Regression clustering

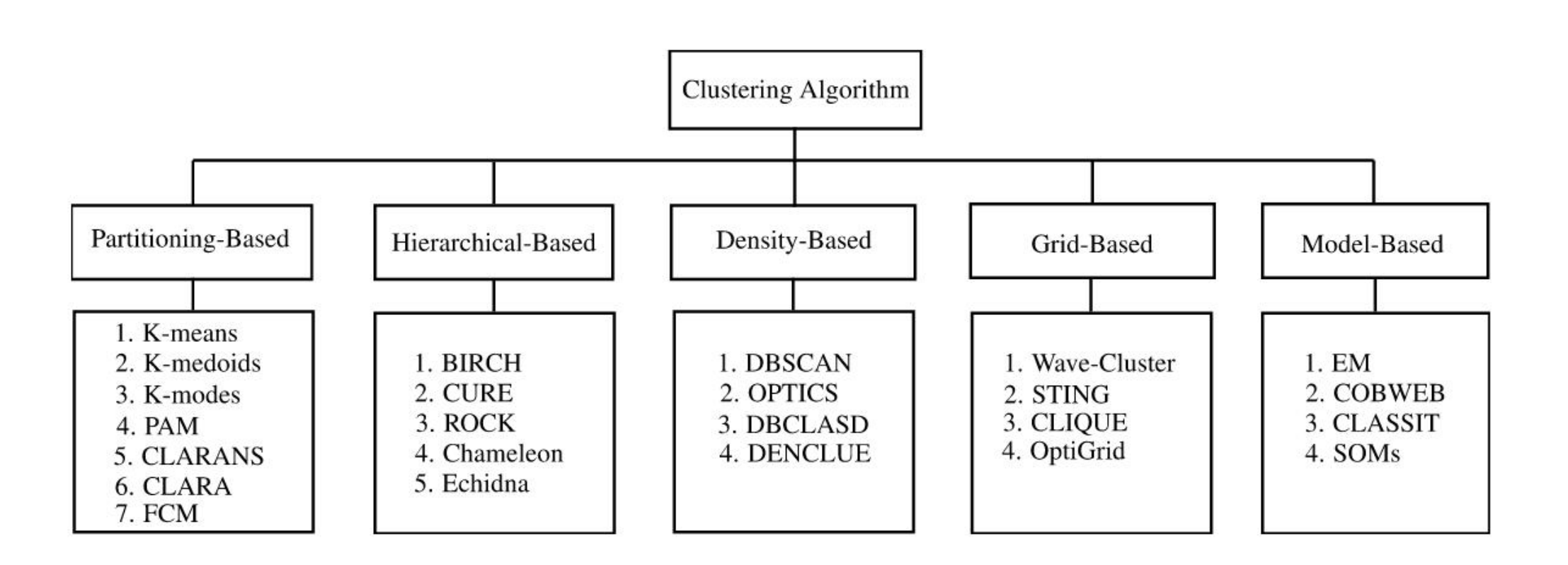
Huang Junjun

聚类

- 聚类是一个无监督学习过程（数据集没有类别标签），旨在将数据集划分成多个簇，使得簇内的相似度尽可能大、簇间的相似度尽可能小。经典的算法如kmeans。(matlab调用格式，[IDX, C] = kmeans(X, K))

- 聚类结果的好坏和往往多种因素相关，如样本量的大小、维度、簇的个数、簇的结构、簇间重叠的程度、子簇各自的大小、异常值个数以及误差扰动程度等等都有关系。

聚类算法



《Data Clustering Algorithms and Applications》,2014
《Handbook of cluster analysis》，2016
《Data Clustering Theory, Algorithms, and Applications》2007

Kmeans聚类(Hard clustering)

K-Means算法是一种无监督分类算法，假设有无标签数据集：

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{bmatrix}$$

该算法的任务是将数据集聚类成 k 个簇 $C = C_1, C_2, \ldots, C_k$ ，最小化损失函数为：

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Kmeans聚类(Hard clustering)

其中 μ_i 为簇 C_i 的中心点：

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in Ci} x$$

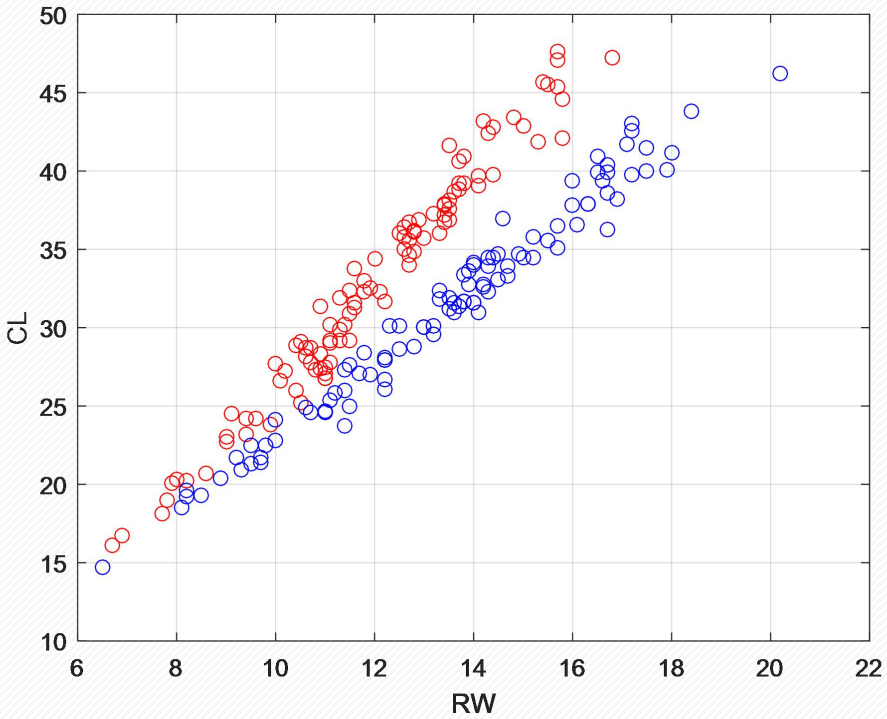
要找到以上问题的最优解需要遍历所有可能的簇划分，K-Mmeans算法使用贪心策略求得一个近似解，具体步骤如下：

