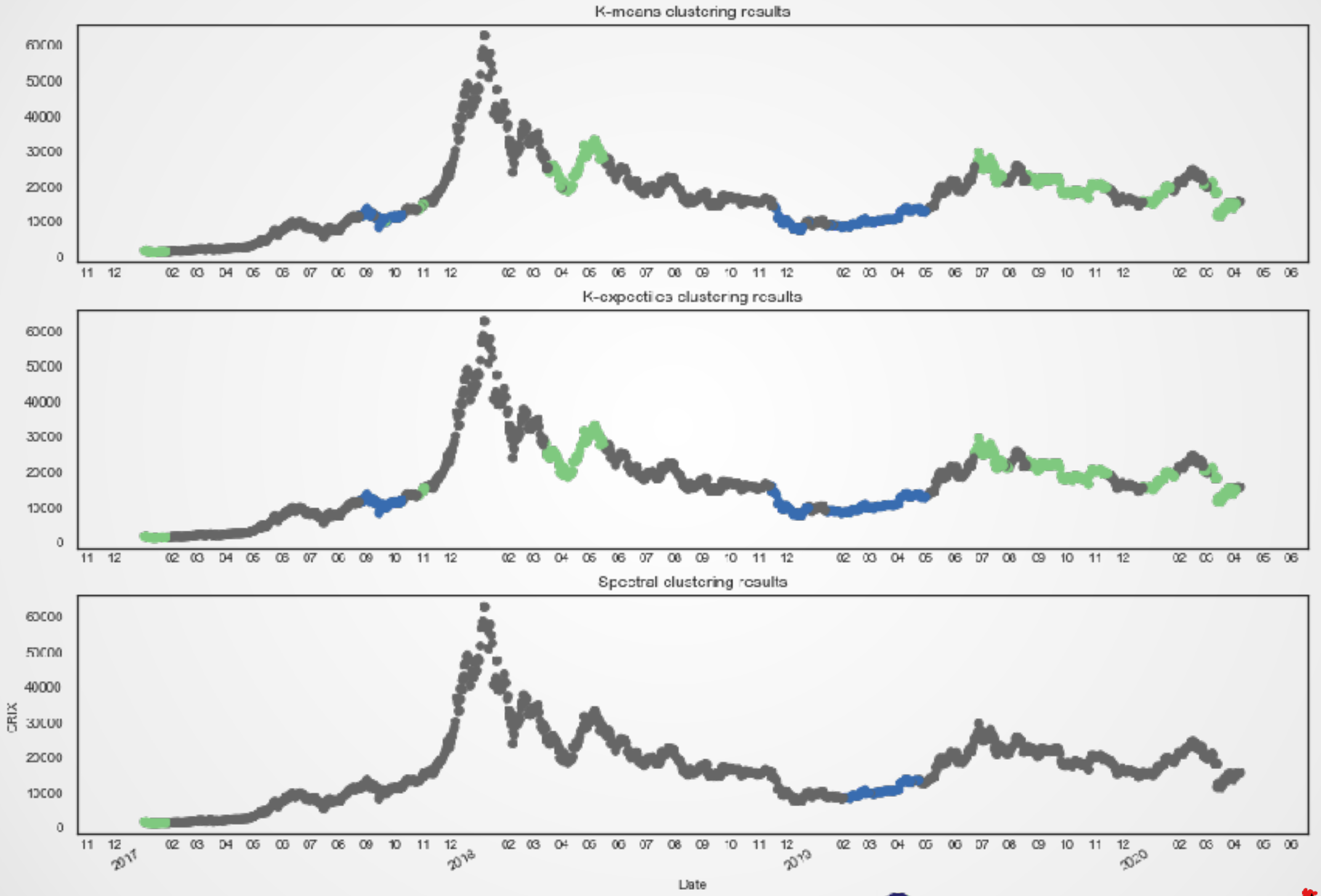
Data view (CRIX-VCRIX 20170102-20200407)