1. 在样本中随机选取 k 个样本点充当各个簇的中心点 $\{\mu_1, \mu_2, \ldots, \mu_k\}$
2. 计算所有样本点与各个簇中心之间的距离 $dist(x^{(i)}, \mu_j)$，然后把样本点划入最近的簇中 $x^{(i)} \in \mu_{nearest}$
3. 根据簇中已有的样本点，重新计算簇中心

$$\mu_i := \frac{1}{|C_i|} \sum_{x \in Ci} x$$

4. 重复2、3

回归聚类



目标：聚类 + 估计K个回归关系

硬聚类To回归聚类

Object function :

$$\min_{\beta_j} \sum_{i=1}^{N} \sum_{(x_i,y_i)\in\beta_j} \left\| y_i - x_i\beta_j \right\|^2 = \sum_{i=1}^{N} \sum_{(x_i,y_i)\in\beta_j} \varphi(y_i - x_i\beta_j) \quad , j = 1,2,...,K \qquad (1)$$

其中φ(*)是某种距离映射; $(x_i,y_i)\in\beta_j$指样本i到超平面j的距离最近（最小），(x_i,y_i)“完全属于”超平面j，或者说样本i隶属于簇j的程度为1，隶属于其他簇的程度均为0。

硬聚类To回归聚类

硬聚类模型和相应的求解算法是最早被使用的模型算法，
求解硬聚类模型（1）通常做法是：
①随机初始划分K类，将其分解为K个子优化问题，并计算这K个子优化问题的参数最优解β_j（j=1,...,K）
注：有众多硬聚类SR的区别，主要是在于如何更新β_j
②根据得到的K个β_j，计算所有样本到这K个超平面距离，将样本分类到某超平面距离最小的那一类中，并依据新分类的结果重新计算K个子优化问题的最优解β_j_new.
③ 计算在新的回归系数下目标函数值与前一步迭代目标函数值的差别，|Obj_new-Obj_old|<ε,则停止迭代；否则，返回到②

模糊聚类To回归聚类

Fuzzy c-regression model（FCRM）：允许样本i隶属于簇 β_j的程度u_{ij}在[0,1]，而不是如硬聚类算法（隶属于某簇的值u_{ij}只能为0或1）。

最先将模糊聚类应用到SR(swtiching Regression)，由HathawayRJ,Bezdek[9]等人，模型如下：

$$\min E_m(U,\{\beta_k\}) = \sum_{k=1}^{K}\sum_{i=1}^{N} U_{ik}^m E_{ik}(\beta_k), \qquad (2)$$

其中 $E_{ik}(\beta_k) = (y_i - x_i\beta_k)^2, \sum_{k=1}^{K} U_{ik} = 1, for\ all\ i = 1,...N.$

Fuzzy Clustering Regression

求解模型(2)，采用的是坐标下降法（组坐标下降）

①随机初始化，分K类，计算得K个初始的回归系数β_k ；

②更新样本i隶属于簇k的隶属度u_{ik}和回归系数β_k；

$$U_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{E_{ik}}{E_{jk}} \right)^{\frac{1}{m-1}}}$$

$$\beta_i = [X^t D_i X]^{-1} X^t D_i Y$$ ，其中D_i为diag(u_{ik})的m次方

③判断是否停止迭代　**存在的问题：**初始化敏感(陷入局部最优)，对异常值或噪声敏感！

有限混合模型To回归聚类

对误差做某种分布假设，将样本看成是带有不同分布的混合分布，然后使用最大似然估计法（MLE）求解，求解MLE模型的算法通常使用的EM算法（Expectation maximum）

$$y_i = x^T \beta_j + \epsilon_{ij} \tag{1.4}$$

其中，$\beta_j = (\beta_{1j}, ..., \beta_{dj}), \epsilon_{ij} \sim N(0, \sigma_j^2)$. y 的条件密度函数为

$$f(y|x) = \sum_{j=1}^{K} \pi_j \phi(y; x^T \beta_j, \sigma_j^2) \tag{1.5}$$

存在同样的问题：初始化敏感(陷入局部最优)，对异常值或噪声敏感

Fuzzy c-Regression Clustering with L1 norm and maximum entropy

$$\min J_{FCRE\_L1}(U, \beta_k) = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}|y_i - x_i\beta_k| + \gamma \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}\log(u_{ik}).$$