3 clusters

KEC_applications

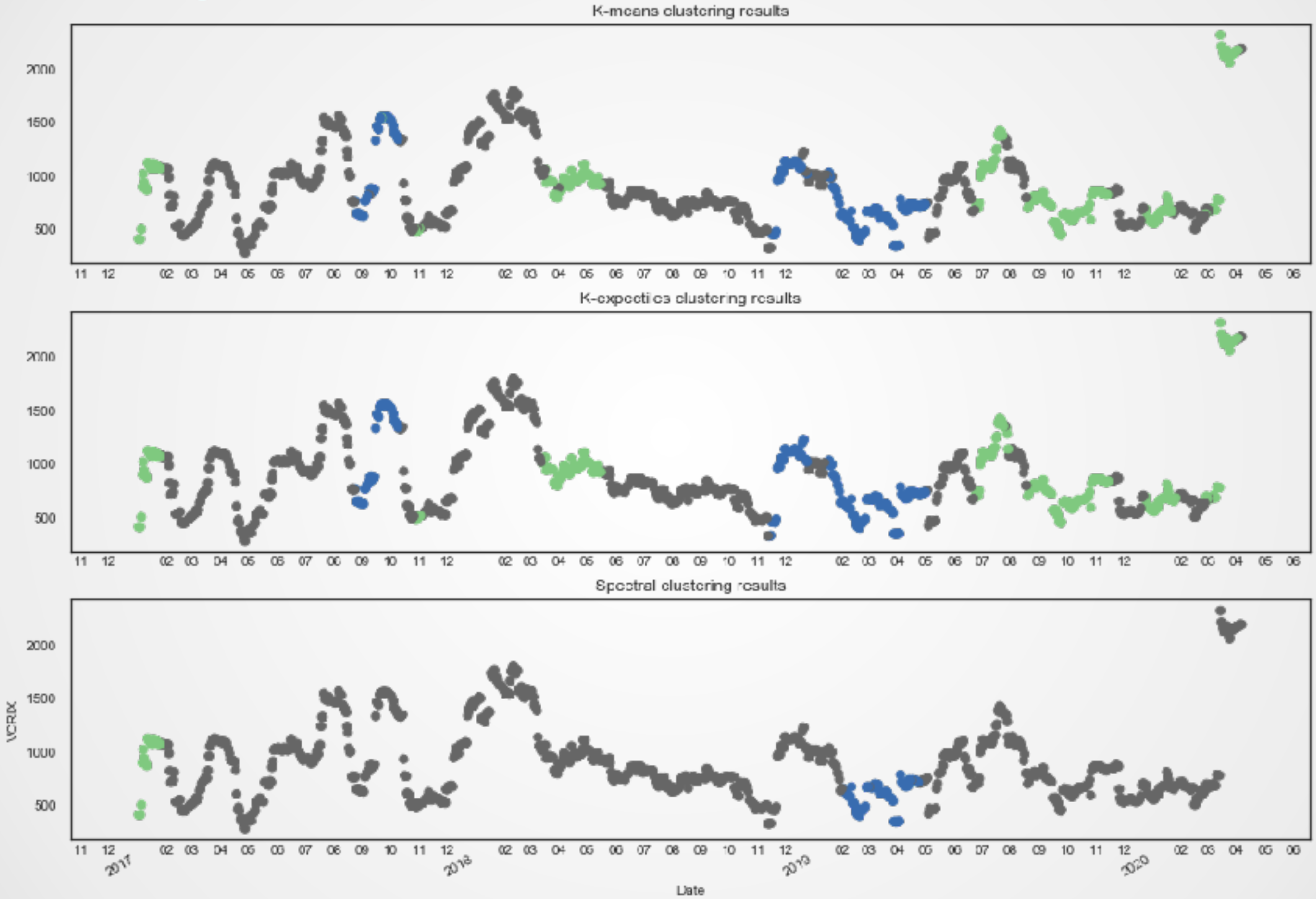
Clustering results (CRIX 20170102-20200407)