- 1：通过引入L1距离度量，抵抗异常值和噪声干扰
- 2：熵正则项，通过调节r矫正模型的病态
- 3：ADMM算法更新迭代回归系数β_k

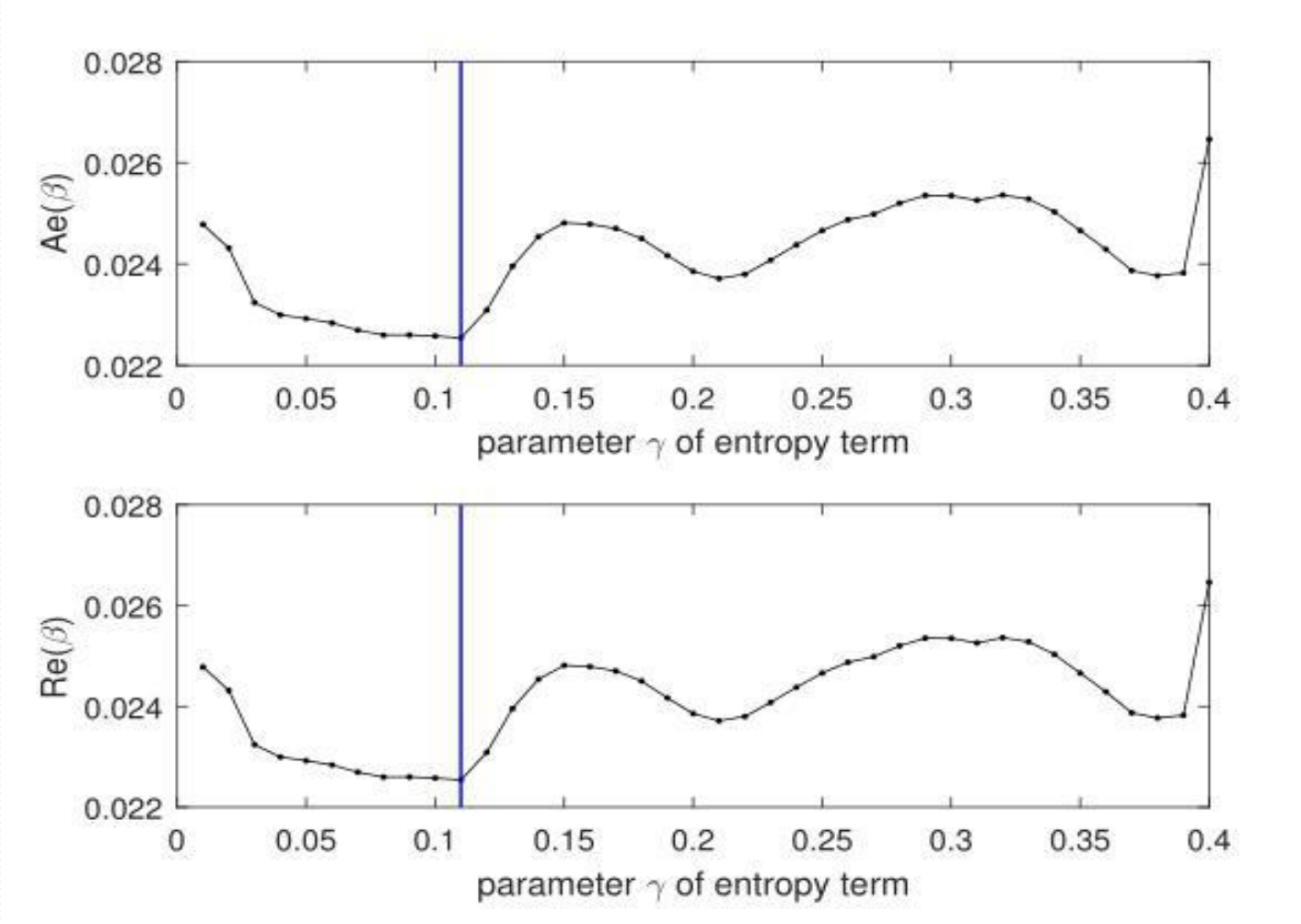
参数r的调节

- 模拟实验：从两方面考量(在不同的r下，随机初始化100次测试聚类结果)
- 1.精确性： 估计的回归系数β_k 与真实值β的绝对误差和相对误差，尽可能小
- 2.稳定性：同一参数r下，随机初始化100次数值结果要稳定（不能时好时坏）