3 clusters

KEC_applications

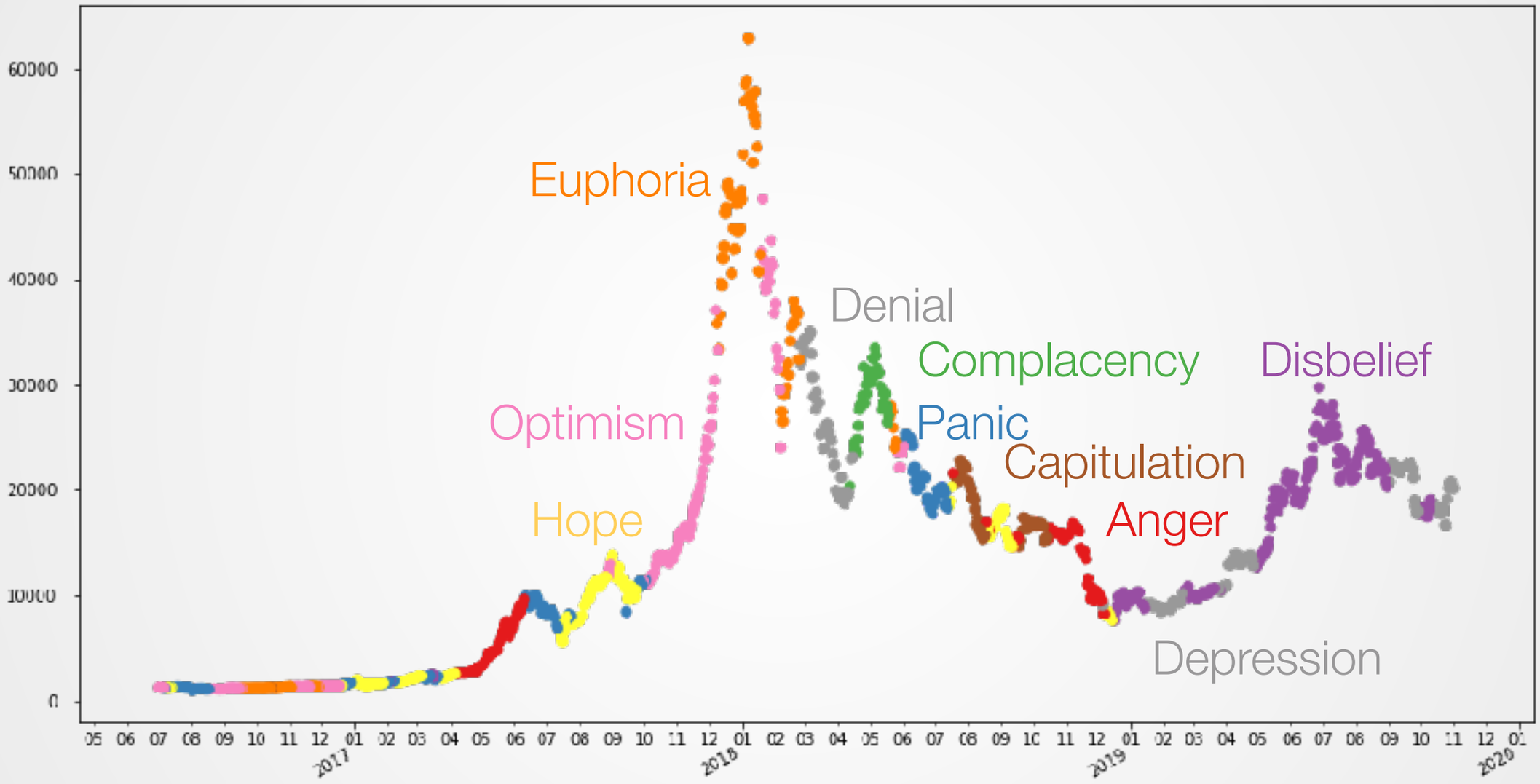
Clustering results (VCRIX 20170102-20200407)



Clustering on VCRIX 20160701-20200407
3 clusters

KEC_applications

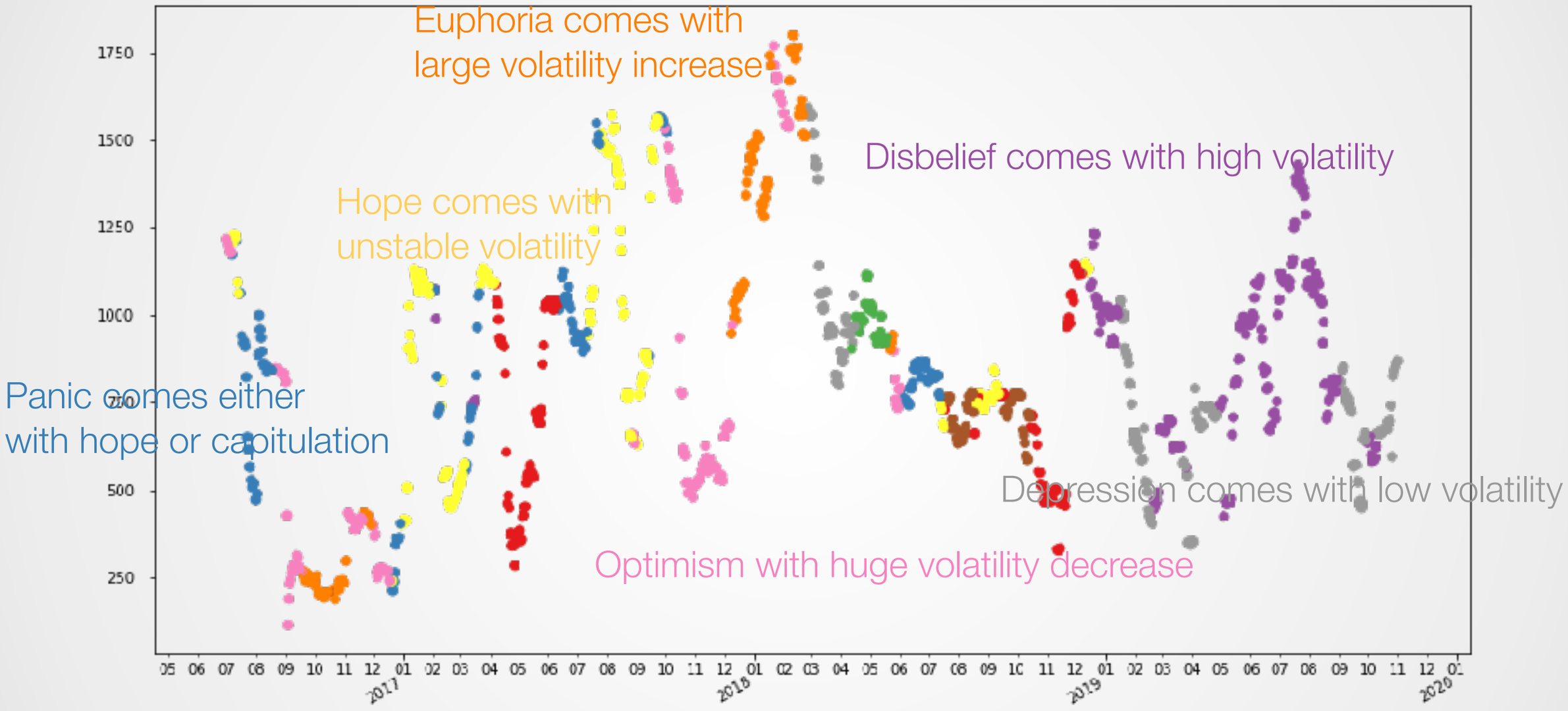
Regime (CRIX 20160701-20191101)