- 结论：当r选择为误差的标准差附近时，二者效果最好

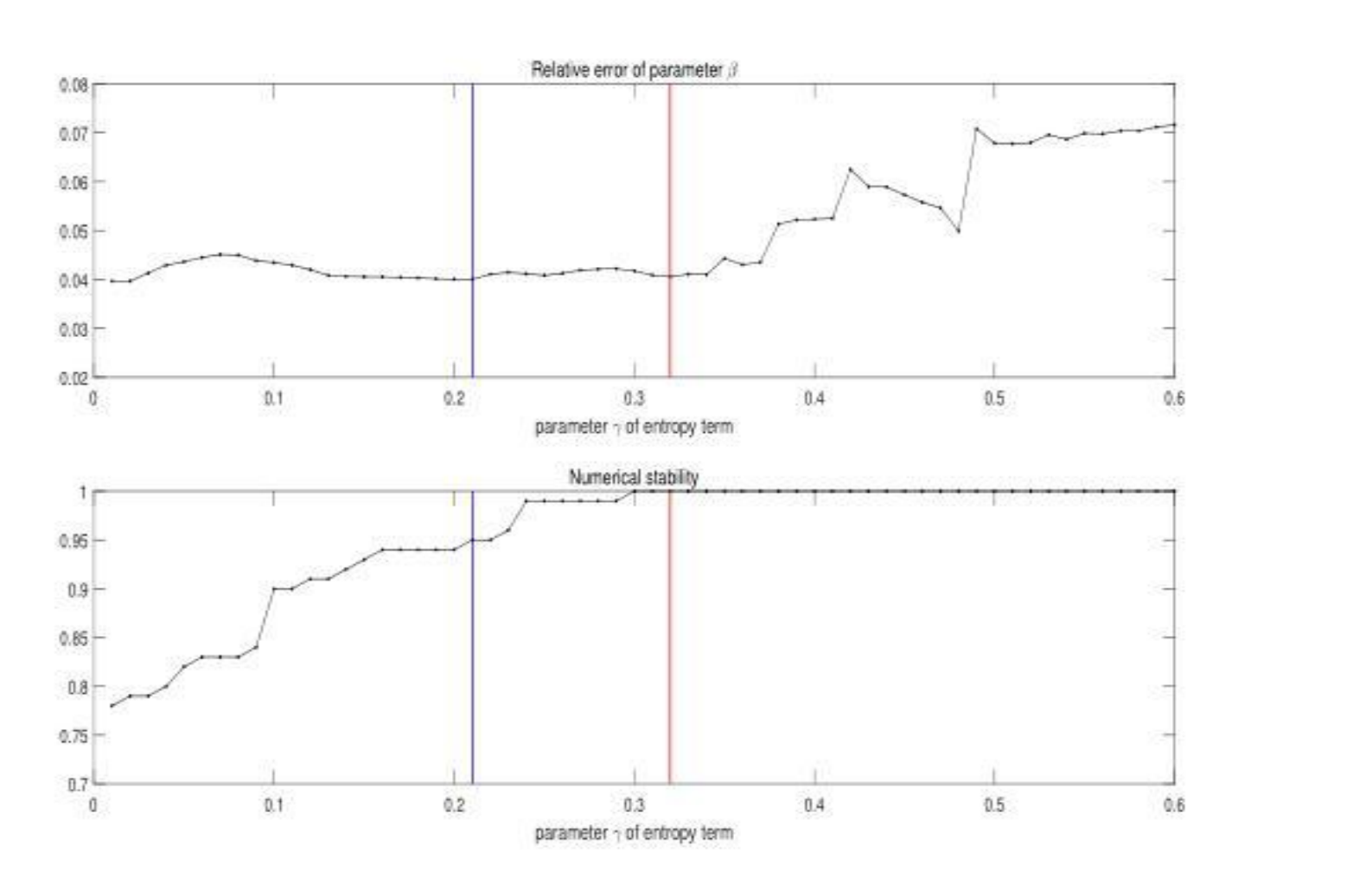
参数r的调节

σ =0.1


0.028
0.026
0.024
0.022
Ae(β)
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
parameter γ of entropy term
0.028
0.026
0.024
0.022
Re(β)
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
parameter γ of entropy term

参数r的调节

σ =0.2


Relative error of parameter β
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0
0.1
0.2
0.3
0.4
0.5
0.6
parameter γ of entropy term
Numerical stability
1
0.95
0.9
0.85
0.8
0.75
0.7
0
0.1
0.2
0.3
0.4
0.5
0.6
parameter γ of entropy term

实验比较

TABLE I: Best output comparison: FCRM($m = 2$),FCRa($m = 4$,$\alpha = 0.65$),FCRE($\gamma = 0.04$)

	true parameter	kmeans	kmeans_L1	FCRM	FCRM_L1	FCRa	FCRa_L1	FCRE	FCRE_L1	EM	EM_Lap
β_{10}	2	1.8768	1.9701	1.8301	1.9695	1.9375	1.9695	1.8733	1.9689	1.9165	1.9591
β_{11}	0	0.0705	0.0217	0.0938	0.0219	0.0346	0.0219	0.0710	0.0216	0.0428	0.0246
β_{20}	0	-0.7160	0.0035	-0.9824	0.0053	-0.1341	0.0053	-0.7147	0.0027	-0.6623	0.0027
β_{21}	1	1.3572	0.0035	1.6191	0.9978	1.0616	0.9978	1.3593	0.9989	1.3079	0.9991

实验比较

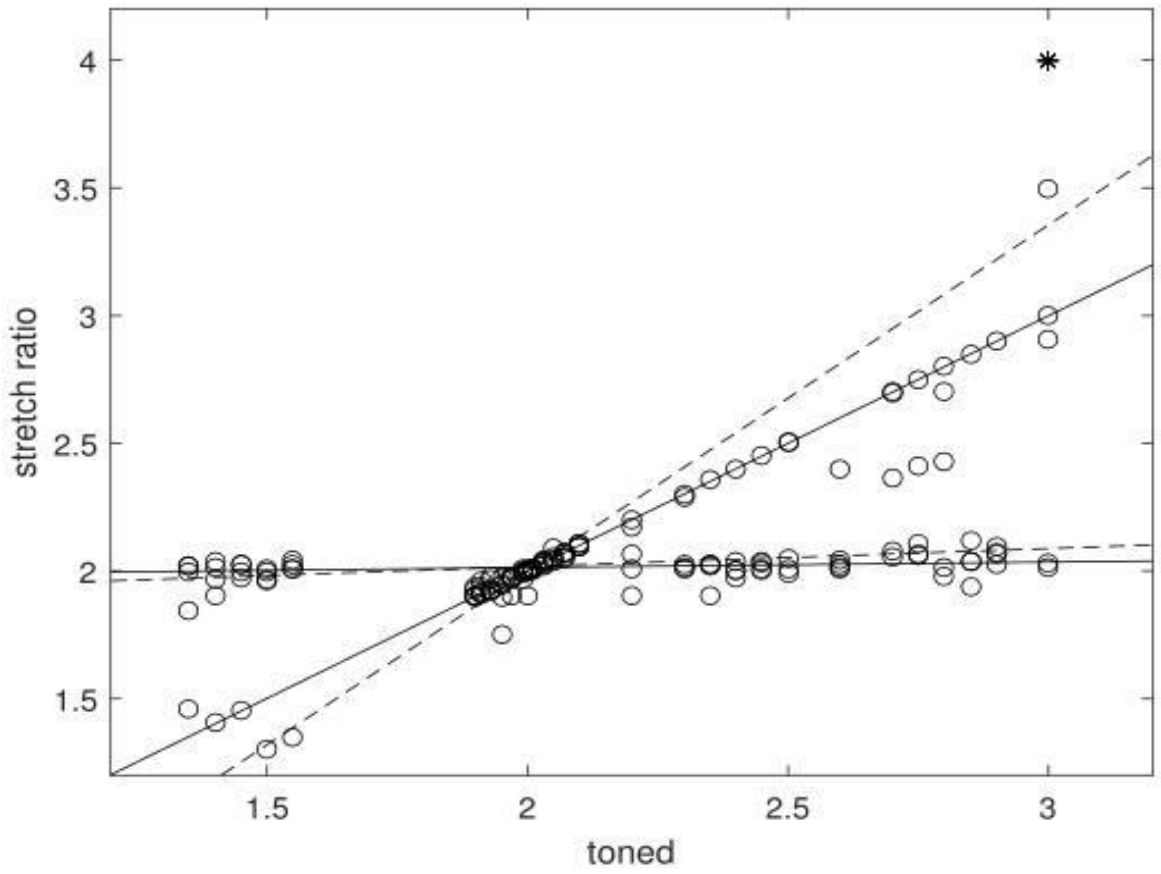


Fig. 3: Tonedata:kmeans(dash lines) and kmeans_L1(solid lines)

实验比较

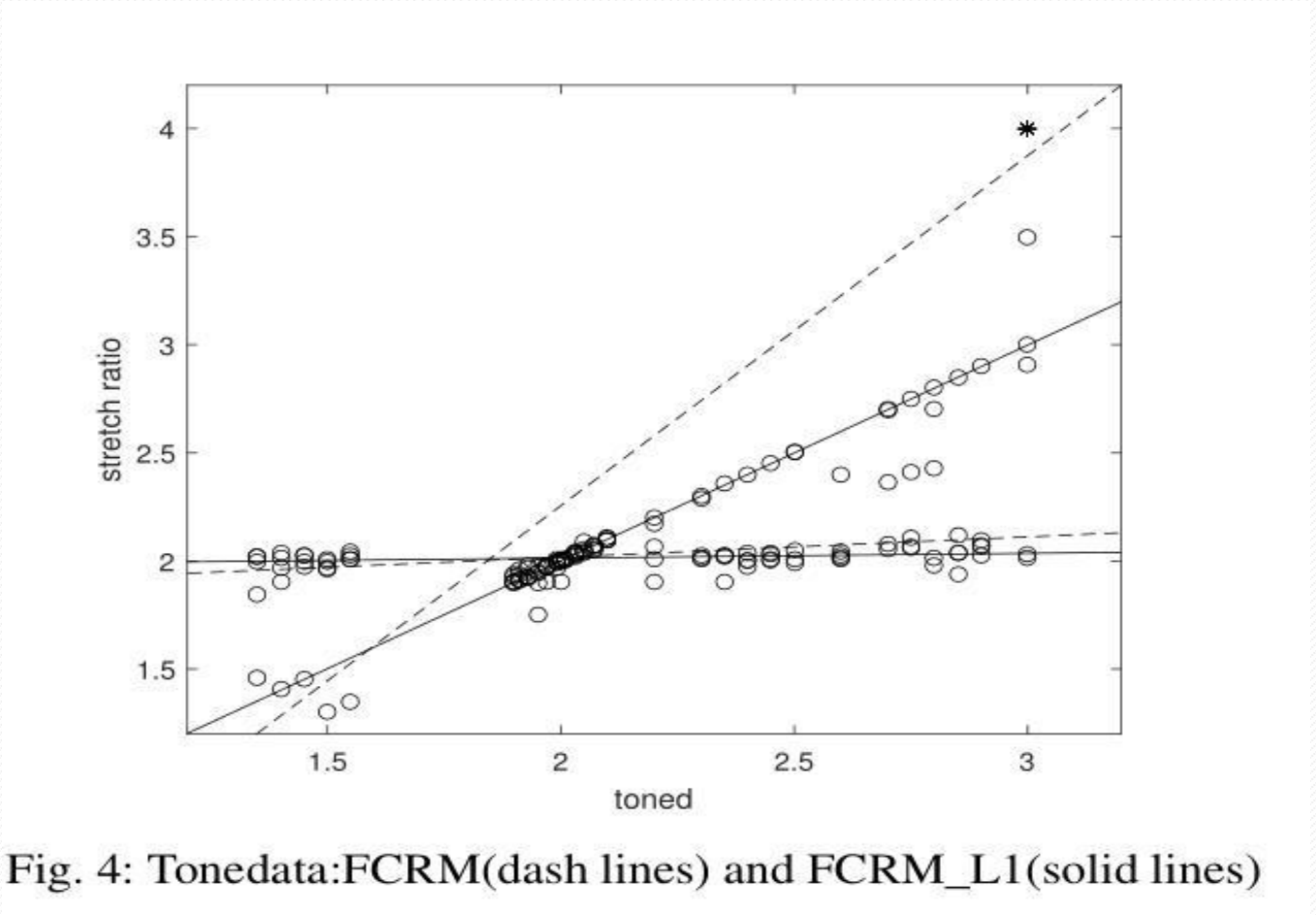


Fig. 4: Tonedata:FCRM(dash lines) and FCRM_L1(solid lines)