Clustering on CRIX-VCRIX 20160701-20191101
10 clusters

KEC_applications

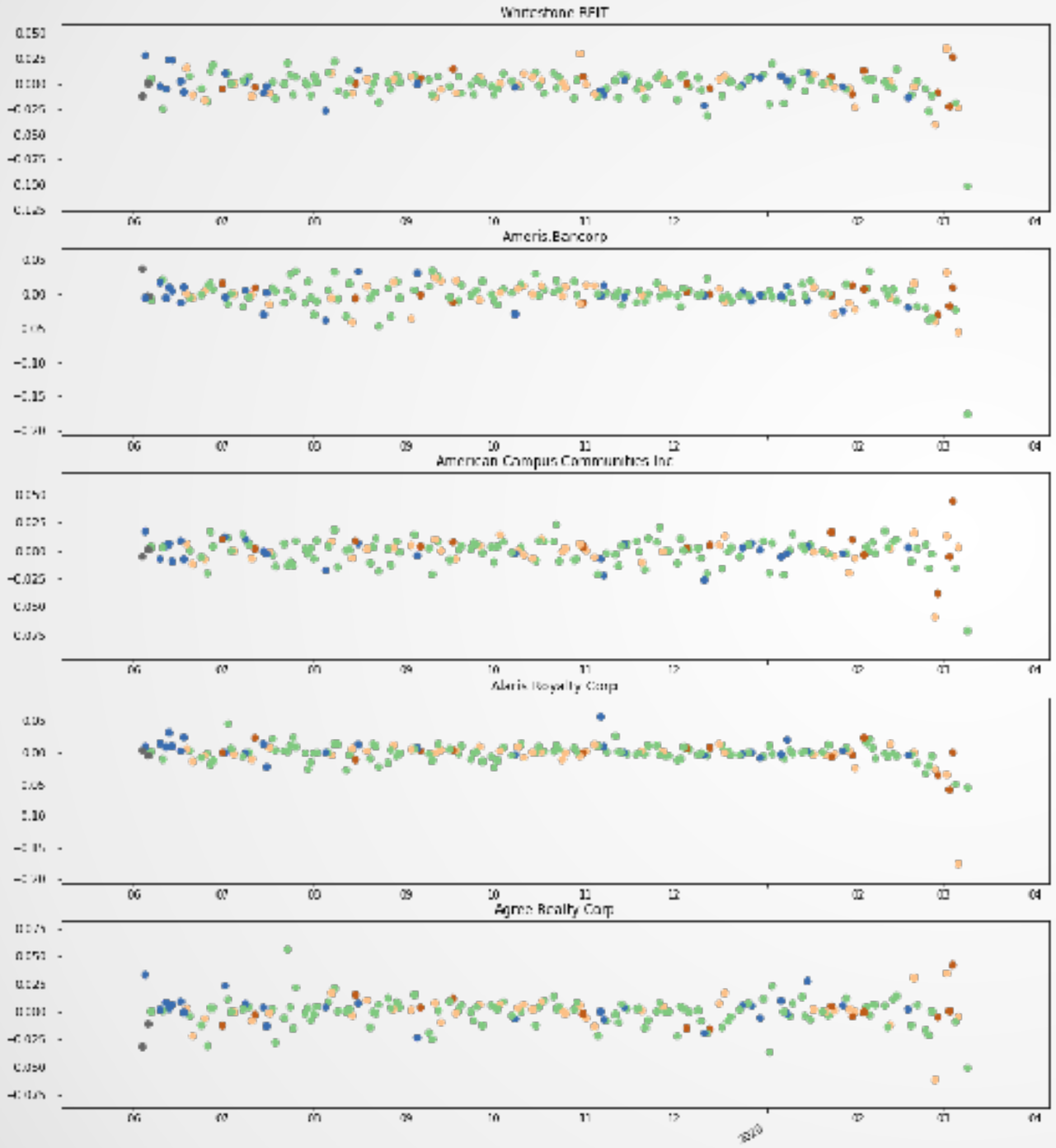
Regime (VCRIX 20160701-20191101)