实验比较

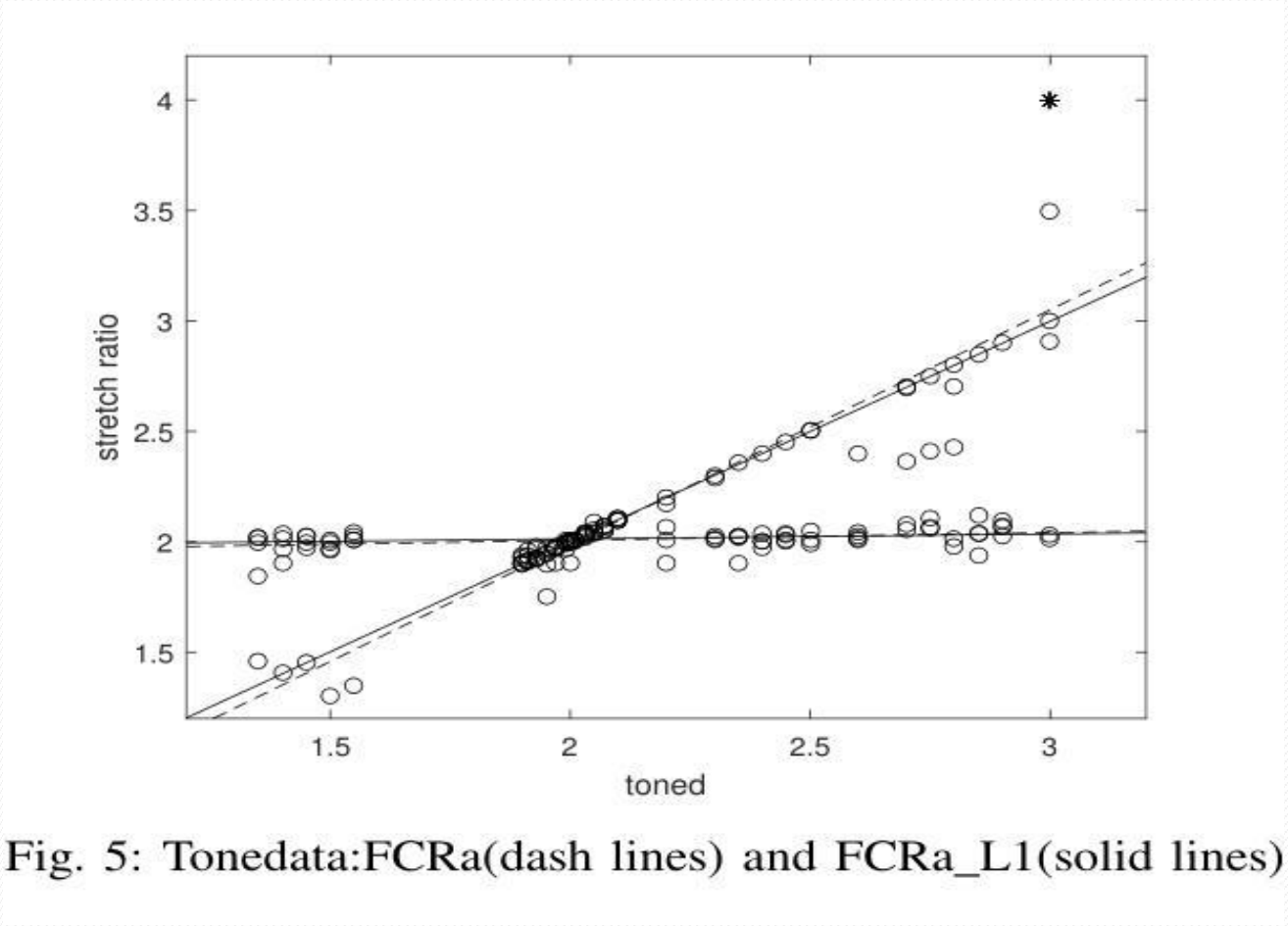


Fig. 5: Tonedata:FCRa(dash lines) and FCRa_L1(solid lines)

实验比较

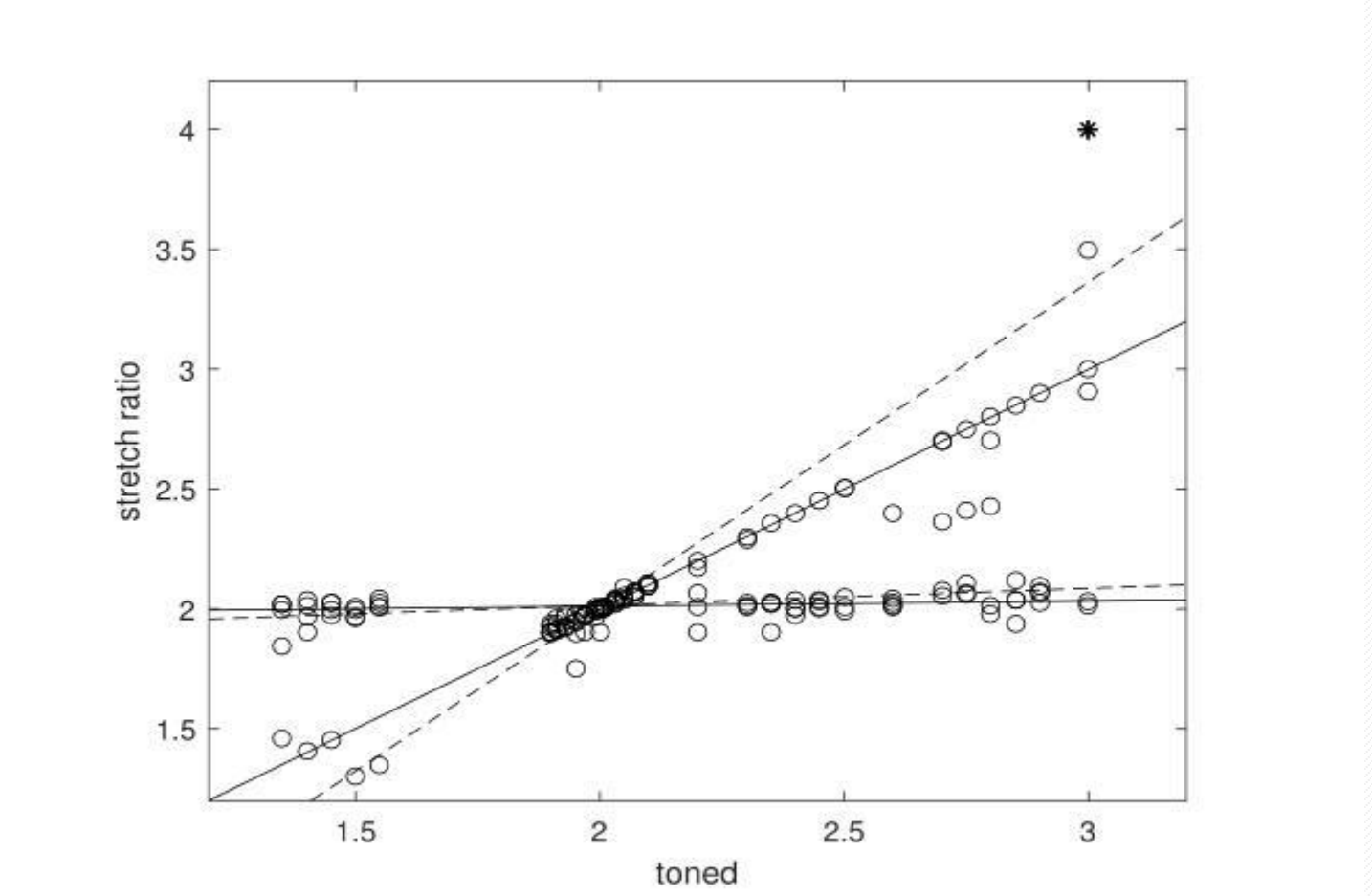


Fig. 6: Tonedata:FCRE(dash lines) and FCRE_L1(solid lines)

实验比较

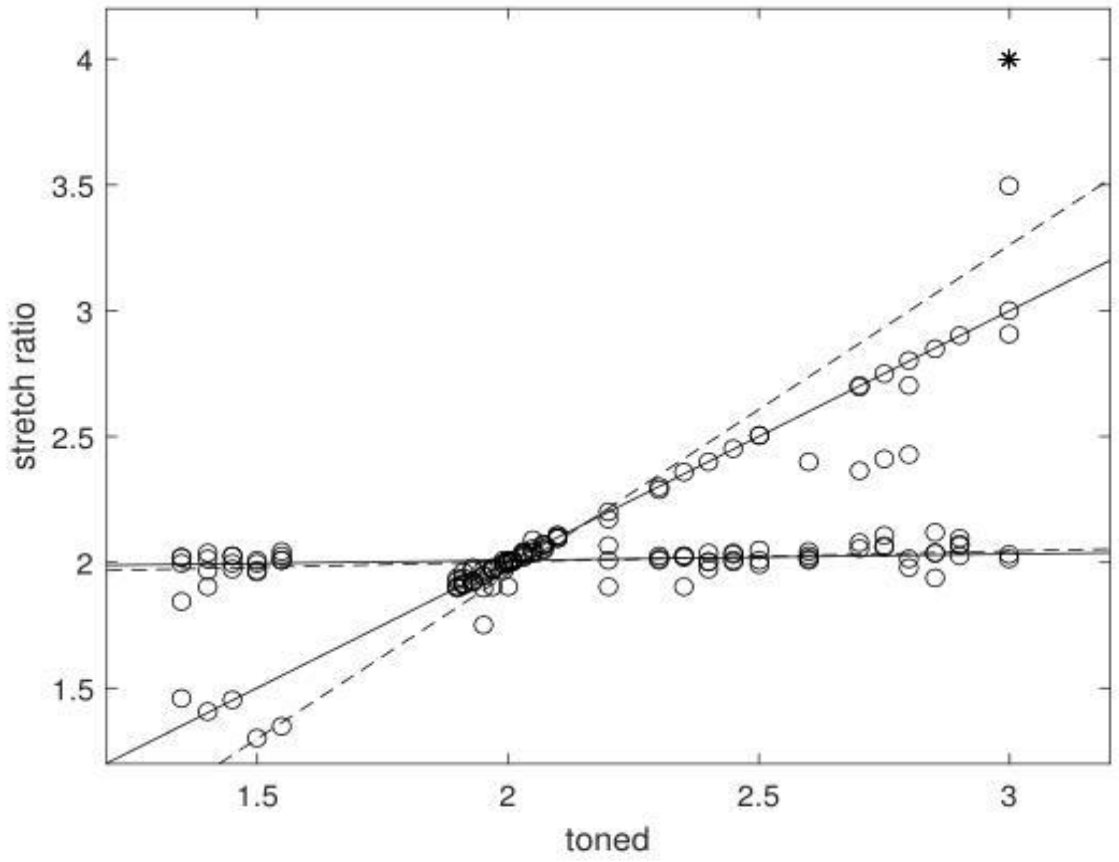


Fig. 7: Tonedata:EM(dash lines) and EM_Lap(solid lines)