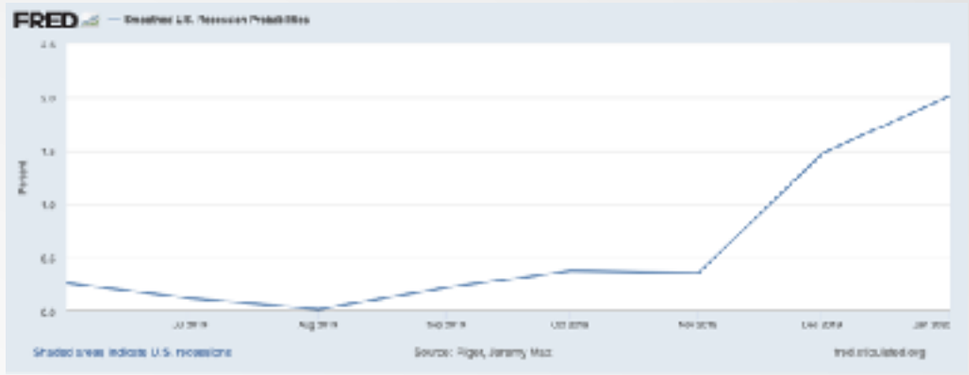


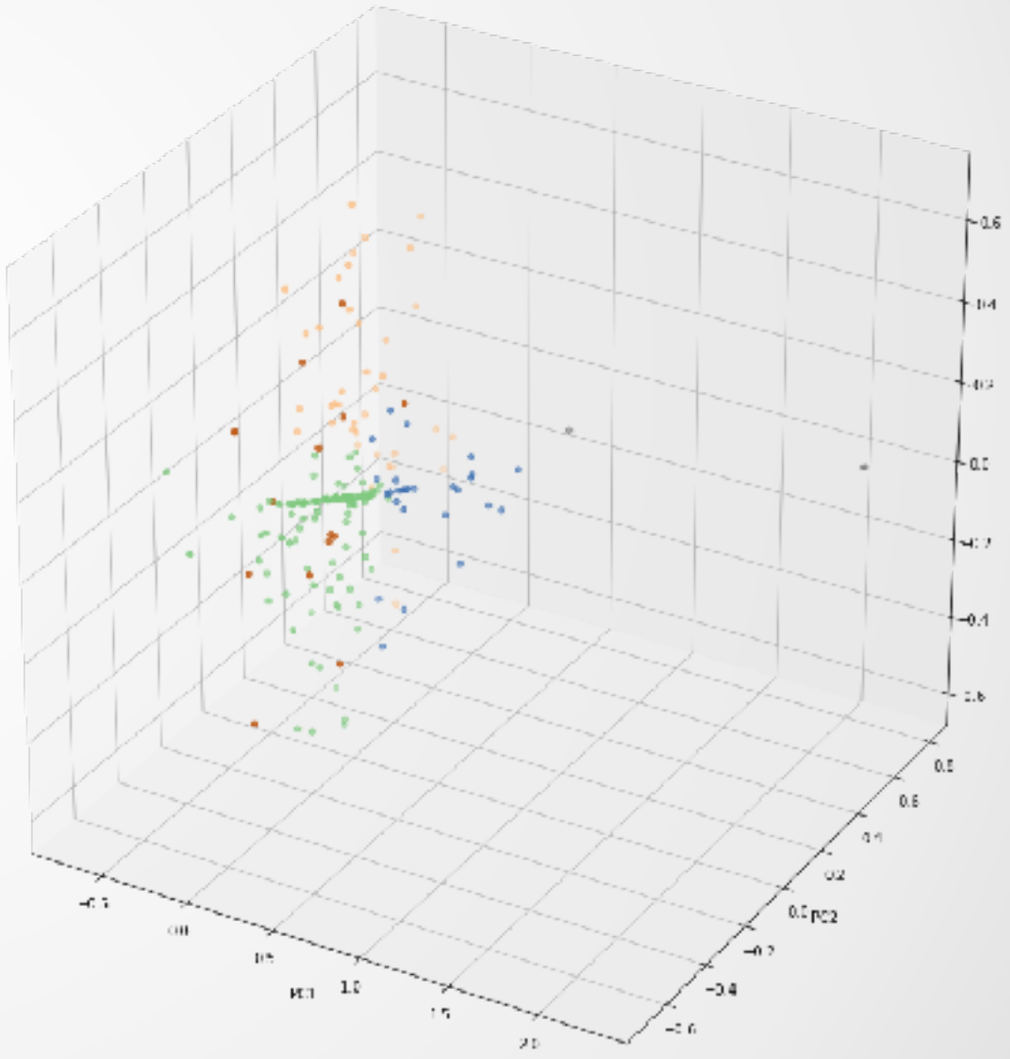
Clustering on CRIX-VCRIX 20160701-20191101
10 clusters