实验比较

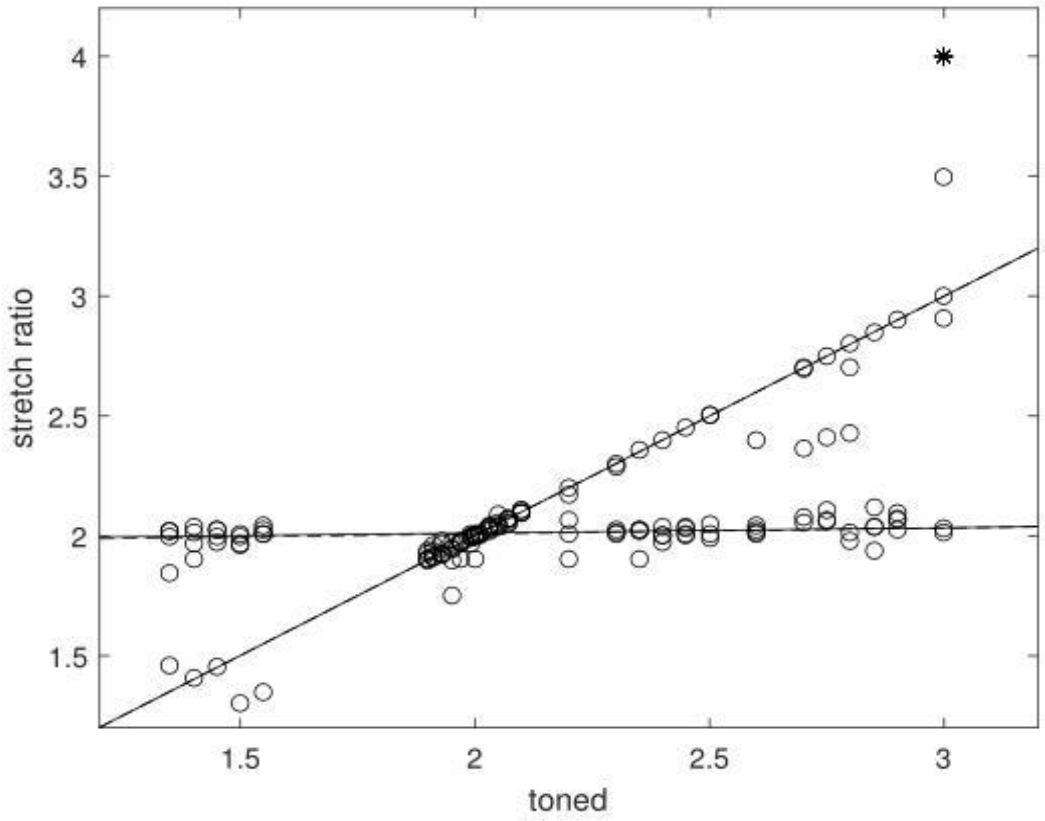


Fig. 8: Tonedata:EM_Lap(dash lines) and FCRE_L1(solid lines)

后续可能工作

- 1. 基于非参的方法做：如基于密度的聚类算法，基于网格的聚类做回归聚类
- 2.多个高斯分布问题，用fuzzy的视角做高斯混合模型(GMM)参数估计
- 3.基于Big Data背景，考量适用于大数据下的聚类算法
- [A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis]
- 4.应用层面：时间序列上的聚类，基因数据聚类，图像分割.....

Reference

[1]H. Späth. Clusterwise Linear Regression[J]. Computing, 1979, 22(4):367-373 4

[2]Zhu Z , Li Y , Kong N . Clusterwise Linear Regression with the Least Sum of Absolute Deviations – An MIP Approach[J]. Int.j.oper.res, 2012(3):162-172.

International Transactions in Operational Research(管理科学类，JCR三区)

[3]Schlittgen R . A weighted least-squares approach to clusterwise regression[J]. Asta Advances in Statistical Analysis, 2011, 95(2):205-217.(数学类，JCR三区)

[4]Réal A. Carbonneau, Caporossi G , Hansen P . Extensions to the repetitive branch and bound algorithm for globally optimal clusterwise regression[J]. Computers & Operations Research, 2012, 39(11):2748-2762. （工程技术3区）

[5]Suk H W , Hwang H . Regularized fuzzy clusterwise ridge regression[J]. Advances in Data Analysis and Classification, 2010, 4(1):35-51. (数学类2区)

[6]Caporossi G , Hansen P , Carbonneau, Réal A. Globally Optimal Clusterwise Regression by Mixed Logical-Quadratic Programming[J]. European Journal of Operational Research, 2011, 212(1):213-222. (管理科学类2区)

Reference

[7] Tan T , Suk H W , Heungsun Hwang…. Functional fuzzy clusterwise regression analysis[J]. Advances in Data Analysis and Classification, 2013, 7(1):57-82.(数学类2区)

[8] Bagirov A M , Ugon J , Mirzayeva H G . An algorithm for clusterwise linear regression based on smoothing techniques[J]. Optimization Letters, 2015, 9(2):

375-390.(数学类三区)

[9] HathawayRJ,Bezdek JC(1993) Switching regression models and fuzzy clustering.IEEETransFuzzySyst1:195–204 (工程技术类1区)

[10] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, “A possibilistic fuzzy

c-means clustering algorithm,” IEEE Trans. Fuzzy Syst., vol. 13, no. 4,

pp. 517–530, Aug. 2005.(工程技术类1区)

[11] W.Pedrycz,“Conditional fuzzy c-means,”PatternRecognit.Lett.,vol.17,

no. 6, pp. 625–632, May 1996.

Reference

[12] N. B. Karayiannis, “MECA: Maximum entropy clustering algorithm,” in Proc. 3rd IEEE Int. Conf. Fuzzy Syst., Orlando, FL, 1994, vol. 1,pp. 630–635...

[13] R. Krishnapuram and J. M. Keller, “A possibilistic approach to clustering,” IEEE Trans. Fuzzy Syst., vol. 1, no. 2, pp. 98–110, May 1993.(工程技术类1区)

[14] D. Özdemir and L. Akarun, “A fuzzy algorithm for color quantization of images,” Pattern Recognit., vol. 35, no. 8, pp. 1785–1791, Aug. 2002.(工程技术类2区)

[15] K. L. Wu, J. Yu, and M. S. Yang, “A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality test,” Pattern Recognit. Lett., vol. 26, no. 5, pp. 639–652, Apr. 2005.(工程技术类2区)

[16] Devijver E . Finite mixture regression: A sparse variable selection by model selection for clustering[J]. Electronic Journal of Statistics, 2014.(math3区)