KEC_applications

FRM US stock market data



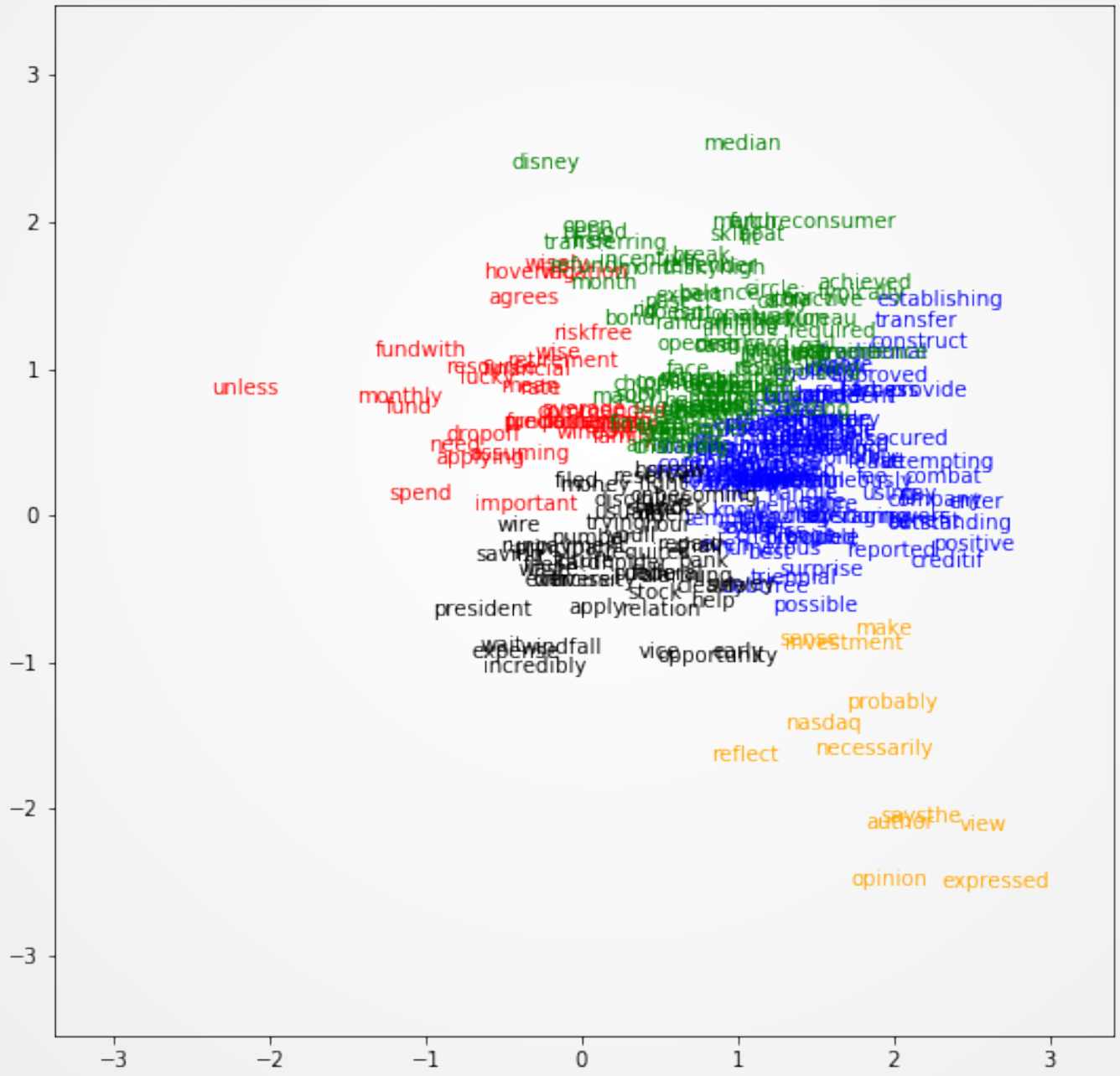




5 clusters on 100 components of log-return of 399 stocks,
$\tau = 0.95$, 20190604-20200309

KEC_applications

Text data (Nasdaq news on Personal finance)



Text data (Nasdaq news on Personal finance)

- **C1:** window eligible wise assuming riskfree mean rate need important precious hovering applying monthly fund lucky unless vacation retirement family financial agrees average resource drop off spend commended certain professor fundamental
- **C2:** closer combat triennial credit establishing new carefully despite obligation bad debt free doe even high guarantee fee away tool chunk company helping helpful offer trouble tempting benefit downside small establish line attempting foundation rebuild loan leave people know type using activity averaging debt surprise prudent post large outstanding deposit putting unsecured rolling welcome approved build plastic ass cruise limit challenge responsibly likely issuer big secured transfer cycle good traditional sneak higher pay step borrowing score advocate consolidation provide access positive example check starting denied numerous reported safe borrower savvy account construct choice possible credit if useful enter handle provided sure file spending simultaneously history repayment nfccs best

Text data (Nasdaq news on Personal finance)

- **C3:** taxpayer deductible maury use past rid opening return building carry month include disney changing tax expert start wont open period eliminate apr soon percent march card convenience suggestion willing refund caught way break randall lender bureau situation dearly office vicious quite doesnt required consumer skiboat circle monthskyhigh gail incentive american result behavior chipping face year student extend examine free finance report balance stop transferring convert fit charging attractive limited central tell median national insteadaccording futureconsumer product dont achieved cash typically bond time
- **C4:** sense nasdaq view make author necessarily investment probably says the expressed opinion reflect
- **C5:** knock the chair clearly stock repaid bank windfall hour reserve borrow run early federal payment apply wait right vice money waste cardholder youll usually filer discipline stand university incredibly on becoming help saving filed requires rider survey public expense alarming principal number opportunity future wire trying exercise relation president

Proof of convergence

Input: $X = (x_1, \ldots, x_n) \in (\mathbb{R}^p)^n; K$

Output: vector $C = (C(1), \ldots, C(n))$, where $C(i) \in (1, \ldots, K)$

Objective function:

$$G(\tau, \Theta, C) = \min_{\tau,\Theta,C} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{j=1}^{p} \left\{ \tau_{kj} + (1 - 2\tau_{kj}) \mathbf{I}_{\left\{ x_{ij} < \theta_{kj} \right\}} \right\} (x_{ij} - \theta_{kj})^2$$

Define

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

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

Proof of convergence

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

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

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

over all possible partitions

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

Hence, $G(C_{(i)}^{(t)}) \leq G(C_{(i)}^{(t-1)})$

Similarly, $G(\tau^{(t)}) \leq G(\tau^{(t-1)})$

Further step

- Local dimensionality reduction using PEC (Tran et al., 2019)
- Given a cluster $G = (\Phi_k, d, \tau_k, \theta_k)$, $\Phi_k = \{\phi_k^{(j)}\}_{j=1}^{d}$ are PEC components of the cluster; d is the reduced dimensionality; $\tau_k \in \mathbb{R}; \theta_k \in \mathbb{R}^d$

$$G^{K-expectiles}(\tau, \Theta, C, X) = \sum_{k=1}^{K} \sum_{j=1}^{p} dist(x_{.j}, \tau_j, \theta_k)$$

$$dist(x_{.j}, \tau_k, \theta_k, d, \Phi_k) = \left\{ \tau_k + (1 - 2\tau_k)\mathbf{I}_{\left\{ x_{.j} < \theta_{k,j} \right\}} \right\} \| x_{.j}\phi_k^{(j)} - \theta_{k,j} \|^2$$

- PEC is defined as maximiser of :

$$\phi_\tau = \arg max_{\phi \in \mathbb{R}^p, \phi^\top \phi} = \frac{1}{n} \sum_{i=1}^{n} (\phi^\top X_i - \hat{\theta})^2 \hat{w}_i$$

Dependence structure

Pearson's correlation only detect linear association

Time series has more complicated dependence structure which cannot be described only by covariance

Copulae generates joint cdfs by binding marginal distributions

$$F_1(x_1), F_2(x_2), \ldots, F_p(x_p)$$

$$H(x_1, x_2, \ldots, x_p) = C(F_1(x_1), F_2(x_2), \ldots, F_p(x_p))$$

Provides a multivariate dependence structure between random variables

Similar to Gaussian mixture model, could we use mixture copulae to estimate the probability of x_i belongs to the cluster k?

The sources

- Sun W, Wang J, Fang Y (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, *6*, 148-167.
- Newey, W. K., & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, 819-847.
- Maume-Deschamps, V., Rullière, D., & Said, K. (2017). Multivariate extensions of expectiles risk measures. *Dependence Modeling*, *5*(1), 20-44.

Further step

The right way of dimensionality reduction?

scaling? (local variance, τ-variance)

Proofs of consistency?

K-expectile clustering

Bingling Wang
Wolfgang Karl Härdle
Yingxing Li

Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
BRC Blockchain Research Center
lvb.wiwi.hu-berlin.de
Charles University, WISE XMU, NCTU 玉山学者